Proceedings of Joint Conference of The Korean Data Analysis Society and Korean Data And Information Science Society, October 17-18, 2003, pp. 75-92

Classification of Diurnal Precipitation Patterns in South Korea using Cluster Analysis

Baek-Jo Kim, Jeong-Hoon Kim, Chun-Ho Cho and Hyo-Sang Chung[1]

1. INTRODUCTION

As the observations, transmission and archiving technology develops, amount of available digital information increases rapidly. For the dynamical methods and numerical weather models, governed by well-settled equations of physics, this "exponential" (i.e., self-enhancing, always faster than before, although no exact time-dependence is known by the authors) development can be used in a fairly straightforward way. Of course, this development requires much innovation in data assimilation, numerical model development, subgrid-scale parameterization, etc.

In statistical approaches, however, we can not directly employ additional laws. The model of the whole unknown complexity is determined by the intuition of the researcher and the statistical recipes recommended by the contemporary computing technology. Both engines of development are limited, however, by the fact that there is one dimension, *the time*, where the increase of information, in most cases is only linear. That is, the more frequent sampling in time does not yield independent data, due to the diurnal cycle and the comparable lifetime of relevant atmospheric phenomena.

The present paper recommends a classification of diurnal precipitation patterns represented by 59 stations of South Korea. The final 15-25 classes transforming the 59 continuous variables into one discrete number are defined by cluster analysis after a preliminary factor analysis.

Precipitation has been selected to demonstrate the following methodology as this variable is mainly related to meso-scale atmospheric phenomena and influenced by physical processes of even smaller scales, including microphysics of cloud droplets and crystals. Hence, deterministic computation of this atmospheric variable is rather limited compared to the requirements of medium-range weather forecasts of especially, climate change scenarios. On the other hand, this important atmospheric variable inter-relates with many environmental factors and has large

[1] Meteorological Research Institute, Korea Meteorological Administration, 460-18, Shindaebang-dong, Dongjak-gu, Seoul 156-720, KOREA.

impact on several sectors of economy. For these two reasons, precipitation is used in statistical research and applications quite frequently.

Of course, neither the factor (principal component) nor the cluster analysis is new tool even in precipitation climatology of South Korea. Several studies have already been performed (e.g. Seo and Joung, 1982; Ho and Kang, 1988; Moon, 1990; Park and Lee, 1993; Lee and Park, 1999). Our study differs from those in its scope, namely the above studies are aimed at objective classification of the stations (sub-regions), whereas this study is focused on determinating of different types of diurnal precipitation pattern. In other words, this study classifies the days of the available archive.

2. DATA

Daily precipitation data observed at 59 stations of South Korea are used, including Cheju Island. The Ullung Island as a single station is omitted in considering the results of factor analysis.

The precipitation classification is performed for 24 years between 1973 and 1996. This period (8760 days, as the 6 leap-days were excluded) is separated into shorter sub-samples to reduce the inhomogeneity caused by the annual course of precipitation (*Figure 1*). More specifically, area averages of the 59 stations are considered including all days, i.e. 24 values on each calendar day and also on the wet days, i.e. on days when the observed precipitation is at least 0.1 mm at the given station. Besides the averages, we also calculated the area mean value of point-wise standard deviation on the wet days.

Although three curves exhibit strong inter-diurnal variability, there is a well distinguished period in the middle of the year when all the curves are in maximum. This is the well-known summer period, connected with bi-directional march of the monsoon ("Changma") front over the Korean Peninsula and with the smaller scale convective activity including the episodic typhoons. This 3-monthly period is selected to be one season, whereas its winter opposite, exhibiting the minimum average and conditional standard deviation, can also be clearly delimited. So, we selected four 3-monthly periods, but one should note that, according to Fig 1., each season starts by 15 days later than the general ones. In the followings, the seasons, called winter, spring, summer and autumn will always mean the sequence of the year between December 16 ~ March 15, March 16 ~ June 15, June 16 ~September 15 and September 16 ~ December 15.

3. METHODS

Classification of diurnal precipitation patterns is performed by cluster analysis (Section 3.2) applied not for the point-wise observations, but for the series of rotated loadings, computed by factor analysis (Section 3.1).

3.1 Factor analysis

Factor analysis is generally used for reduction of data sets, represented in a large number of stations or grid points, still keeping the essence of their common variability, by resolving the initial variables into much fewer non-correlated ones. The monograph by von Storch and Zwiers (1999) gives a comprehensive theoretical overview. Factor analysis can also be used for immediate pattern classification (i.e. Bartzokas and Metaxas, 1993), but this requires larger number of spatially distributed data (stations or grid-points) than cases (days) to be classified.

Each original variable, P_i, i=1, 2, ..., n, can be expressed as $P_i = a_{i1}F_1 + a_{i2}F_2 + a_{i3}F_3 + \ldots + a_{im}F_m$ (m<n), where F_j, j=1,2, ..., m, common for each i, are *the factors* and a_{ij} *the loadings* (scores). First, it is necessary to specify whether the factor analysis is performed on a correlation matrix or a covariance matrix. We selected the covariance matrix to avoid the non-natural transformation related to standard deviation of precipitation at each station. In our view, this is also an immanent feature of the spatial pattern which is meaningful to keep before the analysis. This selection also determined that we perform a principal components analysis, as a specific realization of factor analysis.

An important question is the number of the factors (m) to retain. On this matter, many criteria have been proposed. It is noted that Jolliffe (1993) states "...different objectives for an analysis may lead to different rules being appropriate". In this study, *the Rule 1* or *Guttman criterion* is used, which determines to keep the factors with eigen-values be more than 1 and neglect the ones that do not account for at least the variance of one standardized variable.

Another vital stage is whether, or not, we should rotate the axes (factors). This rocess achieves discrimination among the loadings, making the rotated axes easier to interpret. In this analysis the *Orthogonal Varimax Rotation* is applied, which keeps the factors non-correlated. After rotation, the original precipitation P_i at station, *i*, is $P_i = a'_{i1}Fr_1 + a'_{i2}Fr_2 + a'_{i3}Fr_3 + \ldots\ldots,\ a'_{im}Fr_m$(m<n), where the a'_{i1}, a'_{i2}, ... , a'_{im} *rotated loadings are later used for classification.* The loadings are specified by the regression method to ensure the best fit of the initial data at each station.

