

Selection of Variables for Regional Frequency Analysis of Annual Maximum Precipitation Using Multivariate Techniques

Woosung Nam[1], Ju-Young Shin[2], Hongjoon Shin[3], Jun-Haeng Heo[4]

[1]Ph. D. Candidate, School of Civil and Environmental Engineering, Yonsei University, 134 Sinchon-Dong, Seodaemun-Gu, Seoul, 120-749, Korea

(Tel: +82-2-393-1597, Fax: +82-2-393-1597, e-mail: nws77@yonsei.ac.kr)

[2]Master Program, School of Civil and Environmental Engineering, Yonsei University, 134 Sinchon-Dong, Seodaemun-Gu, Seoul, 120-749, Korea

(Tel: +82-2-393-1597, Fax: +82-2-393-1597, e-mail: ausran@yonsei.ac.kr)

[3]Ph. D. Candidate, School of Civil and Environmental Engineering, Yonsei University, 134 Sinchon-Dong, Seodaemun-Gu, Seoul, 120-749, Korea

(Tel: +82-2-393-1597, Fax: +82-2-393-1597, e-mail: sinong@yonsei.ac.kr)

[4]Professor, School of Civil and Environmental Engineering, Yonsei University, 134 Sinchon-Dong, Seodaemun-Gu, Seoul, 120-749, Korea

(Tel: +82-2-2123-2805, Fax: +82-2-364-5300, e-mail: jhheo@yonsei.ac.kr)

Abstract

The regional frequency analysis is useful to estimate more accurate precipitation quantiles than the at-site frequency analysis, especially in case of regions with short record length like South Korea. In this study, the regionalization of annual maximum precipitation in South Korea was considered. The identification of homogeneous regions has a significant effect on quantile estimation in the regional frequency analysis. Various variables related to precipitation can be used to form regions. Since the type and number of variables may lead to improve the efficiency of partitioning, it is important to select those precipitation related variables, which represent most of the information from all candidate variables. Multivariate analysis techniques such as principal component analysis, factor analysis, and Procrustes analysis were used for this purpose. Finally, 33 variables were selected from the 42 candidate variables using multivariate techniques. A big loss of information due to dimension reduction was not found. Therefore, dimension reduction can promote the efficiency of cluster analysis. The selected variables can be successfully used to form regions for regional frequency analysis of annual maximum precipitation in South Korea.

1. INTRODUCTION

Regionalization of precipitation climates is useful for the estimation of more

accurate quantile for return periods more than 100 years. It is especially useful to South Korea with short record length. Several studies have used different type and number of variables to identify hydrologically homogeneous regions. For example, Guttman (1993) used 7 geographic and climatic variables while Mallants and Feyen (1990) did only daily precipitation data. There is no general guideline for delineating homogeneous regions.

Multivariate analysis techniques have applied to this field in recent. Zhang and Hall (2004) used some cluster analysis techniques. Dinpashoh et al. (2004) applied principal component analysis (PCA), Procrustes analysis (PA) and factor analysis (FA) to improve the effectiveness of identification of homogeneous regions.

The objective of the present study is selection of representative variables accounting for precipitation climate of South Korea by means of PCA, PA and FA.

2. MATERIALS AND METHODS

2.1 Study Area and station selection

The study area includes South Korea where the monsoon causes heavy rain mainly from June to September. The main rainfall is caused by typhoon and convection. Data from 60 stations over South Korea were gathered. There are 4 sites with record length less than 15 years and the data with at least one month of missing were excluded. Data from 39 climatological variables (Table 1) and 3 geographical ones including latitude, longitude and altitude, were obtained for 60 stations.

Table 1. Climatological variables used in the study

Variable(s) symbol(s)	Description
MAP	Mean Annual Precipitation
DayP	number of Days with Precipitation in a year
APM_i, i = 1, 2, ⋯, 12	Average Precipitation in a Month
DP_i, i = 1, 2, ⋯, 12	number of Days with Precipitation in a month
MDP_i, i = 1, 2, ⋯, 12	Average Maximum Daily Precipitation in a month
AMaxMDP	Average Maximum of Maximum Daily Precipitation in a month

2.2 Procrustes analysis

The general features of data can be expressed by principal components (PC).

$k(< p)$ of PC scores involving p original variables summarizes the data. A PC is defined as a linear combination of the original variables. Since the number of loadings of a PC is equal to the number of variables and a PC is usually interpreted on the basis of corresponding loadings, its interpretation may not be easy. Therefore, it is reasonable to reduce the number of variables to some number of $q(\geq k)$. Because the PA is an efficient method in selecting the most important variables, which preserve the configuration of whole data (Krzanowski, 1987), it is used to find the best set of q variables.

The PA is based on PCA and singular value decomposition (SVD). Figure 1 shows the schematic diagram. The standardized data matrix, $X(n \times p)$, consists of the measured values of p variables for each of n sites. Data are first standardized to take into account different scale unites. For the purpose of predicting the essential dimensionality of the data, PCA is performed initially on $X(n \times p)$. The matrix of PCS for the first k selected PCs is denoted by $Y(n \times k)$. It is possible to select any subset of q variables from p, which satisfies the two conditions: (1) $q < p$ and (2) $q \geq k$. Suppose that $\underline{X}$ denotes the $n \times q$ reduced data matrix and Z is the corresponding $n \times k$ matrix of PCS. If the true dimensionality of the data is indeed k, then Y can be viewed as the true configuration , while Z is the corresponding approximate configuration based on only the q important retained variables.

