Statiscal Analysis on Stream Pollution
I. Factor Analysis on Stream Pollution of the Yodo River

By

Hikaru SHOJI* and Takeo YAMAMOTO*

(Received April 30, 1962)

In order to obtain the composite pollution index which may be available as the evaluation of the degree of gross stream pollution, factor analysis was carried out using monthly water examination data from 1923 to 1958 at Kunijima intake crib. Ten items i.e. turbibity, pottasium permanganate consumed, color, general bacteria count, residue by evaporation, total nitrogen, chlor ion concentration, hardness, stream flow rate and stream water temperature were adopted as variables in factor analysis. From the results of factor analysis, three definite factors i.e. pollution factor, rainfall factor and air temerature factor were identified, and correlations between these three factors and ten variables were elucidated. Computing the β weights for pollution factor, the composite pollution index was obtained.

1. Introduction

Since the "Water Quality Conservation Act of Public Water Basins" and the "Industrial Wastes Regulation Act" were promulgated by the Japanese Government in 1958, conservation measures in regard to river water quality have gained great interest in this country.

In order to establish conservation measures for rivers, it is necessary to make clear the present status and the causes of river pollution. The authors carried out a statistical analysis of the water pollution of the Yodo River using monthly data from the Kunijima intake crib water qualities from 1923 to 1958. The Kunijima intake crib water has been purified and supplied as the drinking water for the whole area of Osaka City.

Factor analysis and time series analysis were adopted as the analytical methods. In this paper, the results of factor analysis on river pollution were reported and the results of the time series analysis will be published in the second paper. Factor analysis has been developed by Spearman, Pearson, Burt, Thomson, Garnet, Holizinger, Thurstone, Harmann etc[1]., and its main applications

* Department of Sanitary Engineering

have been made public in the psychological field. So far as we know, the work reported here is the first trial designed to explain the various causes of river pollution by the present method and to obtain a composite pollution index of river water quality.

2. Analytical Method

In the present analysis, the data of the water examination by the Osaka Municipal Hygienic Laboratory and Purification Plant of the Osaka Municipal Water Works were applied. Ten items i.e. turbidity, pottasium permanganate consumed, color, general bacteria count (common agar plate count), residue by evaporation, total nitogen (albuminoid+ammonia nitrogen), chlor ion concentration, hardness, stream water temperature and stream flow rate were selected as the variables in the factor analysis. After the observed values of each variable (each testing item) were transformed to the standardized values, the correlation coefficients among the variables were calculated by means of equation (1):

$$r_{jk} = \sum_{i=1}^{N} Z_{ji} Z_{ki} / N \qquad (1)$$

where Z_{ji} and Z_{ki} are the standardized values of testing item j and k for observed individual data, and N is a total number for samples.

As a procedure for estimating communalities, Guttman's iterative approximation method[2] (convergent method) was applied. Guttman's method is as follows.

If the reduced correlation matrix after t-th repeated calculation with proper estimate of communalites is expressed in the form

$$\boldsymbol{R}_t = \boldsymbol{R}_0 + \boldsymbol{C}_t \qquad (t = 1, 2, \cdots) \qquad (2)$$

where $\boldsymbol{R}_0$ is the n-th order (in this case, $n=10$) correlation matrix whose diagonal elements are all zero, and $\boldsymbol{C}_t$ is the diagonal matrix whose elements are approximate communalities after t-th repeated calculation, and if the principal diagonal of $\boldsymbol{R}_t^{-1}$ is designated as $\boldsymbol{D}_t$, then the final desired communalities are obtained in the ensuing formula:

$$\boldsymbol{C}_{t+1} = \boldsymbol{C}_t - \epsilon \boldsymbol{D}_t^{-1} \qquad (t = 1, 2, \cdots). \qquad (3)$$

Guttman proposed to take $\epsilon = 1/2$ and $\boldsymbol{C}_1$ as the diagonal matrix of SMC (squared multiple correlation of each variable with remaining $n-1$ observed variables).

Among different types of factor solutions, principal factor solution[1] was adopted as a preliminary solution, because it has a rigorous mathematical basis and the most useful property for the application of electronic computers.

