AN EVALUATION OF WATER QUALITY INFORMATION: A CASE STUDY OF STREAMS IN METROPOLITAN NEW JERSEY

Presented at the
National Symposium on Societal Problems of Water Resources
at the
65th Annual Meeting of the Illinois State Academy of Science
Western Illinois University, Macomb, Illinois
April 28-29, 1972

ROBERT M. HORDON
Department of Geography
Rutgers University
New Brunswick, New Jersey 08903

ABSTRACT.—Some of the best water quality data sets available in New Jersey are routinely collected by large public potable water supply agencies. One such set consisted of eleven variables (temperature, dissolved oxygen, turbidity, color, discharge, percent saturation, hardness, alkalinity, BOD, pH, and bacteria) collected every week at three sites in the Raritan River Basin. Factor analysis using the Varimax rotation resulted in three factors: an oxygen-related factor (high inverse loadings on dissolved oxygen and temperature), an appearance factor (turbidity, color and discharge), and a third variable factor occasionally loading on percent saturation. The cumulative percentage explained by the rotated factors declined during the decade, suggesting an increase in independence developing among the variables. It is hypothesized that urbanization with its commensurate changes in land use and runoff patterns might be intervening in the natural ecology of the stream. The effects of urbanization, then, would be to interfere with instream interaction among the variables. Factor structures were quantitatively compared from year to year and from site to site by a computer program called RELATE.

The purpose of this study is to evaluate the available water quality data in metropolitan New Jersey by the application of a factor analytic methodology. The decade of the 1960's was selected, as it included the continuation of urbanization with its resultant impact on watershed quality and periodic events such as the record drought of 1962-66. It was hoped that the relative effects of these trends and events could be isolated by factor analysis.

Factor analysis is one form of multivariate analysis wherein a set of intercorrelated variables is collapsed to form a smaller number of composite variables, or factors (King, 1969; Rummel, 1970). These are then rotated to yield a set of independent, uncorrelated factors.

Numerous investigators in a variety of disciplines have used factor analysis in their studies. For example, Weaver (1954) studied crop-combination patterns in the Midwest, using 88 counties and seven variables. Carey (1966) offered an interesting interpretation of the housing patterns of the population in Manhattan, based upon 33 socioeconomic variables and 269 census tracts.

In the field of hydrology, Wong (1963) used a modified principal components analysis on 12 variables and 90 basins in New England in order to estimate the magnitude of the mean annual flood. Wallis (1965) used factor analysis in a study of the agents that contribute to soil erosion and stream sedimentation in northern California. Dawdy and Feth (1967) applied factor analysis to results of chemical analyses of ground water samples from the Mojave valley in California. Matalas and Reiher (1967) felt that factor analysis was technically underdeveloped, a view seconded by Wallis (1968) who also felt, however, that the technique was a powerful tool for screening variables. Rice (1967) suggested Varimax rotation as an aid in rationally interpreting factors resulting from a study of hydrologic relationships. Eiselstein (1967) used a principal component regression analysis with Varimax rotation of the factor weight matrix as a means of synthesizing flow records for small ungaged watersheds. Annual precipitation and runoff data from 14 watersheds in Ohio and 7 watersheds in Texas were subjected to principal component and factor analysis by Diaz, Sewell, and Shelton (1968) in an attempt to identify the factors affecting the water yield of the basins. Knisel (1970) used factor analysis on 13 variables for five reservoirs in Texas to determine which variables were significant for estimating reservoir losses on the fissured limestone terrain.

METHODOLOGY

The factor analyses discussed in this paper were predicated on a number of conservative assumptions. First, the communality was estimated by using the square of the multiple correlation coefficient between each variable and all other variables in the data set. These communalities would appear along the principal diagonal of the correlation matrix.

Alternative estimates of the communality could have been chosen; for example, ones could be inserted in the principal diagonal. This procedure then assumes that all of the variance of each variable is related to the common factors of the data set. The true communality presumably lies between the R^2 and unity estimates; thus, the R^2 estimate is probably more realistic.