An important feature of the non-rotated a_{ij} loadings is that they represent a decreasing order of variance as j increases. This variance is equal to the eigen-values of the analyzed covariance matrix. Moreover, the factors is the most effective set of orthogonal functions in the sense, that they "explain" the highest portion of variance retaining any fixed number of m. After rotation we loose this optimum feature and the variance is distributed more evenly among the retained loadings. Nevertheless, increasing sequence of loadings mean decreasing importance in explaining the variance, which is a key feature, used in cluster analysis with special emphasis. Application of rotation is not compulsory, but in our case it is also explained by the strongly skewed distribution of the loadings of first (non-rotated) factors (see Section 4.1).

Factor analysis was already used for annual precipitation of Korea by Moon (1990) to classify the territory of the country. Similar interpretation is briefly exposed for the unified annual sample in Section 4.1, too. The point of this application is the mapping of the maximum a'_{ij} loading at the station, i, which selects that Fr_j rotated factor for which the loading is >0.7. (According to the general experience, this threshold designates maximum one region to belong to.)

Another interpretation of the factor analysis is related to the explained communalities, i.e. that part of the initial variance which can be "explained"by linear combination of the retained factors and the loadings. If this mean square difference between the original and the estimated values is low for a given station, than this station exhibits large individual variations not correlated to those at the other stations. This interpretation is also illustrated in Section 4.1.

3.2 Cluster analysis

Cluster analysis produces hierarchical clusters of cases based on distance measures of dissimilarity or similarity (e.g. Anderberg, 1973). This method is employed to classify the diurnal precipitation patterns, represented by series of rotated factor loadings. The latter ones exhibit larger variance at the lower serial number of factors, proportionally to the explained variance, i.e. stronger difference generally occurs in most important components. Besides the 5-8 factors, explaining 81-89 % of the total variance (Section 4.1), the analysis is also performed for that number of loadings, which commonly explain 95 % of the initial variance.

Since we do not *a priori* know, how many clusters to define, hierarchical joining is applied. This requires preliminary definition of distance measures for any pair of cases and specification of the algorithm, which unequivocally selects the two clusters (or single cases) to be unified at a given stage of amalgamation. The latter is based on a distance index to be minimized, which characterizes common dissimilarity.

Having tried several possible versions, in majority of the classifications we use Euclidean distance measure, i.e. the square root of the sum of the squared differences between the components of each case. Unification of the clusters follows the Ward's method (Ward, 1963), which minimizes the sum of all within-cluster variances.

For discussion purposes, pattern correlation between two diurnal sets of loadings is also applied, which sorts the days with similar spatial distribution into one group, even if the amounts are different (Section 4.2). This approach is accompanied by method of Furthest Neighbors, which identifies the distance of two classes by the maximum possible distance between the corresponding pairs of points. The latter method will sometimes be referred as relative classification to demonstrate the importance of the area mean amount of precipitation in the resulted clusters.

The most difficult step of cluster analysis is the decision about the number of clusters to retain and interpret as the final solution. This decision should consider the retained number of classes and the efficiency of the classification. Both conditions can be considered by analyzing the agglomeration schedule, which shows the order and distances at which cases and clusters are combined into a new cluster. Fast orientation is generally supported by qualitative representation of agglomeration in form of horizontal or vertical icicle plots and dendrograms, but in our analysis it was not feasible, due to the large number of cases (1402~ 1952 days in the different seasons).

If the function of distance on the number of cases exhibit sudden changes (breaks), than the clustering can be naturally terminated before such an increasing jump. One should note however, that this distance is related to the smoothed factorial representation of precipitation field, not to the total variance of the original patterns. Hence, a more established solution should be based on computation of the quality of representation for the original fields. The explained variance of the clustering, EV(k), depending on the number of clusters, k, is a key characteristic for this, defined as

$$\mathrm{EV(k)} = \sum_{i=1}^{k} \frac{\sum_{j=1}^{N_i} \frac{1}{58} \sum_{s=1}^{59} (P_{ijs} - \langle P_{is} \rangle)^2}{N_i - 1}$$

where $\langle P_{is} \rangle$ *is the cluster-mean precipitation at station, s, derived from all* P_{ijs}values of the N_i days, that belong to the i-th cluster. For better interpretation, this explained variance is compared to the k = 1 version, EV(1), and the RV(k) = EV(k)/EV(1) relative variance is expressed in %. The lower this ratio, the more effective the clustering.

Since in our case both quality indices behaved rather smoothly, not allowing an optimum selection by this criterion, another point of view, i.e. the minimum number of cases in the smallest cluster, was also considered. Sothe selection of the final clustering is performed in three steps:

1.) Candidates for termination were selected according to the N_{min} size of the smallest cluster, i.e. $N_{min} > 1$, $N_{min} \geq 5$, $N_{min} \geq 1\%$ and the last stage before $N_{min} \geq 5\%$.

2.) The $RV(k)$ relative variances were determined for the candidates. The second one was selected, as it was just slightly worse than the first one and they represented the a priori expectations about the seasonal differences: i.e. the most clusters occurred in summer and no big differences took place among the three other seasons.

3.) This selection was finally corrected by decreasing the number of clusters by one in spring and summer, which gave further slight improvement in the explained variance.

The final clustering was also characterized by a modified formula, taking the differences of the cluster-mean precipitation into consideration by inversely weighting the squared deviations, as

$$WV(k) = \sum_{i=1}^{k} \frac{\sum_{j=1}^{N_i} \frac{1}{58} \sum_{s=1}^{59} (P_{ijs} - \langle P_{is} \rangle)^2 / \langle P_{is} \rangle}{N_i - 1}$$

This weighted variance is also standardized by the $WV(1)$, no-clustering reference value. The idea of this alternative index is to consider if the variance is dominant in the high or low-precipi-tation clusters. Of course this parameter can not be applied for the dry (no precipitation) cluster.

