water

MDPI

Article

Regionalization of Drought across South Korea Using Multivariate Methods

Muhammad Azam [1,*], Hyung Keun Park [2,*], Seung Jin Maeng [1] and Hyung San Kim [3]

[1] Department of Agricultural and Rural Engineering, Chungbuk National University, Cheongju 28644, Korea; maeng@chungbuk.ac.kr

[2] Department of Civil Engineering, Chungbuk National University, Cheongju 28644, Korea

[3] K-Water Research Institute, Daejeon 34045, Korea; kamakim486@naver.com

* Correspondence: azam@chungbuk.ac.kr (M.A.); parkhk@chungbuk.ac.kr (H.K.P.); Tel.: +82-043-261-3587 (M.A. & H.K.P.)

Received: 24 November 2017; Accepted: 27 December 2017; Published: 29 December 2017

Abstract: Topographic and hydro-climatic features of South Korea are highly heterogeneous and able to influence the drought phenomena in the region. The complex topographical and hydro-climatic features of South Korea need a statistically accurate method to find homogeneous regions. Regionalization of drought in a bivariate framework has scarcely been applied in South Korea before. Hierarchical Classification on Principal Components (HCPC) algorithm together with Principal Component Analysis (PCA) method and cluster validation indices were investigated and used for the regionalization of drought across the South Korean region. Statistical homogeneity and discordancy of the region was tested on univariate and bivariate frameworks. HCPC indicate that South Korea should be divided into four regions which are closer to being homogeneous. Univariate and bivariate homogeneity and discordancy tests showed the significant difference in their results due to the inability of univariate homogeneity and discordancy measures to consider the joint behavior of duration and severity. Regionalization of drought for SPI time scale of 1, 3, 6, 12, and 24 months showed significant variation in discordancy and homogeneity of the region with the change in SPI time scale. The results of this study can be used as basic data required to establish a drought mitigation plan on regional scales.

Keywords: regionalization; topography; hydro-climatic features; HCPC; PCA; homogeneity; discordancy

1. Introduction

Recent variation of climate on an interdecadal time scale due to global warming across South Korea has had a large impact on agriculture production and water resource management [1], and may cause floods [2,3] and drought conditions. The occurrence of extreme drought events has increased abruptly in the East Asian region since 1980 [4]. The spatial and temporal relationships of drought occurrence and intensity between South Korea and East Asia has showed strong correlation [5]. Drought is expected to become more frequent and severe, with increasing water shortage due to increase in population and uncertainty in water supplies [6]. Since large urban and industrial areas in the Korean peninsula have showed a significant increase in annual temperature, which may lead to changes in precipitation pattern [1], the meteorological droughts in the Korean peninsula can be correlated with the attributes such as hydrological, climatic, and physiographic characteristics [7]. Therefore, drought events could be characterized by the joint behavior of variables which are not usually independent.

In regional drought modeling, multivariate analysis is a collection of procedures to analyze the association between climatic, hydro-meteorological and physiographic variables, which are known to be strongly correlated. Since collection and analysis of a large number of drought variables are often time consuming and cumbersome to handle, adequate simulation of these variables needs a reduction in number of variables. This screening of data is able to provide a rational basis for a multi-dimensional classification of variables and can be used as a basis to perform regional drought frequency analysis [8].

The main purpose of the multivariate analysis is to study resemblances and differences between dataset individuals using dimension reduced subspace. To accomplish this task there are some traditional methods available, such as Principal Component Analysis (PCA) and Factor Analysis (FA), both of which reduce dimensionality by forming linear combinations of the features. PCA looks at the representation of lower dimensions that account for the most variance of the variables, while FA seeks the most correlation among the variables. PCA has been widely used to determine the spatio-temporal patterns of climatic variables [9]. The relationship between water quality parameters and land use types were evaluated by correlation analysis and PCA in South Korea [10]. PCA analysis was also used to regionalize the seasonal variation in the monthly precipitation data from 1931 to 1990 at 85 stations in Turkey [11].

Multivariate datasets need to be classified in groups using some clustering algorithm after the dimensional reduction. Various clustering methods such as fuzzy c-means (FCM) clustering, self-organizing feature map (SOFM) [12], hierarchical clustering, and K-means clustering have been used for the delineation of hydrological homogeneous regions. In hierarchical clustering and partitioning methods, sites are divided into different clusters based on distance measurements (e.g., Euclidean distance).

There are several studies devoted to the regionalization of drought and other climatic variables. Many are based on stochastic processes for simulating the time-space variability of drought [13]. Basist et al. [14] developed statistical relationships between topography and the spatial distribution of mean annual precipitation. These relationships were derived using linear bivariate and multivariate analysis. In previous studies, regionalization across South Korea was done on the basis of specific variables derived from precipitation [15,16]. Regionalization was also performed by using drought variables [15]. However, precipitation or drought phenomena can be influenced by topographic or other climatic variables. Precipitation patterns in South Korea are highly influenced by the summer monsoon, known as the "changma front", and thus rainy season during summer becomes short with an increase in the amount of rainfall and number of heavy rainfall days [1,17]. Overall, the annual precipitation trend across South Korea showed the decline in annual precipitation [18]. Since South Korea has a mountainous environment (Section 2.1), there is a possibility of more rain at the areas located at higher altitude, as the temperature gets cooler with the increase in altitude. Therefore, seasonal and annual precipitation is far from being spatially homogeneous. Therefore, it will be preferential to regionalize the drought after the delineation of regions based on physiographic and climatic variables, since the Korean peninsula is characterized by heterogeneous topography, vegetation, and geology. Proper delineation of homogeneous regions helps us to improve the quantile estimation in regional drought frequency analysis, and can be utilized in drought risk planning and management on regional scales.

The major problem of drought frequency analysis is the lack of availability of the lengthy records that limit the reliability of statistical estimates of drought quantiles. To cope with this problem, "pooling" or "regionalization" of information from multiple sites is often used. To overcome the shortage of observed data, the regional drought analysis attempts to collect the similar sites in one region. The "L-moments" statistical approach is recommended for regional analysis because it has many theoretical advantages such as characterization of a wide range of distributions, able to consider correlations between samples and robust to the presence of outliers [19]. However, this approach is limited to the performance of heterogeneity tests only at the univariate framework. The major problem with this approach is that each variable represents its own homogeneous region. Therefore, univariate regional heterogeneity tests can only provide a limited evaluation of drought at ungauged sites and are not sufficient to fully represent multivariate phenomena of drought (drought duration and severity). To cope with the above stated problem, the approach is extended to the multivariate framework, incorporating a Multivariate Homogeneity Test in order to check the homogeneity of the region with several characteristics [20]. The multivariate homogeneity test has several advantages over univariate analysis, such as including the control of the first kind error and the consideration of the correlation between variables etc., [20,21]. Most literature related to hydrological multivariate analysis dealt with the at-site (local) multivariate analysis using topological, climatic or drought attributes. However, very little effort has been done for the joint representation of drought characteristics in regional drought modeling at ungauged sites. The purpose of this study is (i) to suggest a statistically accurate clustering method to cope with complex topographical and climatic features of South Korea (ii) to regionalize drought using extended bivariate discordancy and heterogeneity measures by identifying spatial changing properties of drought, (iii) analyze the effect of the Standardized Precipitation Index (SPI) time scale on bivariate regionalization of drought.

Following the introduction, the remaining parts of this paper are organized as follows. Section 2.1 explains the overview of the location of study area and introduces the criteria used to define the drought and the data used in this study, Section 2.2 explains the mathematical overview used as a preprocessor for cluster analysis, Section 2.3 provides an explanation about the clustering algorithm, Section 2.4 explains the indices used for cluster validation, Section 3 provides the results of the effect of different SPI time scale on drought variables, cluster analysis, cluster validating indices and bivariate regionalization of drought using SPI multi time scale and finally, conclusions are presented in Section 4.