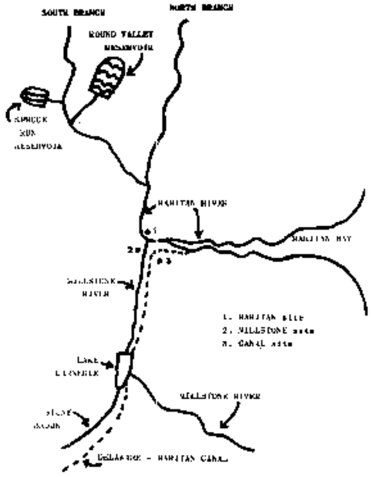
Secondly, the general factor analysis program used (BMD 03M) extracts a set of factors by the principal components method. These factors were then rotated so as to concentrate the loadings on a few of the factors. The rotation used was the Varimax solution, which involves a series of orthogonal transformations of factor pairs. The Varimax routine is perhaps the most commonly used form of rotation.

Rotations were performed on factors only if the eigenvalues exceeded unity. In almost all cases, this arbitrary but conservative rule ensured that all factors accounted for at least as much as one of the original variables. In order to ensure uniformity among factor structures on some data sets, some factor analyses were repeated with a stipulated but reasonable number of factors to be rotated.

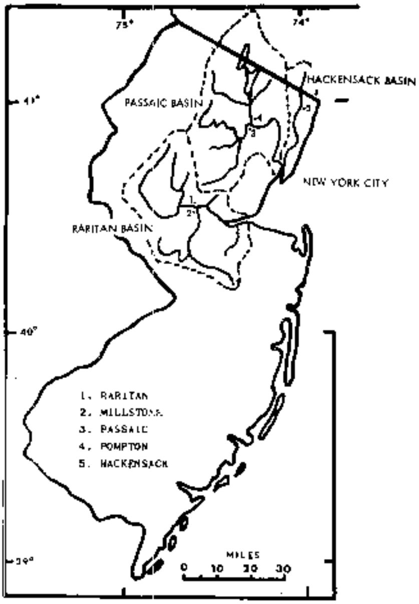
Quantitative factor structure comparisons can be achieved by using a computer program called RELATE. As discussed by Veldman (1967) this program accepts as input the factor loading matrices that were obtained from an orthogonal factor analysis of identical sets of variables. Then the factor axes are rotated until maximum overlap between corresponding test vectors in the two structures is attained. The degree of rotation required is expressed as the cosine of the angle between the factor axes. These cosines may be interpreted as correlations between the factor variables.

Acquisition of the requisite water quality data sets was an immediate and troublesome task. Preliminary investigation revealed sampling inadequacies in frequency, duration, and number of variables. Data storage and retrieval problems were also evident. From the mix of public and private agencies that operate stream sampling stations in New Jersey, it was determined that some of the best sets of water quality information were routinely collected by large potable water supply agencies.

Fortunately, an excellent set of data for the Raritan Basin was available in the files of the Elizabethtown Water Company (EWC), one of the largest private water purveyors in the country. In 1971, EWC distributed 117 mgd (441,090 m³/day) to 44 municipalities in central New Jersey (EWC Annual Report, 1971). Most of the water (about 75%) is diverted from the Raritan and Millstone Rivers near Bound Brook (see Figs. 1 and 2) and the Delaware-Raritan Canal which flows parallel to the Millstone.



2. Schematic Map of the Raritan River Basin.



1. Major Watersheds in Northeastern New Jersey.

EWC samples its raw water on a daily, weekly, and monthly basis for 6, 9, and 17 variables, respectively, for each of the three intakes near the filter plant. With the addition of discharge and percent saturation, the best mix of frequency of observation and total number of variables is attained with the weekly series.

The Raritan River Basin is one of the largest in New Jersey, covering 1,100 square miles (2849 sq. km.). The Raritan River rises in the Highlands of New Jersey and then flows eastward through the Piedmont Province. The total length of the Raritan is over 74 miles (119 km.), with the tidal portion accounting for 19 miles (31 km.) or 26%. The drainage area just above the confluence with the Millstone River is 490 square miles (1269 sq. km.). The average discharge for 52 years of record is 716 cfs (20.3m³/sec.), or 1.46 cfs/square mile (0.016 m³/km²).

The Millstone River rises in the Coastal Plain Province, flows westward and then northward through the Piedmont Province where it joins the Raritan near Bound Brook (see Fig. 2). The drainage area is over 260 square miles (673 sq. km.) with a main channel length of 28 miles (45 km.). The average discharge for 49 years of record is 356 cfs (10.1 m^3/sec.), or 1.38 cfs/square mile (0.015 m^3/km^2).