4. RESULTS

4.1 Factor scores for further analysis

The loadings, resulted by factor analysis are just input variables for the pattern classification. The factors and the seasonal loadings are analyzed in a separate study (Mika et al, 2001). Here we focus on four aspects of these computations, evaluating:

- How detailed is the representation of the original patterns by the applied 5 8 loadings?
- Can factor analysis or rotation change the strongly skewed distribution of precipitation?
- How can we interpret the geometry of the rotated loadings, applied for regionalization?
- How can factor analysis be used to decide about inclusion of Cheju and Ullung Islands?

The retained factors and the explained variances are displayed in *Table 1* to illustrate the answer to the first question. The retained 5 (winter), 6 (spring and autumn) or 8 (summer) factors explain the 81 89 % of variance. To explain 95 % of that, we would need much more, 14 30 factors to retain. These seasonal differences correspond to the conditional means, also indicated in the table. Considering the wet days only, the area mean precipitation is almost six times larger in summer (8.56 mm/day) than in winter (1. 52 mm).

Statistical distribution of the area average is rather close to the exponential one, as indicated in *Fig. 2a* without seasonal separation. Overwhelming majority of the wet days represent low precipitation with a steep and monotonous decrease of frequency at higher amounts. Not surprisingly, distribution of the first non-rotated factor loadings (*Fig. 2b*) is rather similar, indicating that this major component, representing 39 64 % of the original variance in the different seasons, are strongly related to the area mean precipitation. Rotation of the factors *(Fig 2c)* can not change the situation too much either. The only difference is that here the kurtosis of the distribution is much larger than that of the normal distribution. The point of the matter, i.e. the strong dominance of the low precipitation averages, remains valid for the rotated components, too. On the other hand, the rotation yields more symmetric distribution of loadings and more even distribution of the variance among the chief retained factors, what increases the freedom of clustering.

Seasonal rotated scores are input variables of the cluster analyses, also representing variously distributing fields (Mika et al., 2001). Here we present another illustration, based on rotated loadings of the all-year factor analysis of the wet days (6268 days in the sample). These factors can also be interpreted as sub-regions it which precipitation variations are similar to each other and, at the same time, relatively different from those in the other regions. *Fig. 3a* indicates the results of this analysis for data of the whole year in 60 stations including Ullung Island. This non-seasonal analysis separates 7 regions with reasonable spatial distribution. The regions are determined by the maxima of the seven rotated loadings at the given station. The majority of stations exhibits nearly 0.7 loading and can be related to one of the seven regions, even if they do not fall into the core of the regions, delimited by the 0.8 loading isolines.

Answer to the fourth question is found in the explained communality, which orders a number to every stations reflecting the proportion of variance statistically explained by linear combination of the retained factors and loadings. If the communality is not close to 1, we establish high proportion of individual variance. For this reason, Ullung Island will be excluded from the country-wide classification of the patterns (*Fig 3b*), since only 38 % of its variance is related to the common information represented by the factors. Cheju Island is worth keeping since the communality is equal or higher here than in the internal part of the peninsula, which might be due to the three stations located on that island.

4.2 Alternative results of cluster analysis

As already mentioned in Section 3.2, function of the between-cluster distance on the decreasing number of clusters does not yield any break, but it represents a smoothly increasing dissimilarity. *Figure 5* indicates the behavior of the distances for the last 30 steps in autumn. In its upper module, "Rule 1"indicates the 6 rotated loadings and "95 %"reflects the case of 17 loadings, as bases of the clustering (compare with Table 1). Naturally, the more detailed representation of the diurnal patterns is somewhat more difficult to compress into a fixed number of classes, which is valid also in case of the overall distance (one joint class, i.e. no clustering), too.

In case of the relative classification (pattern correlation, Furthest Neighbors, see lower panel of Fig. 4), however, there are some brakes in the index, which allow to select a natural termination of clustering. This implies that most likely the continuum of the area-mean precipitation is the main reason of the smooth behavior in the Euclidean distance-based way of clustering.

According to the methodology, described in Section 3.2, number of clusters and the relative variances of the four candidates and the final selection are comprehended in *Table 2.* Note that they are related to the wet clusters and cases, without inclusion of the dry days (i.e. cases and clusters characterized by no measurable precipitation at any of the 59 stations). Figures of the Table can be interpreted, as follows:

i) Number of clusters when not any single, non-clustered case remains is rather large and it were difficult to interpret why we had more clusters in spring (29) than in summer (25), also, much more in the transition seasons than in winter. Hence, this candidate should be rejected.

ii) Number and seasonal distribution of the $N_{\min} \geq 5$ candidate is fairly reasonable and the relative variances are also convincing. Hence, this candidate is worth considering, although the number of clusters is still high.

iii) The $N_{\min} \geq 1\%$ candidate for clustering is already characterized by convenient numbers of clusters, but the relative variances are not attractive. Especially the summer and autumn values are too big, larger than 50 %.

iv) As concerns the last stage before $N_{\min} \geq 5$ %, its number of wet clusters is very practical (3-6 clusters), but the unexplained variance is even higher, than in the previous case.

Hence, although there might be applications, especially, if related to short samples, when the low number of classes are more important than the reduction of variance, it is difficult to recommend a classification for general use, that leaves higher portion of variance unresolved, than explained.

So, there is only one reasonable candidate, the $N_{\min} \geq 5$ one, which can be further polished to some extent. It is performed by a systematic search for lower numbers of clusters, where the explained variance is equal or higher. One should note, again, that it is not principally excluded, since the monotonous increase of the distance function parallel to decreasing number of clusters is strictly related only to the reduced number of loadings, not to the original patterns. But, as expected, we found only two possibilities to improve the pre-selected candidates, since the loadings provided fair representation of the patterns. In both cases the number of clusters decreased by one and the efficiency could even be improved by one percent.