Factor pattern with theoretical communalities h_j^2 in factor analysis can be represented by

$$Z_j = a_{j1}F_1 + \cdots + a_{jm}F_m \qquad (j = 1, 2, \cdots, n) \tag{4}$$

where $F_1, \cdots, F_m$ are common fctors, $a_{j1}, \cdots, a_{jm}$ are factor coefficients (factor loadings) and Z_j is a variable of j-testing item. In equation (4) the unique factor is omitted for simplicity. The first step of the principal factor method is to obtain the first factor coefficient a_{j1} which make the total contribution of the first factor F_1 to the communality of the variable Z_j a maximum, under the conditions of

$$r_{jk} = \sum_{p=1}^{m} a_{jp}a_{kp} \qquad (j, k = 1, 2, \cdots, n) \tag{5}$$

where the total contribution to the first factor is given by

$$V_1 = \sum_{j=1}^{n} a_{j1}^2 . \tag{6}$$

The next step is to find the second factor F_2 whose total contribution to residual communality is maximum. Then, it is necessary to obtain the second factor coefficients a_{j2}, so as to satisfy the conditions above mentioned, under the restrictions of

$$_1r_{jk} = r_{jk} - a_{j1}a_{k1} \tag{7}$$

where $_1r_{jk}$ are the first-factor residuals. The total contribution to the second factor F_2 is given by

$$V_2 = \sum_{j=1}^{n} a_{j2}^2 . \tag{8}$$

Generally, under the same procedures, the computation processes are continued until the m-th factor coefficients a_{jm} are extracted in regard to m common factors. When the method of Langrange multipliers is applied to obtain the above solutions, it is found that the final results are lead to determine the first, second, $\cdots$ and m-th largest eigenvalues and associated eigenvectors of the reduced corralation matrix $\boldsymbol{R}$. Then, the factor coefficients can be obtained from the multiplication of the square root of the eigenvalues and associated eigenvectors. As the computing procedure in the determination of the eigenvalues and eigenvectors, the modified Jacobi method was applied by using a High Speed Digital Computer. The outline of the progromming of digital computers for pricncipal factor solution was introduced by Harman.[1] We continued the above mentioned computation processes until the third factor was extracted. The extractions of sequential factors were neglected, because the sum of the first four eigenvalues becomes nearly equal to the original total communality.

Subsequently, to obtain the final solution (multiple-factor solution), the orthogonal transformations of the factor axes was carried out, applying the normal

varimax method derived by Kaiser[3], to satisfy Thurstone's simple structure principle. The essence of Kaiser's normal varimax method is to find the factor coefficients so as to make the following function a maximum with suitable axis rotations,

$$V = n\sum_{p=1}^{m}\sum_{j=1}^{n}(b_{jp}/h_j)^4 - \sum_{p=1}^{m}(\sum_{j=1}^{n}b_{jp}^2/h_j^2)^2 \qquad (9)$$

where b_{jp} are final factor coefficients, h_j^2 are communalities, n is number of testing items and m is number of common factors (in this case $n=10$, $m=3$). The desired angle of axis rotation to satisfy the equation (9), was derived by Kaiser[4].

Finally, we solved the following normal equations derived from multivariate regression theory, to find the estimate of the most important factor (in this case, pollution factor)

$$\begin{aligned}
&\beta_{p1}+r_{12}\beta_{p2}+\cdots+r_{1n}\beta_{pn} = r_{z_1F_p}\\
&r_{21}\beta_{p1}+\beta_{p2}+\cdots+r_{2n}\beta_{pn} = r_{z_2F_p}\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&r_{n1}\beta_{p1}+r_{n2}\beta_{p2}+\cdots+\beta_{pn} = r_{z_nF_p}
\end{aligned} \qquad (10)$$

where $\beta_{p1}\cdots\beta_{pn}$ are unknown weighting values (frequently called β weight) to be determined and $r_{z_jF_p}=a_{jp}$ in the orthogonal axis rotation. The magnitude of any factor $\bar{F}_p$ can be expressed by

$$\bar{F}_p = \beta_{p1}Z_1+\beta_{p2}Z_2+\cdots+\beta_{pn}Z_n \qquad (p=1, 2, \cdots, m)\,. \qquad (11)$$

In our case, the evaluation of the pollution factor is considered as the composite pollution index which represents the degree of water pollution collectively.

All foregoing computing procedures were performed by using KDC-I, Kyoto University High Speed Digital Computer with suitable programmings.