Following adjacent to the Millstone and then the Raritan River is the Delaware-Raritan Canal. By virtue of a U.S. Supreme Court decision in 1954, New Jersey can divert up to 100 mgd (377,000 m^3/day) of raw water from the Delaware River at Raven Rock, about 20 miles (32 km.) upstream of Trenton (New Jersey, Commission on Efficiency, 1967). The water then flows by gravity through the 55-mile (88 km.) long canal until it empties into the tidal Raritan at New Brunswick. Consequently, Canal water comes from a basin of approximately 6,350 square miles (16,447 sq. km.), about 13 and 24 times larger than the gaged portions of the Raritan and Millstone basins, respectively.

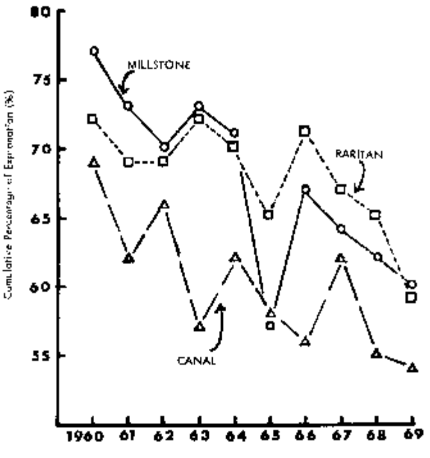
ANALYSIS OF THE DATA

Eleven variables were available at each of the three EWC stations: temperature, pH, DO, turbidity, BOD, color, alkalinity, hardness, bacteria, discharge, and percent saturation. The ten years of weekly observations comprise 5,720 bits of information at each site, or over 17,000 bits for the three sites. The data were arranged in calendar year sets.

The first run of factor analyses resulted in structures about evenly divided between two and three factors. In order to preserve year to year consistency in interpretation, the data were factor analyzed again, this time specifying that three factors were to be rotated. The following discussion is based on the ordered three factor runs.

The cumulative percentage of explanation of the rotated factors is plotted in Figure 3. The Millstone decreases from a peak of 77% in 1960 to a low of 57% in 1965, averaging 67% for the decade. The Canal structures yield the lowest percentages of explanation, ranging from 69% in 1960 to 54% in 1969, averaging 60% for the decade. Although the decline is not steady from year to year, a pronounced secular decrease in explanation is apparent for all three intakes. This decline suggests an interference in in-stream interactions among the variables. It is hypothesized that human interferference might be the causative agent, in the form of increased effluent discharges and accelerated runoff commensurate with urbanization and land use changes.

The rotated factor loadings for selected years are shown in Table 1. For clarity, only the highest loading for each variable is shown. Thus, a



3. Cumulative Percentage of Explanation of the Rotated Factors, Elizabethtown Water Company.

Table 1.—*Rotated Factor Loadings for Selected Years of the Elizabethtown Water Company Data*

Intake	Factor	Percent[1] Expl.	Temp.[2]	DO	BOD	pH	Alk.[3]	Q[4]	Turb.[5]	Color	Hard.[6]	Percent[7] Sat.	Bact.[8]
Canal	1	30	—.92	.85	.76			—.69			.72		
1960	2	24				.64	.76		—.77	—.80			
	3	15										.94	—.31
	All	69											
Canal	1	23	—.93	.86	.70								
1969	2	16				.55	.61		—.64	—.67	.39		
	3	15						—.45			—.37	.94	—.38
	All	54											
Millstone	1	34	—.93	.99	.69		—.67					.84	
1960	2	31						.80	.97	.96			.75
	3	12				.64					.59		
	All	77											
Millstone	1	24		.87	.74							.97	
1965	2	21	—.81			—.58	—.62	.51					—.64
	3	12							.73	.66	—.27		
	All	57											
Raritan	1	29	—.93	.85	.69	—.55	—.58						
1960	2	28						.75	.89	.94			.61
	3	15									.56	—.82	
	All	72											
Raritan	1	24	—.94	.94	.74	.46	.52						—.25
1969	2	24						—.88	—.78	—.81			
	3	11									.37	—.91	
	All	59											

Notes:
[1] Percent Expl.—Percent Explanation
[2] Temp.—Water Temperature
[3] Alk.—Alkalinity
[4] Q—Discharge
[5] Turb.—Turbidity
[6] Hard.—Hardness
[7] Percent Sat.—Percent Saturation
[8] Bact.—Bacteria

blank indicates a factor loading less than .70, a conservative threshold for significance. The remaining 24 factor structures that are not shown in Table 1 are broadly similar to the six representative structures. The similarities and differences between and among the entire 30 factor structures will be discussed in a later section.