2. Materials and Methods

2.1. Study Area

South Korea is located in the northeastern part of the Asian continent between 33–43° N and 124–131° E (Figure 1). More than 70% of the land, especially in the north and east, is covered with mountains. South Korea is influenced by Asian monsoon, with annual mean temperature of 12.3 °C, and average temperature varies from 6.6 °C (winter) to 16.6 °C (summer). In some dry regions, precipitation is less than 1000 mm due to topographical effects, and many parts of South Korea are characterized by the precipitation range of 1200 mm to 1400 mm, which is about 30% greater than the worldwide average of 973 mm [22]. Annual maximum precipitation is usually recorded from late Jun through July, and precipitation recorded during this period accounts for more than 40% of the annual precipitation. The climate in South Korea has complex spatial and temporal variation because of the climate change effects and topographic characteristics consists of complicated mountainous terrain.

Drought Definition and Data

The Korean Meteorological Administration (KMA) serves as a basis to extract monthly precipitation data for 70 rainfall stations across South Korea. The randomness of the monthly precipitation data was investigated using homogeneity, the absence of artificial trends and spurious auto-correlation tests. Three non-parametric tests [23], Mann-Whitney homogeneity test, Mann-Kendall trend test, and Kendall's, autocorrelation test, were applied on all rainfall stations. These trend tests were performed using the "Trend" and "Kendall" packages in R programming [24,25]. A detailed description of each non-parametric tests is provided by [26]. Previous work was used as reference to check the detailed overview of the rainfall trends in South Korea [18]. April precipitation decreased between 15% and 74% for basins located in the south-western part of the Korean peninsula. June precipitation increased between 18% and 180% for most basins. The Mann-Kendall test has been applied by many researchers to detect the seasonal and annual precipitation trends in South Korea [27–29]. The test results showed that more than 15 rainfall stations should be discarded because of the low-quality data and more than 5% missing values. The remaining 55 rainfall stations across South Korea cover more than 35 years (1980–2015) of data, and were used for further analysis. For example, trend tests applied at Sokcho station are shown in Figure 2. Annual precipitation increased from 1980 to 2004 and decreased afterwards. The Mann-Kendall test statistic value of 0.0715 confirms the presence of upward trend at Sokcho station. The Move4 technique (Maintenance of

Variance Extension) proposed by [30] was used to fill the gap of missing values. This method is adopted because extended records are generated while maintaining the variance of the data series.

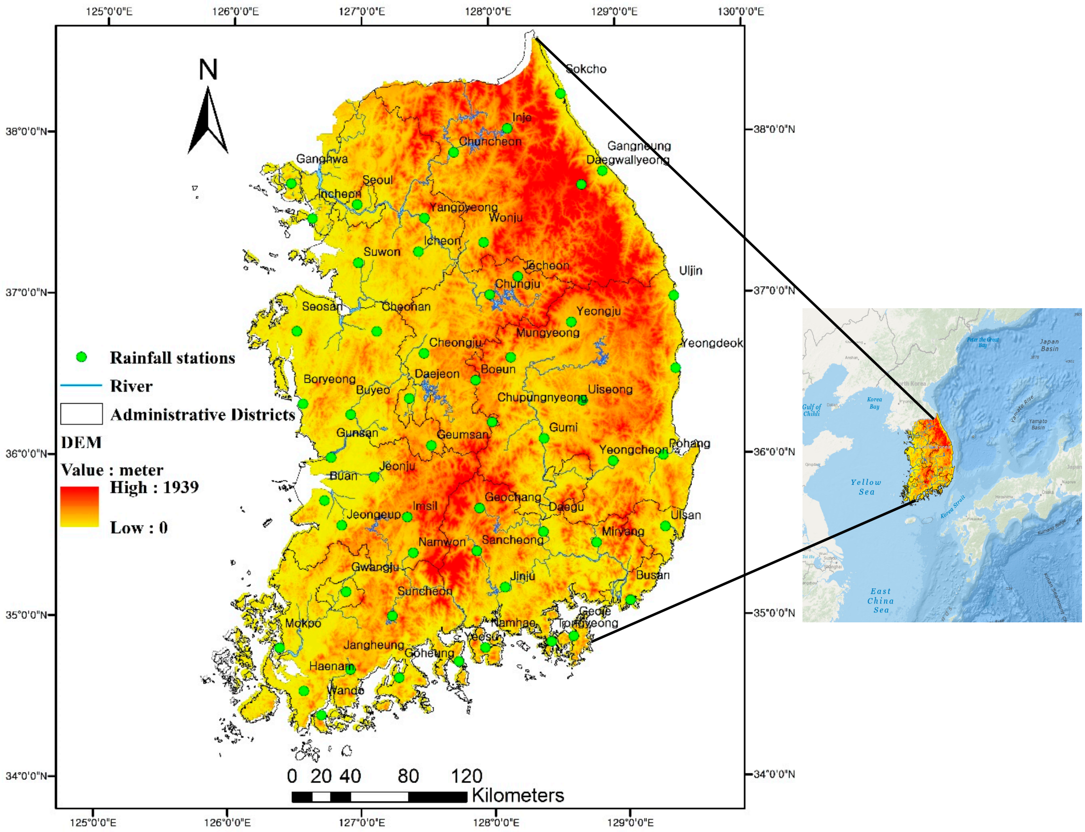


Figure 1. Location of study area and rainfall stations.

SPI originally proposed by [31] was used in this study for regional assessment of the drought in South Korea. SPI is the most widely used drought index [32] because of several advantages, such as (1) flexibility that can be applied on different time scales [33]; (2) being less complex and requiring relatively simple and well-set calculations [34]; (3) being able to adopt different hydroclimatic variables besides precipitation [35]. SPI was computed by fitting gamma distribution on rainfall data at any desired time scale (1, 3, 6, 12 and 24 months). In this study, regionalization of 55 station was done by using all desired time scales from 1 to 24 months. It is found that the gamma distribution fit more closely to the precipitation data of 55 stations across South Korea [36]. Since there are a number of zero-bounded continuous variables in climatology, it is important to give a distribution that may be used for such variables. The gamma distribution has a zero lower bound has been found to fit several such variables well [37]. Additionally, gamma distribution has been recommended by many researchers for SPI analysis on different time scales across South Korea [38,39]. In this context, and for the purposes of the present study, the following two operational definitions were established, in relation to the phenomenon of drought analyzed from a bivariate framework: drought duration and drought severity. Drought duration is the period when the SPI value was below −0.99, and drought severity is the cumulative deficit during that drought event.

The hydro-meteorological features of the Korean peninsula are complex, and it is generally affected by East Asian monsoon, with the heavier precipitation in summer during a short rainy season, and winter temperatures are higher along the southern coast and considerably lower in the mountainous interior. Since the Korean peninsula has complex topographical features, hydro-meteorological characteristics of the region play a vital role in the drought conditions of the region. Therefore, eight important hydrological, climatic (weather regimes), and physiographic (basin) variables (that can affect droughts in South Korea) were selected for delineation of spatially homogeneous regions. Complex phenomena of drought is considered to be correlated with the above-stated eight attributes of the region and were, therefore, included in multivariate analysis. Data was extracted using KMA, GIS maps, and information acquisition. Summary statistics of the 55 rainfall stations across South Korea are presented in Table 1.