3. Results

(1) Reduced Correlation Matrix

The correlation matrix calculated from observed values was tabulated in Table 1. In principal diagonal place (printed in Gothic) there entered communalities approximated from equation (2), (3). Such matrices having communalities as diagonal elements are generally described as reduced correlation matrices. In Table 1, eight testing items, excepting stream flow rate and stream water temperature are the variables which indicate the degree of water pollution. According to the values of correlation coefficients, these testing items can be classified into the following three groups. The first group which has a characteristic feature of high positive inter-correlation coefficients, consists of turbidity, pottasium permanganate consumed, general bacteria count and residue by

Table 1. Reduced correlation matrix.

Test	1	2	3	4	5	6	7	8	9	10
1	**.714**									
2	.647	**.640**								
3	.534	.350	**.383**							
4	.516	.417	.228	**.571**						
5	.426	.594	.310	.380	**.441**					
6	.324	.047	.182	.090	−.083	**.341**				
7	.167	.056	.041	.068	.051	−.150	**.139**			
8	.037	−.048	.226	.488	.129	.164	−.185	**.476**		
9	.003	.371	.022	.173	.398	−.356	−.093	.096	**.449**	
10	−.112	.262	−.087	.140	.288	−.297	−.056	.054	.534	**.366**

1. Turbidity, 2. Pottasium permanganate consumed, 3. Color, 4. General bacteria count, 5. Residue by evaporation, 6. Stream flow rate, 7. Total nitrogen, 8. Stream water temperature, 9. Chlor ion concentration, 10. Hardness.

evaporation. The second group is total nitrogen which has negative correlation with stream flow rate, stream water temperature, chlor ion concentration and hardness. The third group consits of chlor ion concentration and hardness: they have high positive correlation coefficients between each other and positive coefficients with stream water temperature and negative coefficients with stream flow rate.

(2) Eigenvalues, Coefficients of Factors and Identifications of Factors

Eigenvalues computed by Jacobi's method are illustrated in Table 2. There are five positive and five negative values. Since the sum of the first three eigenvalues (4.857) was slightly above the original total communality, only these three eigenvalues were adopted.

Table 2. Eigenvalues of principal factors.

1	2.594	6	−.070
2	1.438	7	−.107
3	.825	8	−.146
4	.291	9	−.198
5	.156	10	−.263
Sum of positive eigenvalues			5.304
Sum of negative eigenvalues			−.784
Total			4.520
Original total communality			4.520

Computing the multiplication of square root of the three eigenvalues and associated eigenvectors (i.e. factor coefficients), the results were illustrated in Table 3 and Fig. 1.

Since the coefficients of the first factor F_1 are all positive, this factor may

Table 3. First three principal factors for ten items.

Test	Pollution factor F_1	Rainfall factor F_2	Air temperature factor F_3	Communality	
				Original	Calculated
1	.755	−.427	−.234	.714	.807
2	.784	.100	−.275	.640	.700
3	.497	−.296	−.009	.383	.335
4	.655	−.084	.341	.571	.552
5	.670	.214	−.064	.441	.499
6	.090	−.571	.114	.341	.347
7	.064	−.048	−.276	.139	.083
8	.266	−.096	.688	.476	.553
9	.374	.630	.061	.449	.540
10	.245	.610	.077	.366	.438
V_p	2.592	1.439	.823	4.520	4.854

1. Turbidity, 2. Pottasium permanganate consumed, 3. Color, 4. General bacteria count, 5. Residue by evaporation, 6. Stream flow rate, 7. Total nitrogen, 8. Stream water temperature, 9. Chlor ion concentration, 10. Hardness.

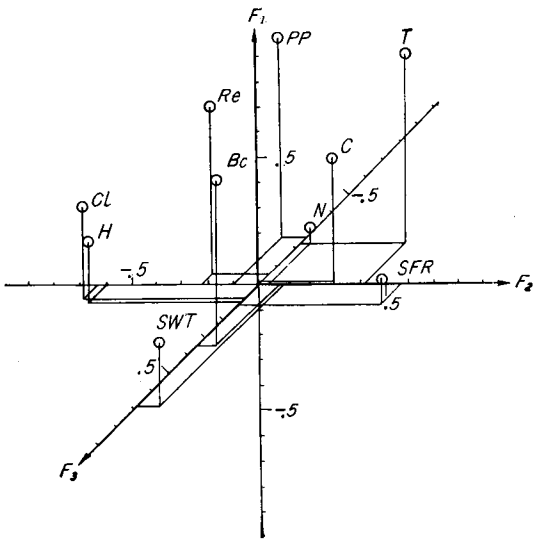

Fig. 1. Three-dimensional plot of the factor loadings of each item with each of three common factors. T. Turbidity, PP. Pottasium permanganate consumed, C. Color, Bc. General bacteria count, Re. Residue by evrporation, SFR. Stream flow rate, N. Total nitrogen, SWT. Stream water temperature, Cl. Chlor ion concentration, H. Hardness.