4.3 The final classification

The final classification exhibits 14+1; 15+1; 24+1 and 15+1 clusters, in the above-defined winter, spring, summer and autumn periods, respectively. (The+1 cluster indicates the totally dry days with no measurable precipitation at any station.) The cases (days) are distributed very unevenly among the classes, as demonstrated in *Figure 5*. (This figure shows the wet days only.) Having the clusters sorted according to their area-mean precipitation in an increasing order, the frequency of clusters exhibits nearly the opposite distribution. The most frequent wet clusters are characterized by very low amounts of are-mean precipitation in each season. Together with the dry days, the two clusters represent 68, 60, 41 and 66 % of all cases of the seasons starting from winter to autumn, in the above given sequence. In another comparison, the dry days represent 28.4 % of the 24 years, whereas the largest wet clusters cover 30.2 % of that.

In each season there are some clusters with relatively large area-mean precipitation, but their frequency is not high, except for the summer season. The importance of cluster analysis is demonstrated by these high-precipitation clusters: Despite their low general frequency, claiming for equalization by amalgamation into one cluster, strong internal differences of the patterns belonging to one or the other cluster do not allow this unification. In many relations the pattern correlation between the cluster centers is strongly negative. Hence, they should be kept separately.

There is no place to present all the 67 wet clusters, hence we limit ourselves to illustrate the big differences among the clusters. Each season is represented by 4 clusters in *Figure 6*, according to the following selection: Besides the clusters with the two lowest and the two highest area-mean cluster centers, two other pairs are selected from the middle of the distributions, characterized by strongly negative pattern-correlation, but similar area mean values. The four corresponding pairs are positioned beneath each other with an increasing order of cluster centers, from the left to the right. So, differences of the clusters are recommended to consider mainly betweenthe upper and lower figures, within a given season. The cluster-centers exhibit fairly small-scale patterns of precipitation maxima, which might, however, be related to one or few extreme events, especially in case of small clusters. (See the number of clusters in the headings).

Main statistical characteristics of the final classification are presented in *Table 3*, incorporating the dry days, as well. Comparing frequencies of the dry cluster to the most frequent wet cluster, the latter ones exhibit higherpercentages in winter and summer, but they appear more rarely in the transition seasons. Size of the smallest clusters is always between 5 and 9 members.

The relative variance, explained by the clustering becomes slightly better if we include the dry days into the analysis. Average performance of the classification is as good as 37 %, with better capability in winter and spring (31 and 34 %), but weaker than average in summer and autumn (41and 40 %). In other words this means that the classification is able to explain the complementary part of variance, i.e. 63 % in average (59-69 % in seasonal extremes).

Speaking more practically, if having been informed about the prevailing cluster in a given day, one substitutes the actual precipitation pattern by the cluster centers, the average squared error of estimation is only 37 (3141) % of the initial uncertainty, determined just by the knowledge of climatic mean patterns. Relying at a distant analogy, the case of linear regression, in that case similar reduction of uncertainty is achieved by 0.79 (0.770.83) correlation coefficients.

Applying weighted relative variances, i.e. dividing the squared deviations by the cluster centers (see Section 3.2), we obtain even better figures. This weighted average uncertainty is only 22 % of the unclassified one with a variation between 17 % (autumn) and 27 % (summer).

The last paragraphs demonstrate fairly encouraging numbers, derived from point-wise validation of clustering, which represent spatially averaged gain of information in Korea. To demonstrate information about the spatial variability of performance, we recommend the Figure 7, representing frequency distribution of the remained relative (non-weighted) variance among the 59 stations. The distributions are positively skewed in summer and autumn, i.e. there exist a few stations with poorly explained precipitation, but the majority of them belongs to even lower non-explained variance, than the average numbers of Table 3. The two other seasons are more symmetrical and even the worst stations exhibit slightly above 50 % of non-explained variance.

The standard deviation of the relative variance is only 7-8 % around the mean. So, one can conclude that the cluster analysis, represented by fairly low average percentages of non-explained variances, can also be applied for a small sub-set of stations (including individual ones) with substantial reduction of variance, even in the worst possible cases.

Acknowledgements : This paper was performed for the Korea Enhanced Observing Period (KEOP) Project, one of the Principal Projects of Meteorological Research Institute (METRI), KMA.