The factor structures generally reveal the following characteristic pattern: (1) an oxygen-status factor, represented by high loadings on temperature, DO, and BOD; (2) an appearance factor based on turbidity, color, and discharge; and (3) a third variable factor, loading occasionally on percent saturation.

The oxygen-status factor tends to be the component of greatest statistical importance. Temperature and DO almost always have the highest loadings in the factor matrix, usually about .9. The loadings are inversely related, which is in accord with the temperature-DO relationship. Note in Table 1 that in 1965 on the Millstone this relationship is severed, as temperature and DO split their association between two factors. Presumably, this anomaly was caused by drought conditions, as 1965 was the driest year in the 1962-66 drought. Although not shown in Table 1, temperature and DO were also separated in 1965 for the Canal.

The second most important factor in terms of explanation is one based on appearance, *i.e.,* high loadings on turbidity, color, and occasionally discharge. Note that bacteria forms some association with the factor in 1960 for the Millstone and Raritan.

High loadings on percent saturation occasionally form a separate third factor. This last factor generally offers only half the explanation of the first factor.

Bacteria, hardness, pH, and alkalinity generally exhibit the lowest loadings of any of the 11 variables. Indeed, bacteria is the least related to any of the other variables in the data set.

FACTOR STRUCTURE COMPARISONS

In order to assess inter-basin and inter-basin changes in watershed characteristics, the EWC factor structures were compared in space and over time. As mentioned previously, the computer program RELATE (Veldman, 1967) quantitatively compares one factor structure with another as long as the variables are the same and remain in identical order.

RELATE enables one to indicate the degree of similarity between factor structures by obtaining the cosine of the angle between component vectors. These cosine values may then be interpreted as correlation coefficients. Perfect identity in the comparisons would be denoted by an identity matrix I with ones in the principal diagonal. Hughes (1971) used RELATE in his study of major urban metropolitan systems. As Hughes states, "The greater the value of the off diagonal, the greater the difference between the corresponding factors."

The following three examples should clarify the discussion. Consider the following three matrices:

$$\begin{array}{lccc} & & \text{System B} & \\ \text{I. System A} & 1 & 0 & 0 \\ & 0 & 1 & 0 \\ & 0 & 0 & 1 \end{array}$$

$$\begin{array}{lccc} & & \text{System B} & \\ \text{II. System A} & 0 & 1 & 0 \\ & 1 & 0 & 0 \\ & 0 & .0 & 1 \end{array}$$

$$\begin{array}{lccc} & & \text{System B} & \\ \text{III. System A} & .80 & .05 & .60 \\ & .05 & .99 & .01 \\ & .60 & .02 & .80 \end{array}$$

A perfect identity is indicated in example I. The first factor in A is identical to the first factor in B, the second factor in A is identical to the second factor in B, and so on.

Similar factor structures are shown in example II, but the individual factors vary in their importance within each system. Thus, the second factor in A is identical to the first factor in B, since a cosine of 1.0 indicates a zero degree angle between the cosines.

Dissimilar structures are indicated in example III, where the first and third factors of A and B are less than perfectly associated with a value of .80. Only the second factors of A and B with a cosine of .99 come very close to identity.

Summarizing, any value less than unity in the diagonal implies some difference between the corresponding structures.

The discussion in this section will focus on the comparison between factor structures for each intake from year to year and from intake to intake for the same year.

The Canal provides the best year to year comparisons. As shown in Table 2, the mean value for all years exceeds .80. Indeed, only once during the decade (1965-66) did the mean cosine fall as low as .81. For five of the nine years of comparison (1960-61, 1962-63, 1963-64, 1966-67, and 1967-68), the mean value was greater than or equal to .95. The grand mean for the decade was .93, exceeding the values of .91 and .87 for the Millstone and Raritan, respectively. Thus, the Canal provides the most consistent set of factor structures of the three intakes. This consistency is attributed to the regularity of the Canal discharge. Also, Canal water is coming from a large basin which tends to dampen water quality variation.