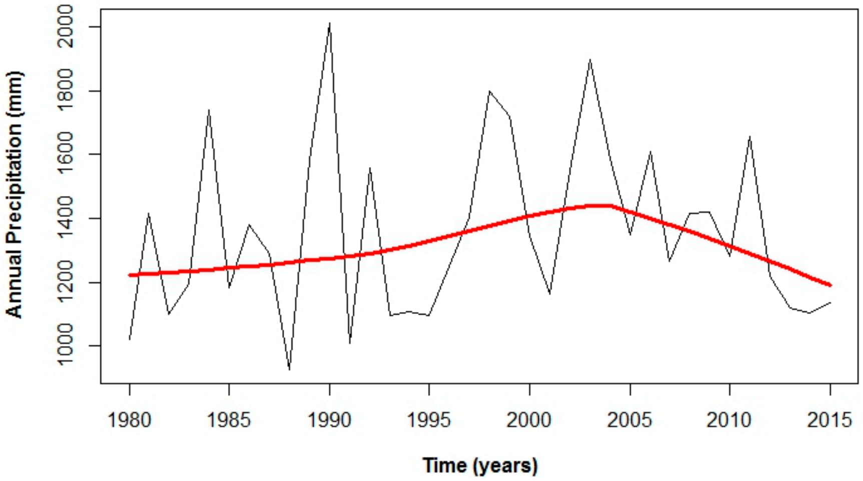


Figure 2. Annual precipitation trend at Sokcho station from 1980 to 2015.

Table 1. Statistical summary of the hydro-meteorological and location variables for the 55 stations selected for cluster analysis.

Variables	Mean	SD [1]	Min [1]	Med [1]	Max [1]	CV [1] (%)
Latitude (N)	36.17	0.85	34.40	36.13	38.25	2.36
Longitude (E)	127.74	0.85	126.38	127.73	129.42	0.67
Elevation above sea level (m)	92.61	114.49	2.90	53.80	772.40	123.63
Mean annual precipitation (mm)	1345.16	193.00	1031.70	1317.30	2007.30	14.35
Mean daily maximum temperature (°C)	18.02	1.30	11.50	18.10	19.90	7.23
Mean daily minimum temperature (°C)	7.82	2.30	2.00	7.60	17.20	29.43
Annual evaporation (mm/year)	1138.72	99.50	956.80	1126.90	1377.60	8.74
Mean relative humidity (%)	677.91	90.07	71.10	694.00	760.00	13.29

Notes: [1] SD = standard deviation; CV = coefficient of variation; Max = Maximum; Min = Minimum; Med = Median.

2.2. Principal Component Analysis

Principal Component Analysis (PCA) is a statistical procedure used to reduce the space into a smaller number of dimensions that can retain most of the information in original space. PCA is able to capture the essential information about the variables and able to determine specific regional characteristics [9,40]. Usually, variables selected for PCA analysis are highly correlated. In this study, the variables indicated in Table 1 were used as input in PCA analysis. The estimation of PCs is the process of reducing inter-correlated variables to some linearly uncorrelated variables. Since the PCs are heavily dependent on the total variation of the hydro-meteorological variables, it is preferred that all variables should be measured in the same unit. Therefore, these variables should be passed through the process of normalization to reduce the effect of units between the variables. A correlation matrix served as a basis to perform the analysis. PCA approach was applied in three steps: (1) standardization of variables and estimation of the correlation matrix, R; (2) estimation of loading matrix using PCA method; (3) Eigenvalue greater one indicates significant PC loading.

All attributes were standardized before the estimation of correlation matrix, using following procedure:

$$X = \frac{x_{ij} - x_i}{S_j} \tag{1}$$

Here value of i and j indicate number of observations (1 to N) and number of hydro-meteorological variables (1 to V) respectively; X denote the matrix of standardized variables; x_{ij} denote ith observation of the jth variable; x_i denotes the mean of the jth variable; S_j indicate standard deviation of the jth variable. The correlation matrix is the minor product moment of the standardized variables and its transpose divided by N and can be expressed as follows

$$R = \frac{X' \times X}{N} \tag{2}$$

Here, X' denotes the transpose of the standardized matrix of the variables. PCs loading matrix was estimated based on correlation matrix of the variables. It shows the degree of correlation of a particular variable with different factors. It is computed by premultiplying the characteristic vector with the square root of the characteristics values of the correlation matrix.

$$A = Q \times D^{0.5} \tag{3}$$

Here A indicates PCs loading matrix; Q and D indicate the characteristic vector and characteristic value respectively, of correlation matrix.

2.3. Hierarchical Classification on Principal Components (HCPC)

HCPC combines PCA, Hierarchical Clustering (HC) and partitional clustering (specifically K-means) [41,42]. HCPC used PCA as pre-processing step to reduce the number of dimensions in parameter space. The attributes mentioned in Table are passed through PCA analysis. Then, HC was applied on PCA to classify individuals into homogeneous groups using Ward criteria. Usually, it is used in complement to factor analysis. The HCPC algorithm gathers the most closed variables on the factorial map in pairs and then aggregates the closest group in pairs until it reaches the proposed level of clustering. The HCPC clustering method is adopted in this study because it has the advantage over the factor analysis that it performs an objective clustering technique to PCA results, which leads to an improvement in the clustering results. Secondly, there is an increase in robustness of the final clusters due to the involvement of mixed algorithm (Ward's classification method with the K-means algorithm). Hence, HCPC method helps to reduce the subjective adjustment in cluster analysis.

2.4. Cluster Validation

Clusters formed by the HCPC approach should be validated using cluster validity indices to determine an optimum number of regions. In this study, four cluster validation indices were used to validate the initial number of clusters. Cluster validation analysis were performed using "fpc" R package.

2.4.1. Connectivity

Suppose N indicates number of observation (rows) and M indicates the number of columns. $nn_{i(j)}$ indicates the jth nearest neighbor of observation i, and let's suppose $x_{i,\, nn_{i(j)}}$ is zero if i and j are in same cluster and $1/j$ otherwise. Then for a clustering partition $\delta = \{C_1, \ldots, C_k\}$ of the N observations into K disjoint clusters, the connectivity is defined as

$$Conn(\delta) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,\, nn_{i(j)}} \tag{4}$$

The parameter L indicates the number of nearest neighbors to use. The connectivity value ranges from 0 to ∞ and should be minimized [43].

2.4.2. Silhouette Width

The silhouette value measures how similar an observation i is to its own cluster compared to other clusters. It ranges from -1 to 1. Higher value indicates better match of an observation to its own cluster. For an observation i [44] proposed as follows

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)} \tag{5}$$

where a_i is the average distance between observation i and all other observations in the same cluster, and b_i is the average distance between i and the observations in the nearest neighboring cluster.

2.4.3. Dunne Index

It is proposed by the [45], is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance.

$$D(\delta) = \frac{min_{C_k, C_l \in \delta, C_k \neq C_l}\left(min_{i \in C_k, j \in C_l} dist(i,j)\right)}{max_{C_m \in \delta} diam(C_m)} \tag{6}$$

where $diam(C_m)$ is the maximum distance between observations in cluster C_m; $dist(i,j)$ is the distance between observation i and j. Dunn Index has the value ranges from zero to ∞, and value should be maximized for better fit.

2.4.4. Calinski and Harabasz Index

It is proposed by [46], It is also called as variance ratio criteria.

$$CH(k) = \left[\frac{B(k)}{W(k)}\right] \times \left[\frac{n-k}{k-1}\right] \tag{7}$$

where n = number of data points, k = number of clusters, $W(k)$ = within cluster variation, $B(k)$ = between cluster variation. Well defined clusters have a large variation between the clusters and small variation within the clusters. Higher value of $CH(k)$ shows the better clusters.