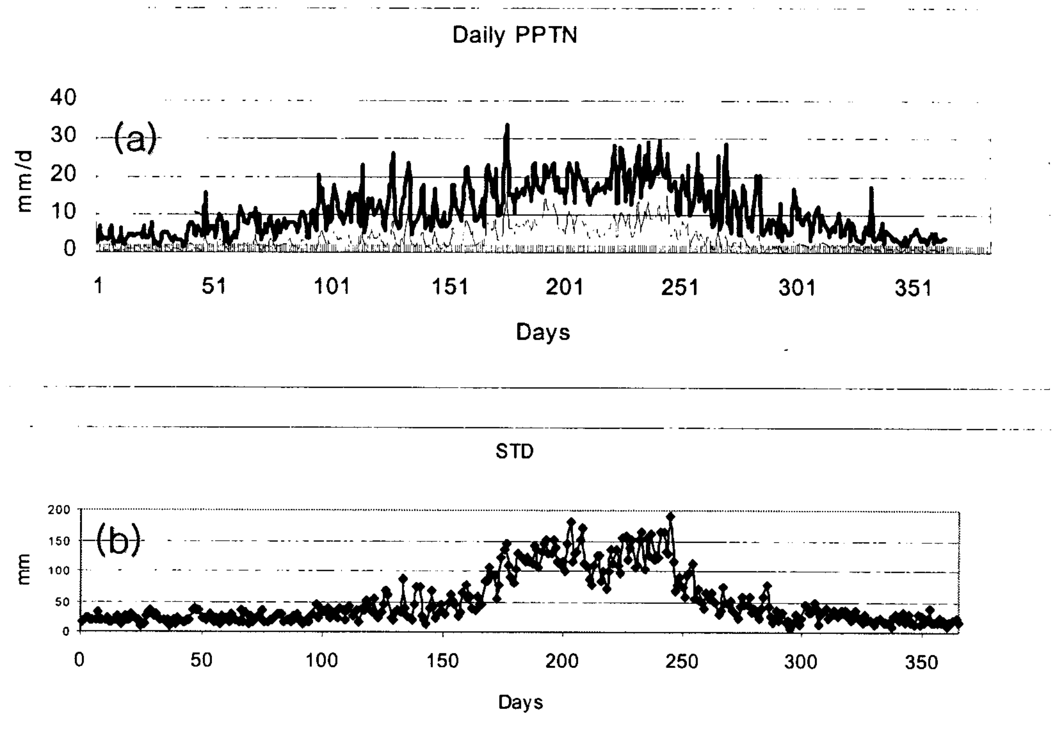


Figure 1. Annual cycle of precipitation in 1973-1996: (a) averages for all days (thin solid line) and wet (>0.1 mm) days(thick solid line), (b) standard deviation on wet days of a station (59 stations, Ullung Island is already excluded). The defined seasons start on Julian day 349 (December 16) in winter, Julian day 75 (March 16) in spring, Julian day 167 (June 16) in summer, and Julian day 258 (September 16) in autumn.

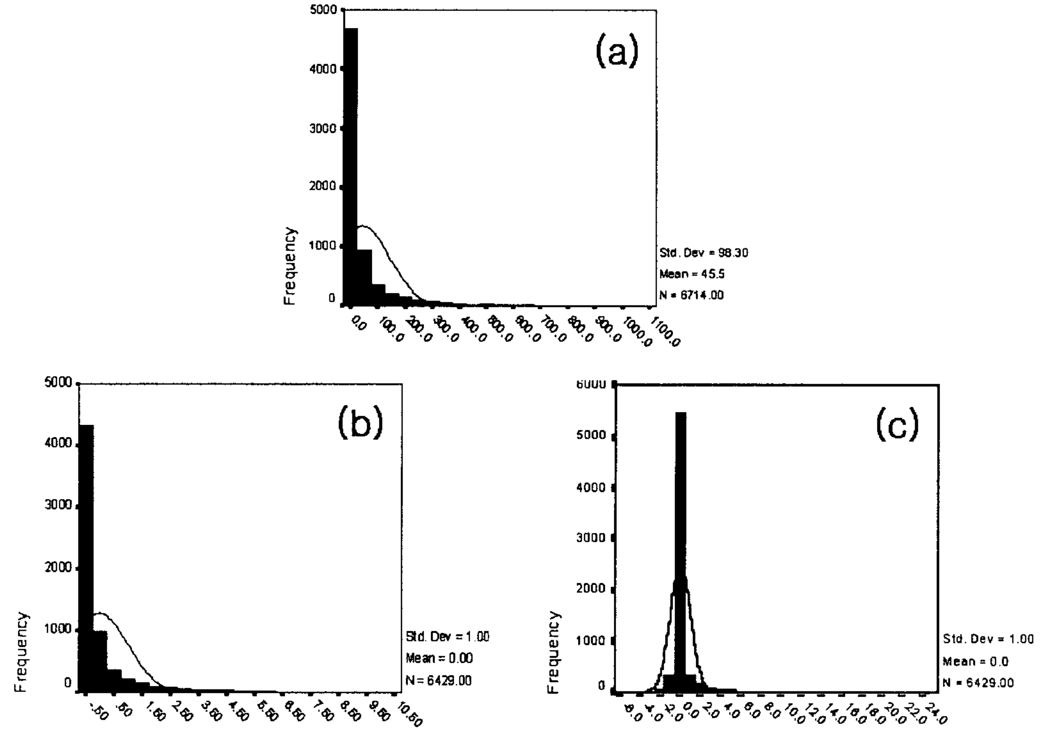


Figure 2. Distribution histograms of (a) area-mean diurnal precipitation at the 59 stations in 1973-1996 in South Korea, (b) the first diurnal non-rotated factor loadings and (c) the rotated factor loadings. Days being dry at all stations are excluded. Normal distribution is graphically fitted.

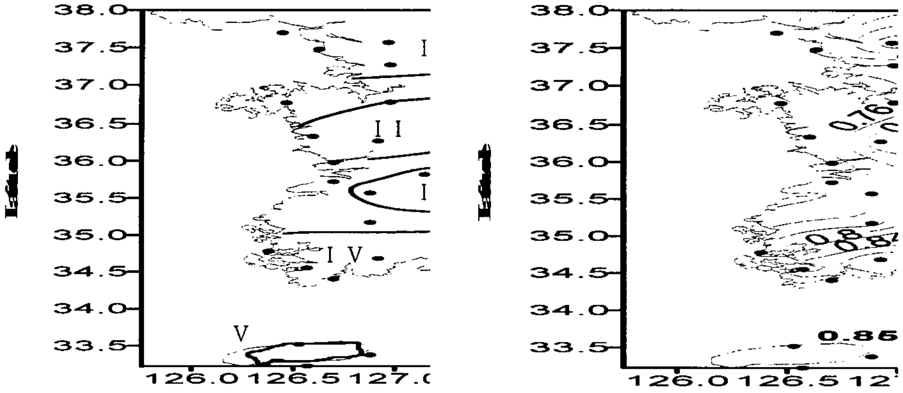


Figure 3. Distributions of (a) seven regions of precipitation from diurnal factor analysis with no separated seasons, (b) communality represented by the retained factors. Dry days are included