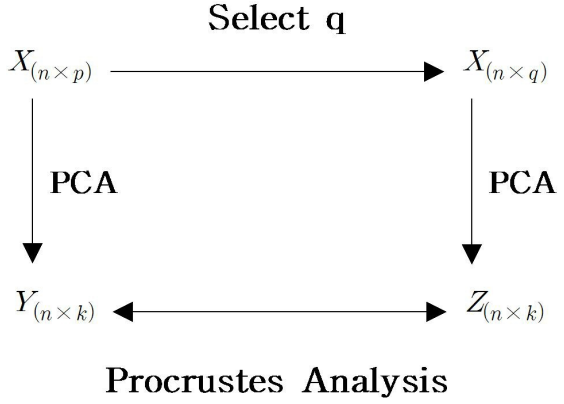

Figure 1. Variable selection by Procustes analysis

The sum of squared differences between the two configurations is as follows:

$$M^2 = Trace\{YY' + ZZ' - 2\Sigma\} \quad (1)$$

In this formula, the prime ' represents transpose and Σ is a diagonal matrix, obtained by SVD of the $Z'Y$ matrix as

$$Z'Y = U\Sigma V' \tag{2}$$

where, $UU' = I_k$, $V'V = VV' = I_k$, I is a unit matrix.

2.3 Factor analysis

FA (Factor analysis) is a statistical technique used to explain variability among observed random variables in terms of fewer unobserved random variables called factors. The observed variables are modeled as linear combinations of the factors, plus error terms. FA has advantages as follows: (1) Reduction of number of variables, by combining two or more variables into a single factor and (2) Identification of groups of inter-related variables, to see how they are related to each other.

3. RESULTS AND DISCUSSIONS

3.1 Selected variables

33 variables were selected by means of PA. 9 excluded variables were longitude, altitude, DP_3, DP_4, DP_8, DP_9, MDP_5, MDP_9 and APM_9. Eigenvalues were changed in accordance with adjustment of number of variables. Eigenvalues of 6 PCs exceeded one in using 33 variables selected by PA while those of 7 PCs did one in case of 42 variables. PC with maximum eigenvalue accounted for 32.7% of total variance before PA. The rate increased to 35.6% after PA. 6 PCs by means of PA accounted for 91.3% of total variance while 7 PCs without PA did for 89.9%. Table 2 shows eigenvalues and ratios of PCs with respect to total variance.

Table 2. Eigenvalues due to adjustment of number of variables

	42 variables				33 variables			
	eigen value	difference	ratio	accumulated ratio	eigen value	difference	ratio	accumulated ratio
1	13.74	6.17	0.327	0.327	11.75	4.83	0.356	0.356
2	7.58	0.96	0.180	0.507	6.92	1.25	0.210	0.566
3	6.62	2.23	0.158	0.665	5.67	2.51	0.172	0.738
4	4.38	1.38	0.104	0.769	3.16	1.55	0.096	0.834
5	3.00	1.63	0.07	0.839	1.61	0.58	0.049	0.883

6	1.36	0.23	0.03	0.869	1.03	0.32	0.031	0.914
7	1.13	0.38	0.03	0.899	0.71	0.36	0.022	0.936

3.2 Factor analysis

Representative factors should be selected by means of FA after selection of variables. At first, the number of factors is chosen by scree plot in usual. Scree plot shows the changes of eigenvales for each factor. In this study, 5 factors were found to be suitable by means of scree plot because 6 factors or more caused radical decrease in eigenvalues.

Then relationship between factors and variables should be considered. Factor pattern can show which variables are represented by each factor. Table 3 shows relationship between factors and variables by factor pattern. Factor 1 has large loadings for variables from late fall to early spring. Factor 2 is related to variables corresponding to spring and early summer. Factor 3 accounts for variables for rainy season. Factor 4 is mainly close to the number of days with precipitation. Factor 5 is closely related to number of days with precipitation in May and June.

Table 3. Distribution of Variables for each factor

Factors	Variables
Factor 1	APM_1, APM_2, APM_{11}, APM_{12} / MDP_1, MDP_2, MDP_3, MDP_{10}, MDP_{11}, MDP_{12}
Factor 2	APM_3, APM_4, APM_5, APM_6 / MDP_4, MDP_6 / LATITUDE
Factor 3	APM_7, APM_8 / MDP_7, MDP_8/ MAP / DP_7 / AMaxMDP
Factor 4	DayP / DP_{10}, DP_{11}, DP_{12} / DP_1, DP_2
Factor 5	DP_5, DP_6

4. CONCLUSIONS

In this study, PA was applied to select variables for cluster analysis. 42 variables were reduced to 33 ones by means of PCA and FA. Despite of decrease in number of variables, it was found that a subset of variables preserved the information of whole data. The number of PCs could be also decreased by using a subset of variables. In cluster analysis, once the sample has been partitioned into clusters, the practitioner is often interested in identifying a small number of variables which can be used to describe cluster membership. The multivariate techniques in the present study such

as PCA, PA and FA could decrease the number of variables without information loss and improve the efficiency of cluster analysis. The selected variables can be successfully used to form regions for regional frequency analysis of annual maximum precipitation in South Korea.

REFERENCES

Dinpashoh, Y., Fakheri-Fard, A., Moghaddam, M., Jahanbakhsh, S., Mirnia, M. (2004). "Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods" *Journal of Hydrology*, 297, 109-123.

Guttman, N.B. (1993). "The use of L-Moments in the determination of regional precipitation climates" *Journal of Climatology*, 6, 2309-2325.

Krzanowski, W.J. (1987). "Selection of variables to preserve multivariate data structure, using principal components" *Applied Statistics*, 36(1), 22-33.

Mallants, D., Feyen, J. (1990). "Defining homogeneous precipitation regions by means of principal component analysis" Journal of Applied Meteorology, 29, 892-901.

Zhang Jingyi, M.J. Hall (2004). "Regional flood frequency analysis for the Gan-Ming River basin in China" *Journal of Hydrology*, 296, 98-117.