2.5. Bivariate L-Moments

Multivariate L-moments are principally developed by [20]. Let $X(j)$ be a random variable with distribution F_j for $j = 1, 2$. By analogy with a covariance representation of L-moments of order $k \geq 1$, multivariate L-moments are matrices $\wedge k$ with L-comoment elements defined by:

$$\lambda_{k[i\,j]} = Cov\left(X^{(i)},\ P^*_{k-1}\left(F_j(X^{(j)})\right)\right),\ i,\ j = 1,\ 2 \text{ and } k\ =\ 2,\ 3,\ \ldots \tag{8}$$

P^*_k is called as shifted Legendre polynomial. Note that the elements $_{[ij]}$ and $_{[ji]}$ are not necessarily equal. Particularly, the first L-comoment elements are

$$\lambda_{2[12]} = 2Cov\left(X^{(1)},\ \left(F_2(X^{(2)})\right)\right) \tag{9}$$

$$\lambda_{3[12]} = 6Cov(X^{(1)},\ (F_2(X^{(2)}) - 1/2)^2) \tag{10}$$

$$\lambda_{4[12]} = 6Cov\left(X^{(1)},\ 20\left(F_2(X^{(2)}) - \frac{1}{2}\right)^3\right) - 3\left(F_2(X^{(2)}) - \frac{1}{2}\right) + 1 \tag{11}$$

Which are the L-covariance, L-coskewness and L-cokurtosis, respectively. The L-comoment coefficients are as follows:

$$\tau_{k[12]} = \frac{\lambda_{k[12]}}{\lambda_2^1},\ k \geq 3;\ \tau_{2[12]} = \frac{\lambda_{2[12]}}{\lambda_1^1} \tag{12}$$

$\lambda_k^{(j)} = \lambda_{k[jj]}$ is the is the classical kth L-moment of variable $X(j)$, j = 1, 2, as defined by [47]. The matrix of L-comoment coefficients is written as:

$$\Lambda_k^* = \left(\tau_{k[ij]}\right)_{i,\ j=1,\ 2} = \begin{pmatrix} \tau_{k[11]} & \tau_{k[12]} \\ \tau_{k[21]} & \tau_{k[22]} \end{pmatrix} \tag{13}$$

Originally L-comoments are similar in structure to the univariate L-moments and able to capture their attractive properties [21].

2.5.1. Bivariate Discordancy Test

Discordancy test originally proposed by [48] is further extended to the multivariate framework by [20,21]. Discordancy test act as a preliminary screening of the data, before the application of homogeneity test. In this study, drought duration and severity are used to evaluate the discordant sites from the region. It consists of identifying discordant sites among a set of N sites. For this purpose, each site i has the matrix $U_i^t = \left[\Lambda_2^{*(i)}\ \Lambda_3^{*(i)}\ \Lambda_4^{*(i)}\right]$ which is composed by three L-moment matrices $\Lambda_2^{*(i)}$, $\Lambda_3^{*(i)}$ and $\Lambda_4^{*(i)}$ defined in Equation (13)

$$\overline{\mathrm{u}} = N^{-1}\sum_{i=1}^{N} u_i, i = 1,\ \ldots,\ \mathrm{N}. \tag{14}$$

The discordancy measure for site i can be defined as

$$D_i = \frac{1}{3}(u_i - \overline{\mathrm{u}})^t \mathrm{S}^{-1}(u_i - \overline{\mathrm{u}}) \tag{15}$$

where S is the matrix of sums of squares and cross-products

$$\mathrm{S} = \frac{1}{N-1}\sum_{i=1}^{N}(u_i - \overline{\mathrm{u}})(u_i - \overline{\mathrm{u}})^{\mathrm{t}} \tag{16}$$

Hence, if $||D_i||$ takes large values, a site i will be discordant with respect to considered set of sites. For large regions, the critical discordancy ($||D_i||$) value of the constant c can be taken as $\chi_{1-0.5}(3)/3 = 2.6049$. Here, $\chi_{1-\alpha}(d)$ is the quantile of a chi-square distribution of order α with d degrees of freedom. In this study, the critical value proposed by [21] ($||D_i|| > 2.6049$) is considered to be criteria to decide discordancy of a region.

2.5.2. Bivariate Homogeneity Test

The homogeneity of the region is tested using drought duration and severity as defined in Section 2.1. The bivariate homogeneity test is proposed by [21] which is a multivariate analogue of the statistic proposed by [48]. It can be summarized as follows: Let statistic $V_{||.||}$ is defined as:

$$V_{||.||} = \left(\left(\sum_{i=1}^{N} n_i\right)^{-1} \sum_{i=1}^{N} n_i \left|\left|\Lambda_2^{*(i)} - \overline{\Lambda_2^*}\right|\right|^2\right)^{1/2} \tag{17}$$

where $||.||$ is the norm defined above, $\overline{\Lambda_2^*} = \left(\sum_{i=1}^{N} n_i\right)^{-1} \sum_{i=1}^{N} n_i \Lambda_2^{*(i)}$, $\Lambda_2^{*(i)}$ is similar as defined in Equation (13), when we consider L-covariance matrix at k = 2. To get interpretable results of the computed value of the statistic $V_{||.||}$ from the observations, it is convenient to standardize it using a large number of simulated homogeneous regions. Thus, the simulated regions are homogeneous with sites having the same record lengths as their observed counterparts. Thus, heterogeneity measure of a group of sites can be expressed as follows:

$$H_{||.||} = \frac{V_{||.||} - \mu_{\text{Vsim}}}{\sigma_{\text{Vsim}}} \tag{18}$$

where μ_{Vsim} and σ_{Vsim} are respectively the mean and standard deviation of the N_{sim} values of $V_{||.||}$ of simulated regions. Similar like univariate heterogeneity measure, bivariate heterogeneity measure also has criteria to decide whether a region should be considered as homogeneous or not. In this study, following criteria was used to decide homogeneity of the observed region: if $H_{||.||} < 1$ region was considered as homogeneous; if $1 \leq H_{||.||} < 2$ region was considered as acceptably homogeneous; if $H_{||.||} \geq 2$ region was considered as heterogeneous.

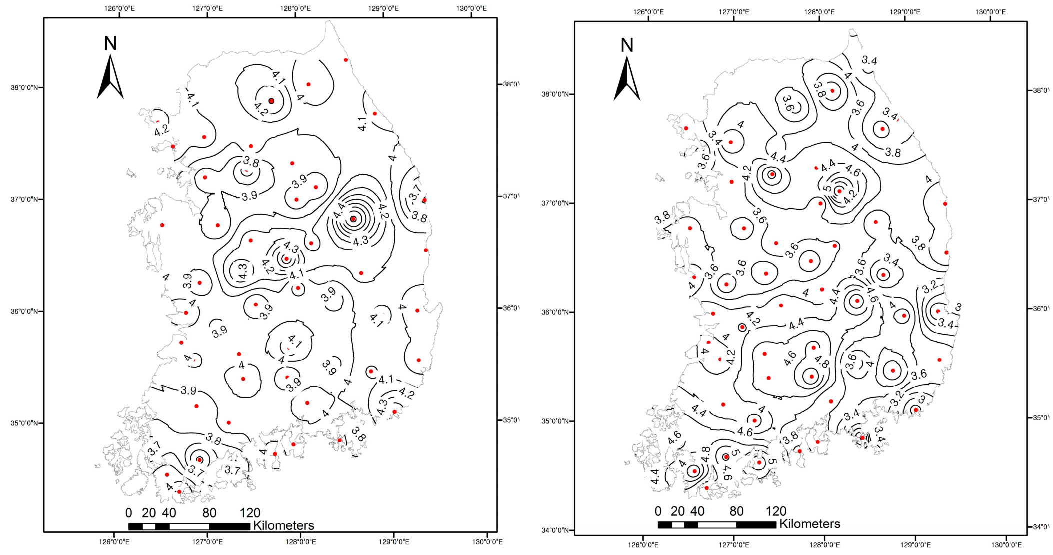
Although it is not part of the objectives of this study, bivariate discordancy and homogeneity measures can be applied for quantile estimation by extending [19] the approach to multivariate domains. This includes the following steps: (i) data preparation, including standardization of variables and application of statistical procedure based on PCA; (ii) identification and acceptance of homogeneous regions based on HCPC and bivariate homogeneity test; (iii) selection of regional frequency distribution for the bivariate case; and (iv) estimation of distribution parameters and quantile function for the bivariate case.