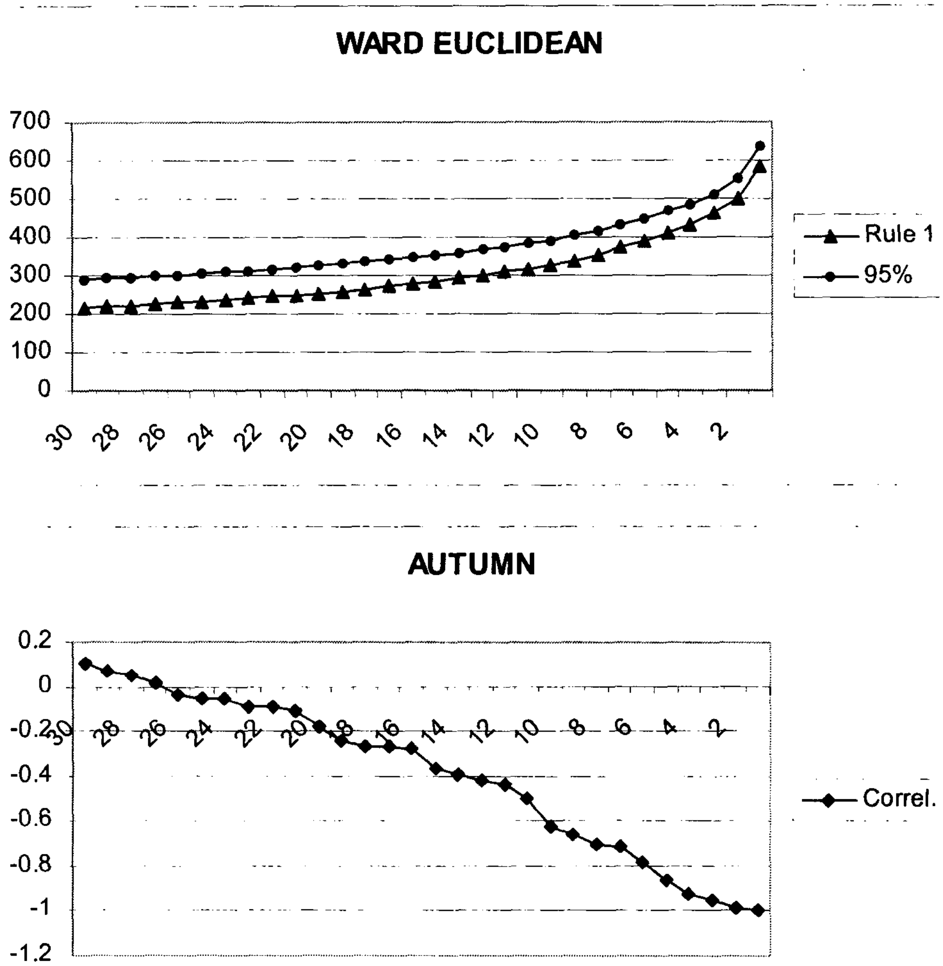


Figure 4. Curves to support clustering termination: a) Ward method, Euclidean distance smooth increase of the distance index, based on the five factors (according to Rule 1) and 17 factors (explaining 95 % of variance). b) relative classification (pattern correlation and method of Furthest Neighbors) with some breaks in the distance index (i.e. lowest correlation).