Conversely, the Raritan structures show the lowest year to year comparisons. Note the .71 and .79 values for 1960-61, and 1962-63 and 1967-68 (Table 2). Interestingly, the peak mean values of .99 for 1964-65 and 1965-66 contrast markedly with the same periods for the Canal, when decade minimums were reached. Pre-

TARLE 2
Intra-Basin Comparison: Mean Cosine Values of Factor Structures for Consecutive Years for Each Intake

		Mean Cosine Value	
Year	Canal	Millstone	Raritan
1960-61	.95	.89	.71
1961-62	.94	.80	.82
1962-63	.97	.89	.79
1963-64	.95	.97	.88
1964-65	.89	.94	.99
1965-66	.81	.84	.99
1966-67	.95	.94	.97
1967-68	.99	.95	.79
1968-69	.89	.99	.87
Decade Mean	.93	.91	.87

sumably, the flow augmentation on the Raritan accounts for the disparity between the two stations. The decade grand mean of .87 is the lowest of the EWC stations.

The Millstone values are intermediate among those of the Canal and Raritan. The lowest mean value of .80 was reached in 1961-62, whereas higher values were attained in the latter part of the decade. In three years out of the nine, mean values were equal to or greater than .95 (Table 2).

Summarizing, although there is some variation in year to year factor comparisons among the three sets, the dominant theme is one of factor stability over time. That is, the similarities from year to year are greater than the differences. Thus, the results from the structure comparison part of the RELATE program (cosine values) support the earlier findings of two or three characteristic water quality factors.

Turning now to inter-basin factor comparisons within the Raritan, we find in Table 3 that the Canal is structurally quite similar to the Millstone. This is surprising, inasmuch as Canal water is really Delaware River water which is coming from a basin about 24 times as large as the Millstone. The mean cosine value varies from a low of .75 in 1960 to a high of .99 in 1965, averaging .89 for the decade. There is noticeable fluctuation from year to year, but the associations remain strong.

Structural congruence between the Canal and Raritan exists, but at a lower level. For three years in the decade, the mean cosine value is less than or equal to .75. The overall decade mean drops to .84, as compared to .89 for the Canal and Millstone. Thus, we may conclude that even though the Canal water is coming from a different basin, the factor structures on the Canal are similar to those of the Millstone and Raritan.

As shown in Table 3, the factor structures between the Millstone and Raritan are strongly associated. In only one year (1967) did the mean cosine value drop as low as .75. For

TABLE 3

Inter-Basin Comparison: Mean Cosine Values for Paired Basins For Each Year

		Mean Cosine Value	
Year	Canal and Millstone	Canal and Raritan	Millstone and Raritan
1960	.75	.86	.91
1961	.98	.75	.92
1962	.87	.84	.94
1963	.80	.96	.81
1964	92	.70	.98
1965	.99	.95	.91
1966	.87	.92	.84
1967	.95	.78	.75
1968	.82	.68	.91
1969	.97	.92	.90
Decade Mean	.89	.84	.89

seven years out of the decade, the mean values were equal to or greater than .90. This structural similarity is to be expected, in view of the fact that both streams drain major subbasins within the same watershed.

The test R portion of the RELATE program measures the degree of association between variables from year to year within a basin and from basin to basin for paired years. Briefly, the highest correlation coefficients are reported for temperature, DO, and percent saturation, whereas hardness, bacteria, and occasionally BOD represent the more erratic variables. Generally, most variables are highly consistent from year to year. A fuller discussion of the test R results is contained in the larger study of water quality in the New York Metropolitan Region by Carey, Zobler, Greenberg, and Hordon (1972).

In conclusion, the factor structures of the Canal show a surprising similarity to those of the Millstone during the decade of the 1960's. The Canal also displays a structural congruence to the Raritan, but to a lesser degree. As might be expected for sub-basins of a larger basin, the Raritan and Millstone are structurally similar. Thus, even though the decade of the 1960's included a record-breaking drought, the similarities among the factor structures over time and through space are greater than the differences.