3. Results

3.1. Selected Attributes

The attributes used for cluster analysis include a reasonable number of hydro-meteorological and physiographic variables that can affect drought phenomena in South Korea. Table 1 showed a statistical summary of these selected variables. The highest Mean Annual Precipitation (MAP; 2007.30 mm) and highest Mean Daily Maximum Temperature (MDMXT; 19.9 °C), shown in Table 1, are recorded at Geoje station and Miryang station, respectively. Both stations are located on the southeast coast of South Korea (Figure 1). This is because of the reason that the synoptic disturbances, typhoons or convective systems within the air mass at south coastal areas cause heavy rainfall during the summer season and low rainfall during the winter season. Thus, the south coast faces extremely unusual rainfall patterns. For example, study based on the precipitation trends across South Korea considering typhoon-induced changes and other climate related risks, and it was concluded that over the 1966–2007 period, typhoons contributed 21–26% of seasonal precipitation and broad patterns towards an increase in the magnitude and frequency of precipitation, especially at the south coast [29]. It can be noted from Figure 3b that the droughts of highest Mean Severities (MS) (round about 5) were recorded at surrounding areas of Jangheung and Goheung stations, located at the southwest coast of South Korea. Furthermore, droughts of longest Mean Duration (MD; round about 4.5 months) and highest MS were recorded at surrounding areas of Jecheon station. Summer is hot and humid and winter is cold and dry across the region due to the influence of Siberian air mass [22]. The average annual relative humidity across South Korea is 68.5% and average monthly relative humidity varies from 61.3% in April to 81.3% in July. Relative humidity is considered to be an important attribute that can influence precipitation climates [49], and has been used for regionalization of precipitation [50]. The attributes used for multivariate analysis are as follows: Latitude (Lat), Longitude (Long), Elevation Above Sea Level (EL), Mean Daily Minimum Temperature (MDMT), Annual Evaporation (AE) and Mean Relative Humidity (MRH).

In previous studies [51,52], at-site statistics were also used as an attribute for regionalization using multivariate statistical analysis. However, [19] suggested that the practical formation of homogeneous regions should be based on at-site characteristics; otherwise there would be a tendency to group together all sites that have high outliers, even though these outliers result from random fluctuations, and testing for the homogeneity of the formed regions by a statistic calculated from the "at-site statistics" would be misleading. Therefore, MD and MS are not included in the initial formation of homogeneous regions because of the involvement of at-site statistics.



(**a**) (**b**)

Figure 3. Spatial variability of drought characteristics from 1980 to 2015, across South Korea using SPI-6 (**a**) mean drought duration (months); (**b**) mean drought severity.

3.2. Variation of Drought on Different Time Scale

SPI values for 55 stations were computed for the time scale of 1, 3, 6, 12, and 24 months. For example, the temporal variation of drought for the time scale of 1 to 24 at Sokcho station from 1980 to 2015 is shown in Figure 4.

It is observed that when the time scale is shorter, frequency of the dry and humid periods is short and relatively high. The SPI for the shorter time scale is usually considered as agriculture drought index [31,53] because it represents the water content of vegetation and the soil conditions [54,55]. The SPI for the longer time scale showed that droughts lasted longer and were less frequent. The SPI for the longer time scales such as 12 months or greater are considered as the hydrological drought index, as it can be used for surface water monitoring, e.g., river flows [53]. A longer time scale such as 24 months showed frequent and longer lasting droughts, with few dry or humid periods recorded.

Table 2 showed the changes in the characteristics of drought with SPI-time scale of 1 to 24 months for 55 stations across South Korea. As the SPI timescale increases, total number of drought events decreased because of the decrease in frequency of drought events. However, Mean, Max and Min statistics of drought characteristics (duration and severity) increased with the increase in SPI-time scale because of the increase in the number of long-lasting drought events.

Table 2. Characteristics of drought with changes in SPI-time scale for 55 stations across South Korea.

Time Scale	Total Drought Events	Duration			Severity		
		Mean	Max	Min	Mean	Max	Min
SPI-1	5549	2.07	18	1	1.67	14.67	0.05
SPI-3	3216	3.71	24	1	2.9	27.19	0.09
SPI-6	1455	5.56	32	1	3.91	31.25	1.05
SPI-12	1093	11.3	84	2	9.17	96.92	1.95
SPI-24	727	20.67	117	5	17.29	124.6	3.65

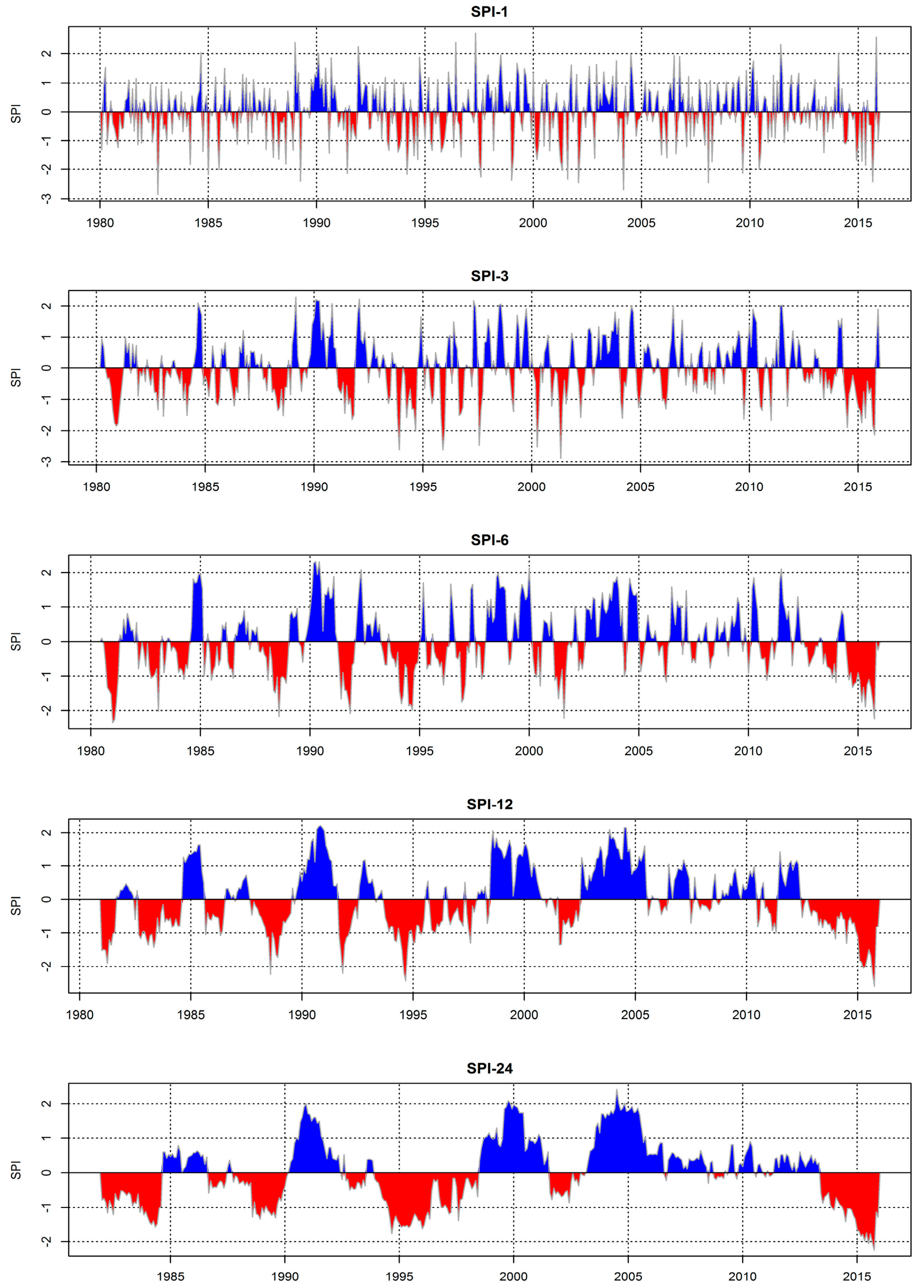


Figure 4. Evaluation of SPI drought index on different time scale.