be regarded as a pollution factor, more specifically, an artificial pollution factor. On the other hand, the coefficients of the other two factors have positive or negative signs. As the second factor F_2 has large coefficients with stream flow rate and turbidity, respectively, so it may be regarded as a rainfall factor. In consideration of the nature of bipolar factors, this factor was shown in Fig. 1 after a conversion of the signs of all the coefficients. Similarly, as the third factor F_3 has large coefficients with stream water temperature and general bacteria count, it may be identified as an air temperature factor.

(3) Transformations of Factor Axes

In order to satisfy Thurstone's

Table 4. Varimax solution for ten items (Varimax criterion $V=40.332$)

Test	Pollution factor F_1	Rainfall factor F_2	Air temperature factor F_3
1	.872	−.188	−.109
2	.753	.325	−.164
3	.557	−.139	.072
4	.594	.107	.433
5	.582	.399	.025
6	.233	−.531	.145
7	.115	−.025	−.262
8	.179	−.020	.722
9	.164	.711	.092
10	.045	.654	.090

1. Turbidity, 2. Pottasium permanganate consumed, 3. Color, 4. General bacteria count, 5. Residue by evaporation, 6. Stream flow rate, 7. Total nitrogen, 8. Stream water temperature, 9. Chlor ion concentration, 10. Hardness.

simple structure criteria, orthogonal transformations of factor axes by Kaiser's normal varimax method were attempted. The varimax criterion V [Eq. (9)] showed fairly good convergence (within the error of three decimal places) in comparison with the proceeding varimax criterion after 3 cycle iterative calculations. The results after orthogonal axis rotations were shown in Table 4 and in Fig. 2. Again, in Fig. 2, the signs of the second factor (rainfall factor) coefficients are all converted. No marked differences are recognized between the preliminary principal factor solution and the final multiple-factor solution. Therefore, the above mentioned identifications of factors can be applied.

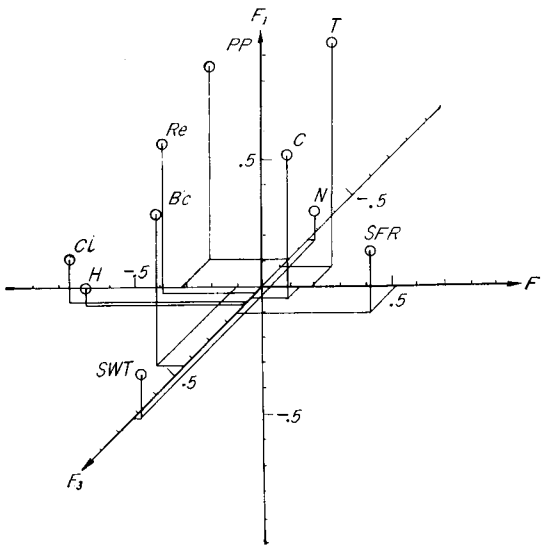

Fig. 2. Three-dimensional plot of the factor loadings of each item with each of three common factors after axis rotations by Kaiser's varimax method.
T. Turbidity, PP. Pottasium permanganate consumed, C. Color, Bc. General bacteria count, Re. Residue by evaporation, SFR. Stream flow rate, N. Total nitrogen, SWT. Stream water temperature, Cl. Chlor ion concentration, H. Hardness.

The coefficients of factors can be interpreted as the correlation coefficients

between the factor and the variables. Five variables i.e. turbidity, pottasium permanganate consumed, color, general bacteria count and residue by evaporation all have large positive correlation coefficients with the pollution factor. Stream flow rate has large positive, chlor ion concentration and hardness have large negative correlation coefficients respectively with the rainfall factor. Stream temperature and general bacteria count have large positive correlation coefficients with the air temperature factor.