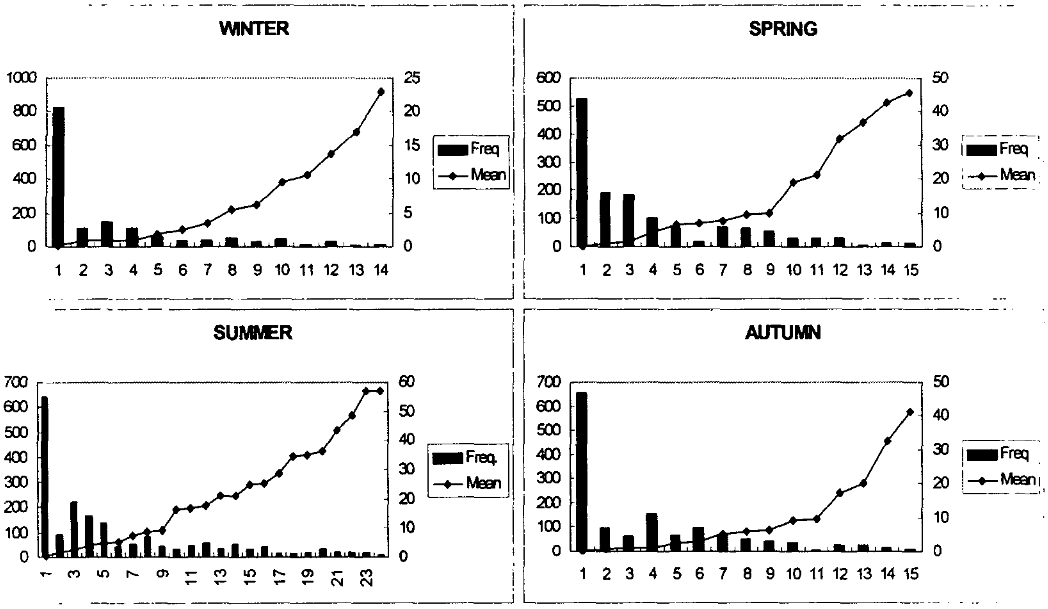
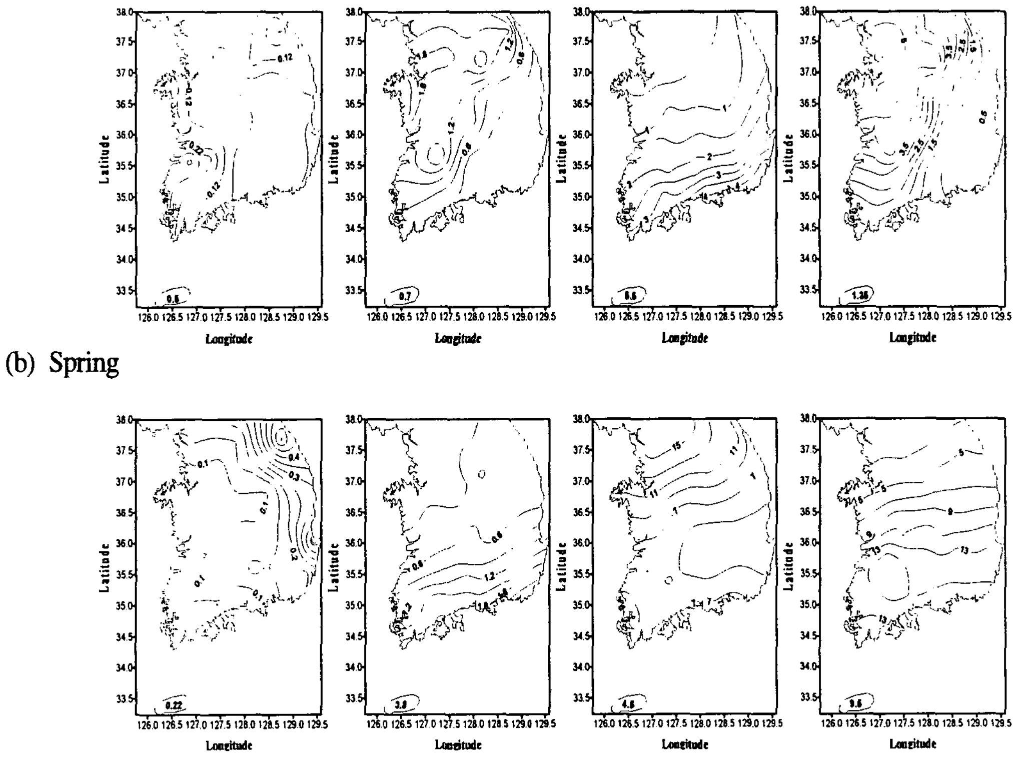
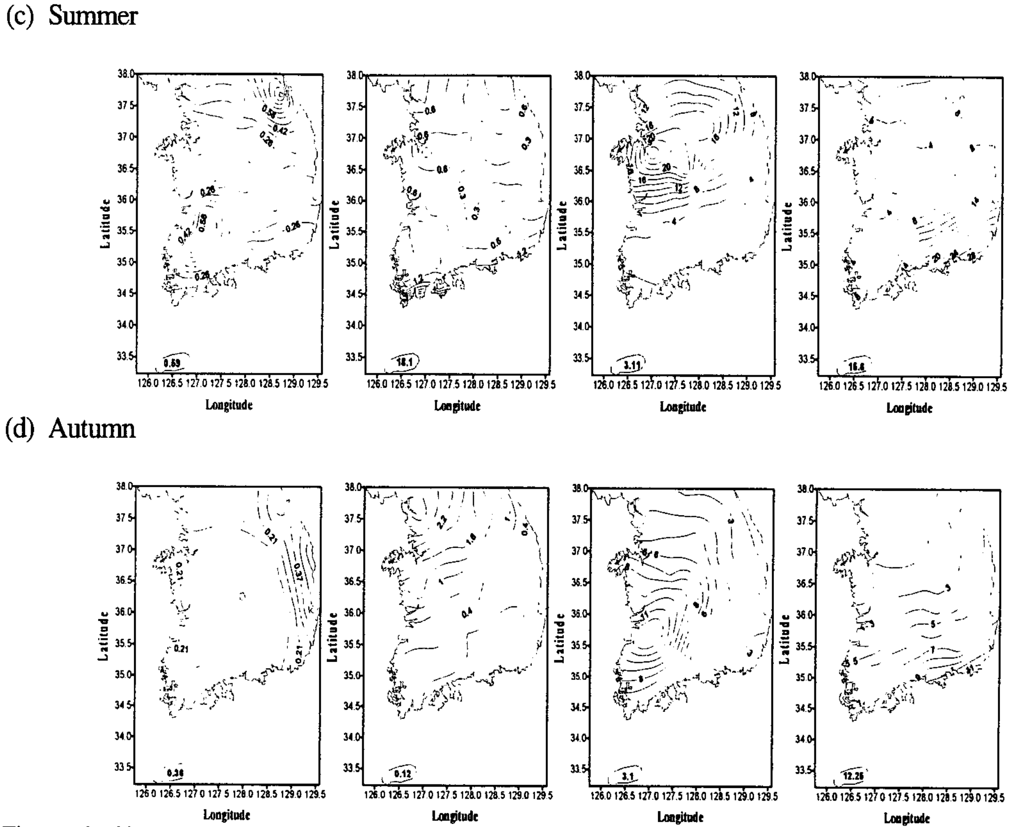


Figure 5. Frequency and area-mean precipitation of the clusters with non-zero precipitation

(a) Winter



(b) Spring


(c) Summer

(d) Autumn

Figure 6. Cluster centers of eight selected maps from each season a) winter, b) spring, c) summer and d) autumn. The eight maps of each season represent clusters with the lowest and highest area averages and also pairs with similar, intermediate averages, but strong negative correlation.