3.3. Principal Component Analysis

In order to reduce the effect of units, each of the eight attributes shown in Table 1 was standardized using Equation (1). Standardization made variable comparable by scaling them to have (i) standard deviation one and (ii) mean zero. Then, intercorrelation matrix of the standardized variables was computed using Equation (2). The spatial pattern of the intercorrelation matrix of the eight attributes reveals that there is a correlation coefficient of 0.6 existing between Long and AE, and MDMXT and MDMNT (Figure 5). A moderate correlation (correlation coefficient more than 0.40) exists between Lat and EL, MDMXT and AE. Since the attributes such as MRH showed relatively weak correlation with any other parameter, it is hard to categorize the parameters into a component or attach any physical significance. Therefore, in the next step, PCA method has been applied on the correlation matrix.

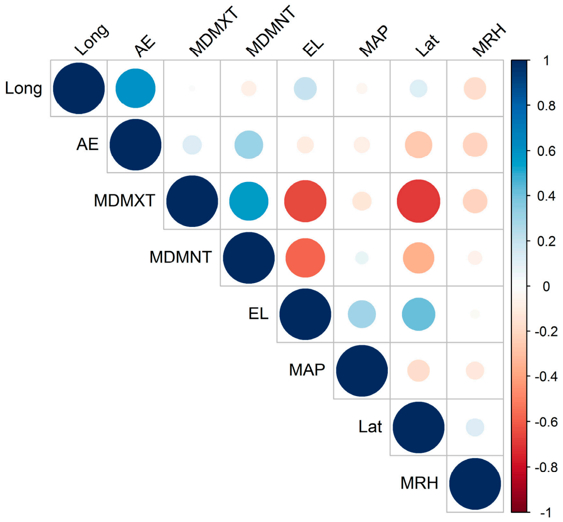


Figure 5. Intercorrelation matrix of the attributes selected in this study.

The principal component analysis applied on the correlation matrix revealed that 78.7% (first component 32.6%, second component 19.4%, third component 15.8% and fourth component 10.9%) of the information (variances) contained in the data were retained by the first four principal components (Figure 6). Since eigenvalues of the first four principal components (first component 2.611, second component 1.55, third component 1.264 and fourth component 0.873) were closer to or greater than 1, the first four principal components were selected for the cluster analysis. The principal component-loading matrix was computed based on correlation matrix using Equation (3). The principal component loading was used to identify the strength of correlation existing between each component and the individual attributes. The contribution of attributes (using principal component loading) in accounting for the variability in principal components is shown in the form of a percentage (Figure 7). Attributes having a high contribution in first and second principal components are considered as most important. However, attributes that do not contribute to any principal components or contributed with the last dimensions are considered as less important in explaining the variability in the dataset. The red dashed line on the graphs in Figure 7 denotes the expected average contribution. Supposing contribution of the attributes is uniform, then expected average contribution value can be computed as follows; $(1/number\ of\ attributes) \times 100$. In our study, 12.5% is considered to be expected average contribution. For a given component, an attribute with a contribution larger than this limit (red dashed line) is considered to be important in contributing to the component. Figure 7 showed that the attributes such as MDMXT, LAT, EL and MDMNT contributed most to first principal component and Long, AE and MRH contributed most to second principal component, and MAP and Lat contributed most to the third principal component, and MRH, MDMNT, MDMXT and AE contributed most to the fourth principal component. It is concluded that MRH is the least important attribute in explaining the variability of the dataset because of their small contribution to the first three principal components. Usually in PCA approach, the attributes that are less significant in explaining the component variance are screened out of analysis. However, MRH still has a significant enough correlation to be included in further analysis.

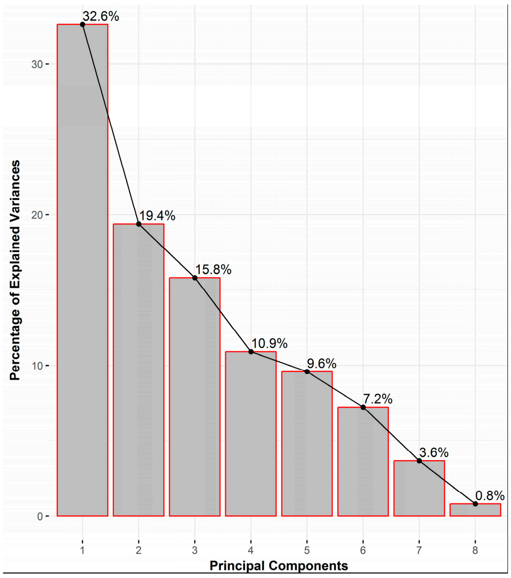


Figure 6. Explained variances of each principal component.