APPLICATIONS

The techniques of factor analysis and RELATE may assist water resource managers in a variety of ways. For example, factor analysis may be used to develop a classification of watercourse types in a region. Such a typology may be based on characteristic factor structures resulting from differences in regularity of discharge, extent of impervious cover, geohydrologic controls, and watershed sensitivity to land use changes.

RELATE may be used to compare existing factor structures against hypothetical structures. The hypothetical or target structure would have factor loadings specified by the user. Indices of similarity or dissimilarity can be developed by comparing the desired or goal structure with the existing factor patterns. These procedures are referred to by Rummel (1970) as "target rotation analysis."

ACKNOWLEDGMENTS

The work upon which this study is based was supported in part by funds provided by the Office of Water Resources Research, U.S. Dept. of the Interior, Grant No. 14-31-0001-3147. Appreciation is expressed for the comments of Professors George W. Carey and Michael R. Greenberg (Urban Studies Division, Livingston College, Rutgers University), Prof. Leonard Zobler (Geography Department, Barnard College, Columbia University), and Mr. Niels West (Geography Department, Hunter College, City University of New York). In particular, I am indebted to Prof. Carey for his invaluable assistance in the utilization and interpretation of the factor analytic models discussed herein.

LITERATURE CITED

CAREY, G: W. 1966. The regional interpretation of Manhattan population and housing patterns through factor analysis. Geographical Review, 56:551-569, 3 tables, 6 figs.

CAREY, G. W., ZOBLER, L., GREENBERG, M. R., and HORDON, R. M. 1972. Urbanization, Water Pollution, and Public Policy Center for Urban Policy Research, Rutgers University, New Brunswick, New Jersey, xiv + 214 pp.

DAWDY, D. R., and FETH, J. H. 1967. Applications of factor analysis in study of chemistry of groundwater quality, Mojave River Valley, California. Water Resources Res., 3(2): 505-510, 3 tables, 1 fig.

DIAZ, G. SEWELL, J. I., and SHELTON, C. H. 1968. An application of principal component analysis and factor analysis in the study of water yield. Water Resources Res., 4(2):299-306, 7 tables.

EISELSTEIN, L. M. 1967. A principal component analysis of surface runoff data from a New Zealand alpine watershed. Proc. Int. Hydrol. Symp., 1:479-489, Fort Collins, Colorado, 7 tables.

ELIZABETHTOWN WATER CO. 1971. Annual report. Elizabeth, New Jersey, 17 pp.

HUGHES, J. W. 1971. Equifinality in major urban metropolitan systems: a cross-cultural factor analytic study. (Unpublished Ph.D. dissertation, Rutgers University.

KING, L. J. 1969. Statistical Analysis in Geography. Prentice-Hall, Inc., Englewood Cliffs, New Jersey. viii & 288 pp.

KNISEL, W. G., JR. 1970. A factor analysis of reservoir losses. Water Resources Res., 6(2):491-498, 4 tables, 3 figs.

MATALAS, N. C., and REIHER, B. J. 1967. Some comments on the use of factor analyses. Water Resources Res., 3(1): 213-223, 1 table.

New Jersey, Commission on Efficiency and Economy in State Government. 1967. Water resources management in New Jersey. 125 pp.

RICE, R. M. 1967. Multivariate methods useful in hydrology. Proc. Int. Hydrol. Symp., 1: 471-478, Fort Collins, Colorado, 4 figs.

RUMMEL, R. J. 1970. Applied Factor Analysis. Northwestern University Press, Evanston, Illinois, xxii & 617 pp.

VELDMAN, D. J. 1967. Fortran Programming for the Behavioral Sciences. Holt, Rinehart and Winston, New York. x & 406 pp.

WALLIS, J. R. 1965. A factor analysis of soil erosion and stream sedimentation in northern California. Unpublished Ph.D. dissertation, University of California, Berkeley.

WALLIS, J. R. 1968. Factor analysis in hydrology—an agnostic view. Water Resources Res., 4(3):521-527, 3 tables.

WEAVER, J. C. 1954. Crop-combination regions in the Middle West. Geographical Review, 44(2):175-200, 1 table, 24 figs.

WONG, S. T. 1963. A multi-variate statistical model for predicting mean annual flood in New England. Annals, Association of American Geographers, 53(3):298-311, 7 tables, 1 fig.