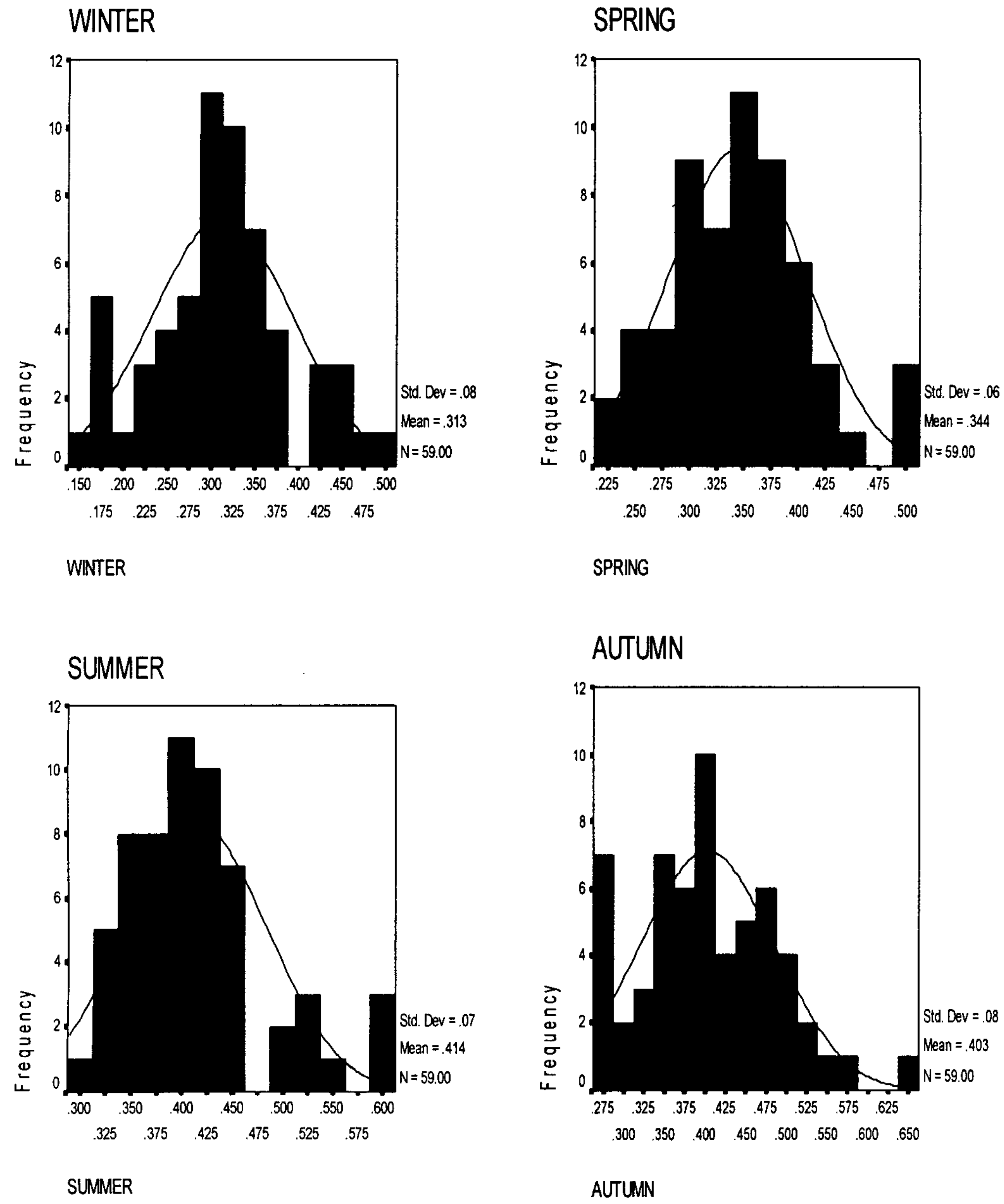


Figure 7. Frequency distribution of the non-explained relative variance among the 59 stations. (Normal distribution is graphically fitted.)

Table 1. Retained original and rotated % eigenvalues of seasonal factor analysis

	Winter		Spring		Summer		Autumn	
Sample size	1508 days		1402 days		1952 days		1406days	
Area mean	1.52 mm		4.74 mm		8.56 mm		2.51mm	
Explained variance	Original	Rotated	Original	Rotated	Original	Rotated	Original	Rotated
1. factor	64	41	67	29	39	15	59	25
2. factor	12	22	9	22	17	15	10	20
3. factor	5	12	5	19	7	14	7	16
4. factor	4	7	4	9	6	14	5	12
5. factor	2	5	2	5	5	11	3	11
6. factor	·	·	2	4	9	4	2	2
7. factor	·	·	·	·	2	6	·	·
8. factor	·	·	·	·	2	2	·	·
$\sum$ indicated	87.3		89.0		81.0		86.4	
95 % expl.	14		15		30		17	

Table 2. Alternative of the final classification with the mean relative variance expressed in the non-dry (N-1 cluster) days.

Cluster member	> 1 member		≥ 5 members		Final selection		≥1%		The largest before≥ 5%	
	Wet	Expl.	Wet	Expl.	Wet	Expl.	Wet	Expl.	Wet	Expl.
	Clsuters	Var. %	Clsuters	Var. %	Clsuters	Var. %	Clsuters	Var. %	Clsuters	Var. %
Winter	14	·	14	32	14	32	9	39	4	51
Spring	29	·	16	37	15	36	10	43	5	48
Summer	25	·	25	43	24	42	12	54	6	63
Autumn	24	·	15	42	15	42	7	57	3	64

Table 3. Main statistical characteristics of the classification. Normalized variance means the proportion of variance to the average, both within a cluster and the whole sample.

	All days of season	Wet days	Dry days	No. of wet days in cluster	Mean variance(%)	Normalized variance(%)
Winter (15 clusters)	2160	1508	652	9-825	31	20
Spring (16 clusters)	2208	1402	806	5-524	34	25
Summer (25 clusters)	2208	1952	256	8-640	41	27
Autumn (16 clusters)	2184	1406	778	5-653	40	17
			Average		37	22

References

[1] Anderberg, M. R. (1973). *Clusteranalysis for Applications*. Academic Press, New York.

[2] Bartzokas A. and Metaxas D. A. (1993). *Covariability and climatic changes of the lower troposphere temperatures over the Northern Hemisphere*. Il Nouvo Cimen., 16C, 359-373.

[3] Hair, J. F., Anderson R. E., Tatham, R. L. and Black, W. C. (1998). *Multivariate data analysis*. (Fifth Ed.) Prentice Hall, New Jersey, p. 730.

[4] Ho Ch.-H. and Kang, I.-S. (1988). The variability of precipitation in Korea. *J. Korean Meteor. Soc.* 24, 38-48 (in Korean).

[5] Lee D.-K. and Park, J.-G. (1999). Regionalization of summer rainfall in South Korea using cluster analysis. *J. Korean Meteor. Soc.* 35, N4, 511-518 (in Korean).

[6] Jolliffe, I. T. (1993). *Principal Component Analysis*: A beginner's guide - II Pitfalls, myths and extensions. Weather, 48, 246-253.

[7] Moon Y.S. (1990). Division of precipitation regions in Korea through cluster analysis. *J. Korean Meteor. Soc.* 26, 203-215.

[8] Park and Lee (1993). A regionalization of annual precipitation over South Korea. *J. Korean Meteor. Soc.* 29, 117-125. (in Korean).

[9] Seo A.-S., and Joung, Ch.-H. (1982). The climatic factor analysis of precipitation temperature and sea-level pressure over Korea using empirical orthogonal functions. *J. Korean Meteor. Soc.* 26, 40-50.

[10] von Storch, H. and Zwiers, F. W. (1999). *Statistical analysis in climate research*. Cambridge Univ. Press, Cambridge, UK, p. 484.

[11] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association*, 58, p. 236.