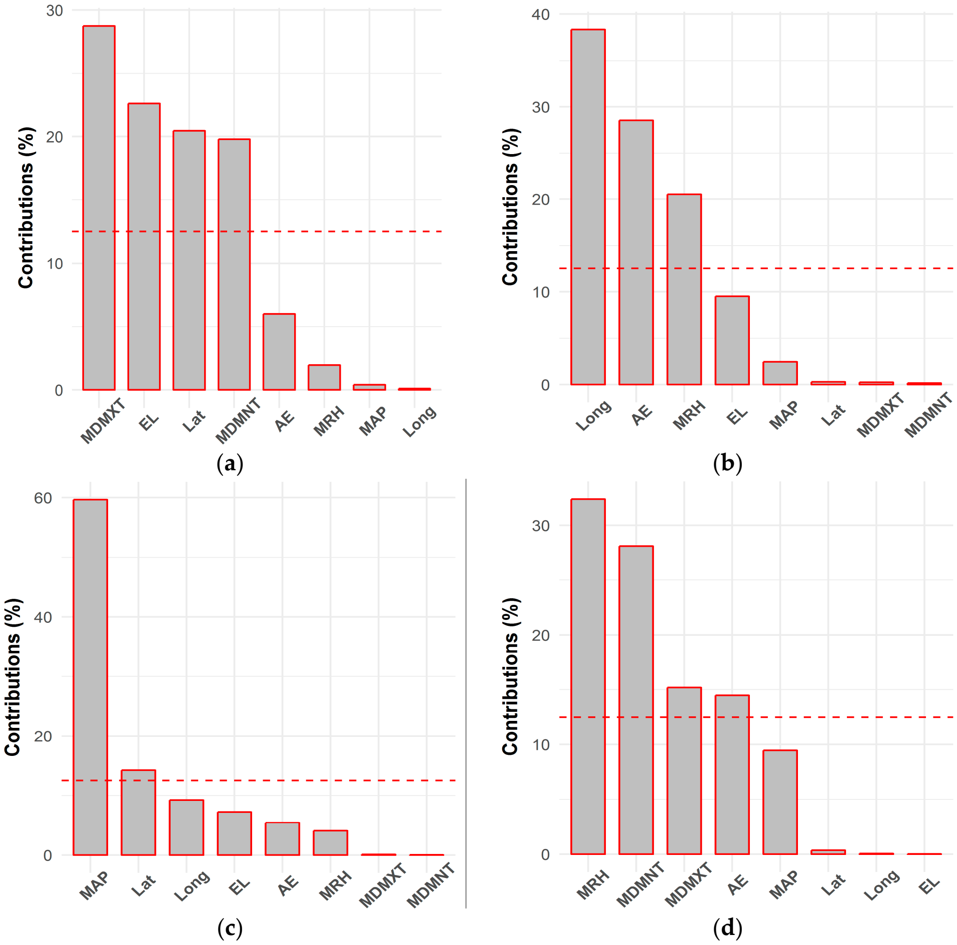
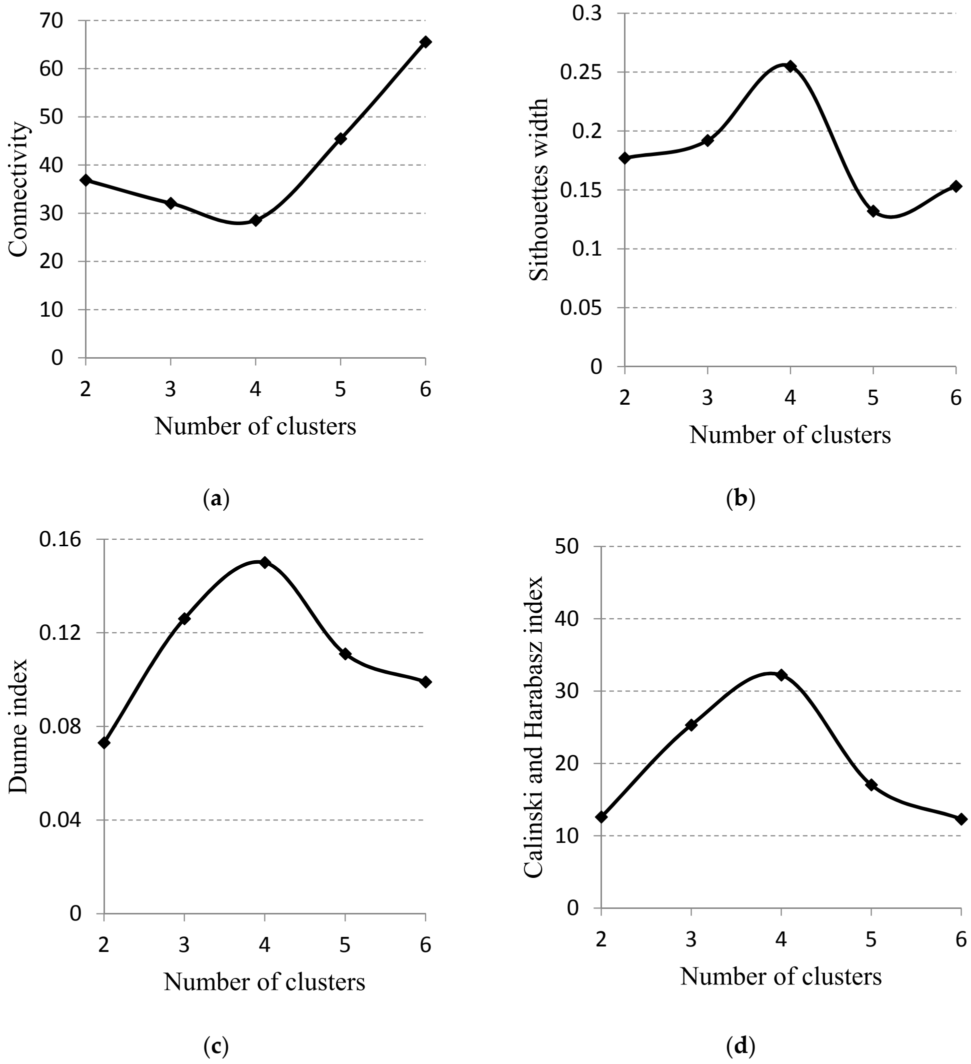


Figure 7. Contribution percentage of eleven attributes to the following principal components: (**a**) first principal component; (**b**) second principal component; (**c**) third principal component; (**d**) fourth principal component.

3.4. Cluster Analysis

HCPC method was applied on the results obtained from PCA. To identify the stability of each region, initially clusters were formed and assessed visually by plotting them on the geographical space of South Korea. It was noted that stable regions do not change their configuration drastically with the change in the number of clusters formed by HCPC method. However, since visual plotting method is subjective and may lead to incorrect estimation of clusters, cluster validation indices were used to aid the selection of an optimal number of clusters.

Four cluster validity indices, connectivity, silhouette width, Dunne index and Calinski and Harabasz index, were tested by varying the possible number of clusters from two onwards to examine an optimal number of clusters. It is evident that the indices, silhouette width, Dunne index and Calinski and Harabasz index maximize their values when the number of clusters reaches four, while connectivity minimizes when the number of clusters reaches four. Hence, the values of all validity indices showed the good agreement when the number of clusters reached four (Figure 8). Cluster validity indices based on the HCPC algorithm showed that the study area should be divided into four regions, which are close to being homogeneous.



(**a**) (**b**)

(**c**) (**d**)

Figure 8. Cluster validation statistics using the plots of (**a**) Connectivity; (**b**) Silhouette width; (**c**) Dunne index and (**d**) Calinski and Harabasz index.

3.5. Bivariate Regionalization of Drought across South Korea

Usually, the regions obtained by clustering algorithms are not statistically homogeneous. Therefore, they must be adjusted to make them homogeneous. This step of regionalization is not properly explained if the set of attributes considered for cluster analysis is exhaustive (i.e., the attributes composed of causal variables which affect drought characteristics of the region under study). However, since it is impossible to have an exhaustive set of attributes to perform regionalization of drought, regions formed by cluster analysis (HCPC algorithm for this study) need not, and are usually not, homogeneous, and therefore regional revision process is required to improve the homogeneity of the region [19]. This is a well-known fact in hydrology and consequently, hydrologists proceed with adjusting the regions to make spatially homogeneous groups. Nevertheless, adjustments need not to be significant if the selected attributes for the cluster analysis is likely to affect drought phenomena in the region and if an appropriate clustering algorithm is used for regionalization of drought [56].

The goal of the adjustment is to make sure that no site is discordant and all regions are homogeneous. Ref. [19] suggested many options to adjust the initial groups formed by cluster analysis. These options include: (i) deletion of a site or few sites from the region; (ii) shifting of a site or few sites from one region to another; (iii) subdivision of regions from one region to two or more regions; (iv) allowance of sites to be shared partially between the regions; (v) the breaking up of regions by shifting their sites to other regions; (vi) merging a region with another or others; (vii) merging two or more regions and redefining the groups; (viii) obtaining more data and redefining the groups.

The regions made by HCPC algorithm are found very useful in adjusting the regions to improve their homogeneity. Particularly, there is no need to devote special effort for the steps (i), (iii), (iv), (v), (vi), (vii) and (vii) described above because the number of clusters is chosen reasonably using the HCPC algorithm. In this study, only option (ii) was used on the clusters formed by the HCPC algorithm to regionalize the drought characteristics across South Korea.

In this study, discordancy and homogeneity of the regions are tested for the SPI-time scale of 1, 3, 6, 12, and 24 months. The total number of drought events differs from shorter time scale to longer time because of the variation in length and frequency of dry and humid periods as shown in Figure 4 and Table 2. Therefore, selected time scale may have direct impact on the homogeneity of the region.

A site is considered to be discordant if it exceeds the critical value of the discordancy measure. Firstly, although [19,48] defined the critical values to declare a site to be discordant, it is preferable to mark the sites having high values of discordancy measure, for example in case of SPI-6 as shown in Table 3. Secondly, since heterogeneity measures (H) of the region change with the addition or removal of any site from the region, it should be continuously examined during regionalization process.

For regionalization of drought, homogeneity of the region needs to be assessed statistically using drought variables, after the formation of homogeneous regions using cluster analysis. An L-moment-based discordancy test proposed by [34] and extended to the multivariate framework by [20] was applied to identify the sites with gross errors in their data or those that are grossly discordant with the region as a whole. Discordancy is measured in terms of sample L-moments ratios (L-CV, L-Skewness, and L-Kurtosis) in the univariate case and the L-comoment ratios (L-covariance, L-coskewness and L-cokurtosis) in the bivariate case as described in Equations (9)–(11).