(4) Composite Pollution Index

Table 5 shows the results of the computation of β weight according to equation (10). Therefore, the composite pollution index is given as follows.

Table 5. Derivation of β weight.

1.	Turbidity	.54
2.	Pottasium permangenate consumed	.26
3.	Color	.09
4.	General bacteria count	.09
5.	Residue by evaporation	.12
6.	Stream flow rate	.02
7.	Total nitrogen	.01
8.	Stream water temperature	.09
9.	Chlor ion concentration	.00
10.	Hardness	.00

Composite pollution index $=0.54Z_1+0.26Z_2+0.09Z_3+0.09Z_4+0.12Z_5+0.02Z_6+0.01Z_7+0.09Z_8$, where Z_1 is turbidity, Z_2 is pottasium permanganate consumed, Z_3 is color, Z_4 is general bacteria count, Z_5 is residue by evaporation, Z_6 is stream flow rate, Z_7 is total nitrogen, Z_8 is stream water temperature. In applying this formula, the measurement data must be converted to the standard normal distribution.

4. Discussion

A characteristic feature of factor analysis consists in describing the properties among each variable (testing items) by means of the minimum number of common factors instead of a large number of variables and to evaluate the measurement of important factors (in our study, the pollution factor) by application of multiple correlation theory. Generally, this is within the bounds of possibility, because the variables are not always independent of each other.

In the present study, we were able to identify the pollution, rainfall and air temperature factor as the three common factors, and to obtain the composite pollution index which may be available as the evaluation of the degree of gross stream pollution.

Generally, it is difficult to identify the common factors in factor analysis, and various trials have been attempted[5]. In our study, since the stream flow rate and stream water temperature, as variables, are introduced into the analysis,

successful results were obtained in the identification of three common factors with comparative ease. Mutual relations between testing items and three common factors are consistent with the general description of the significance of these testing items.

As the indicator of gross stream pollution, BOD is recommended in England and America usually[6]; on the other hand, in our country, Sugito's[7] relative purity index of water which is available for both stream and sewage water, is frequently applied in addition to BOD. However, for the reason that BOD represents the pollution by oxidizable organic matter, it does not seem suitable as the indicator of gross stream pollution. Again, in Sugito's relative purity index, it does not necessarily follow that there is solid basis on the derivation of the coefficients for each testing item.

It was impossible to compare our composite pollution index and Sugito's relative purity index on account of the different kinds and numbers of testing items, but, we consider the composite pollution index contains an excellent feature that β weight is derived in the objective basis, because Guttman's iterative procedure and Kaiser's varimax method have been applied as the estimate of communalities and axis rotations respectively.

5. Conclusion

Factor analysis was carried out, using monthly water examination data from 1923 to 1958 at the Kunijima intake crib. The results obtained were as follows.

1) As the three common factors, the pollution, rainfall and air temperature factor were identified.

2) Four items, i.e. turdidity, pottasium permanganate consumed, color, general bacteria count and residue by evaporation all have large positive correlation coefficients with the pollution factor. It is found that stream flow rate has large positive correlation coefficients with the rainfall factor, chlor ion concentration and hardness have large negative correlation coefficients with the rainfall factor respectively. Stream water temperature and general bacteria count have large positive correlation coefficients with the air temperature factor respectively.

3) The composite pollution index was obtained which may be available as the evaluation of the degree of gross stream pollution.

Acknowledgment

The authors express their sincere appreciation to the Kyoto University Computer Center for permission to use KDC-I. Thanks are also expressed to Mr. T. Nakamura and Miss W. Banno for their generous assitance in carrying out the numerical calculations.

References

1) H. H. Harman: Modern Factor Analysis, Univ. of Chicago Press, Chicago, p. 436 (1960).
2) L. Guttman: Research Report, Univ. of California, no. 12, 13 (1957).
3) H. F. Kaiser: Psych., **23**, 187 (1958).
4) H. F. Kaiser: Ed. Psych. Measurement, **19**, 413 (1959).
5) H. Maruyama and M. Momiyama: A Papers in Meteorol, and Geophys., Meteorol. Res. Inst. Japan, **2**, 311 (1951).
6) E. B. Phelps and J.B. Lackey: Stream Sanitation, 2nd Pr. John Wiley, New York, p. 276 (1947).
7) K. Sugito: J. Jap. Waterworks Sewerage Assn. no. 132, 10, no. 134, 12 (1944).