The results of univariate and bivariate discordancy tests in case of SPI-6 applied initially on clusters formed by the HCPC algorithm is shown in Table 3. In the case of univariate discordancy test applied on drought variables (duration and severity), no site is discordant in cluster 1, 2 and 3 except Yeoungju station in cluster 4. However, in case of bivariate discordancy test, Tongyeong station in cluster 1, and Ganghwa, Jeonju and Boeun stations in cluster 3 are identified as discordant sites. These sites are not discordant in the univariate case because the univariate discordancy test is unable to take account of the correlation between drought variables. Following option (ii) stated above for adjusting the regions, the inclusion of Jeonju station (having the highest value of discordancy) in cluster 2 and then cluster 4 increases the heterogeneity measures of these clusters. However, the inclusion of Jeonju station in cluster 1 reduced the heterogeneity of the region by a significant amount as shown in Table 4, and discordancy measure of Tongyeong station (identified as a discordant station as shown in Table 3 reduced drastically without moving it to any other region (Table 4). Using a similar approach, the moving of other discordant sites continues until these sites were no longer discordant and heterogeneity of the adjusted regions was also reduced. Univariate and bivariate discordancy measures of the final adjusted homogeneous regions are shown in Table 4. To adjust highly heterogeneous cluster 3, all discordant sites were moved to clusters 1, 2 and 4 using option (ii). Final adjusted homogeneous regions in Table 4 and Figure 9c showed that there is an improvement in the size of cluster 2, which was the smallest cluster in the region, and reduction in the size of cluster 3, which was the largest cluster in the region.

Table 3. Results of univariate and bivariate discordancy measure for SPI-6 at 55 stations across South Korea, using the clusters formed by HCPC algorithm, and bold is selected as discordant sites, $D_{i,\,D}$ denotes the discordancy for drought duration, $D_{i,\,S}$ denotes the drought severity and $D_{i,\,SD}$ denotes the discordancy for both drought duration and severity.

Cluster 1							
Site	$\boldsymbol{D_{i,\,D}}$	$\boldsymbol{D_{i,\,S}}$	$\boldsymbol{D_{i,\,SD}}$	**Site**	$\boldsymbol{D_{i,\,D}}$	$\boldsymbol{D_{i,\,S}}$	$\boldsymbol{D_{i,\,SD}}$
Jangheung	1.54	1.85	1.38	Mokpo	1.33	1.28	0.61
Haenam	0.11	0.13	0.17	Jinju	0.96	1.03	1.17
Yeongcheon	0.10	1.19	0.91	Geoje	1.21	1.69	1.90
Miryang	0.95	0.30	0.64	Buan	1.82	2.20	2.29
Sancheong	0.49	0.46	0.74	Namhae	0.14	0.27	0.41
Ulsan	0.43	0.10	0.59	Jeongeup	0.18	0.26	0.65
Gwangju	0.29	0.56	0.69	Goheung	0.80	0.69	0.33
Busan	1.88	1.80	2.10	Yeosu	0.81	0.40	0.51
Tongyeong	1.98	2.30	**3.19**	Wando	1.68	1.13	1.95
				Suncheon	1.96	0.97	1.45
Cluster 2							
Chupungnyeong	1.00	1.00	1.37	Geochang	1.00	1.00	1.19
Geumsan	1.00	1.00	0.09				
Cluster 3							
Chuncheon	0.43	0.71	0.54	Buyeo	0.34	0.40	0.26
Chungju	0.65	0.62	1.10	Suwon	0.74	0.86	0.79
Boeun	2.01	1.95	**3.98**	Imsil	0.44	0.76	0.33
Daejeon	1.69	1.54	1.81	Boryeong	0.38	0.43	0.30
Jecheon	0.25	0.05	0.37	Namwon	0.44	1.10	0.87
Yangpyeong	0.28	0.21	0.89	Seoul	0.77	0.28	0.95
Icheon	0.51	1.76	1.52	Incheon	1.14	1.12	1.96
Ganghwa	2.29	1.91	**3.81**	Wonju	1.32	0.75	1.52
Jeonju	1.75	1.94	**4.10**	Cheongju	0.48	1.05	0.73
Cheonan	1.08	1.30	1.47	Gunsan	0.95	0.06	1.62
				Seosan	0.40	0.72	0.95
Cluster 4							
Sokcho	0.31	1.24	0.47	Uiseong	1.08	1.11	0.92
Daegwallyeong	0.16	0.3	0.55	Gumi	1.11	1.11	1.29
Gangneung	0.85	1.01	1.23	Yeongju	**2.97**	2.06	1.52
Uljin	2.06	1.73	1.96	Mungyeong	0.68	0.1	0.23
Pohang	1.34	1.27	1.16	Yeongdeok	0.51	0.84	0.61
Daegu	0.86	0.24	0.23	Inje	0.61	0.98	1.21

Table 4. Results of univariate and bivariate discordancy measure for the final adjusted homogeneous regions using SPI-6. C_u and C_b indicate the critical values of discordancy measure for univariate and bivariate cases, respectively.

Cluster 1 (C_u = 2.63, C_b = 2.60)							
Site	$D_{i,\,D}$	$D_{i,\,S}$	$D_{i,\,SD}$	**Site**	$D_{i,\,D}$	$D_{i,\,S}$	$D_{i,\,SD}$
Jangheung	1.64	1.68	1.35	Jinju	0.92	1.09	1.32
Haenam	0.13	0.11	0.31	Geoje	1.23	1.76	1.66
Yeongcheon	0.12	1.22	0.81	Buan	1.91	2.45	1.89
Miryang	0.64	0.32	0.52	Namhae	0.15	0.2	0.56
Sancheong	0.36	0.42	0.09	Jeongeup	0.18	0.21	0.33
Ulsan	0.46	0.11	0.67	Goheung	0.73	0.58	0.81
Gwangju	0.25	0.59	0.42	Yeosu	0.81	0.44	0.51
Busan	1.86	1.91	0.95	Wando	1.78	1.19	1.41
Tongyeong	1.94	2.06	1.65	Suncheon	1.62	1.04	1.75
Mokpo	1.21	1.16	0.88	Jeonju	1.77	0.98	1.29
Cluster 2 (C_u = 1.33, C_b = 2.60)							
Chupungnyeong	0.49	0.74	1.22	Boeun	1.29	1.32	1.66
Geumsan	0.83	0.89	1.65	Yeongju	1.08	1.02	0.74
Geochang	1.33	1.03	0.07				
Cluster 3 (C_u = 3.00, C_b = 2.60)							
Chuncheon	0.61	1.35	1.65	Imsil	0.83	1.02	1.19
Chungju	0.58	0.96	1.10	Boryeong	0.48	0.59	0.94
Daejeon	2.9	1.94	1.74	Namwon	0.66	1.51	1.65
Jecheon	0.42	0.02	0.56	Seoul	1.45	0.44	1.52
Yangpyeong	0.63	0.33	0.23	Incheon	2.13	1.69	1.71
Icheon	1.04	2.42	1.11	Wonju	1.46	0.96	0.59
Cheonan	1.22	1.16	1.06	Cheongju	0.37	0.99	1.23
Buyeo	0.47	0.53	0.65	Gunsan	1.25	0.09	0.27
Suwon	0.99	1.12	1.32	Seosan	0.50	0.86	0.59
Cluster 4 (C_u = 2.86, C_b = 2.60)							
Sokcho	0.20	1.45	1.29	Uiseong	1.04	1.51	1.41
Daegwallyeong	0.13	0.26	0.61	Gumi	1.09	1.17	0.55
Gangneung	1.01	0.88	1.10	Mungyeong	0.57	0.05	0.47
Uljin	2.05	1.47	1.88	Yeongdeok	0.48	0.74	0.19
Pohang	1.41	1.34	1.21	Inje	0.61	0.98	1.21
Daegu	1.01	0.18	0.71	Ganghwa	2.39	1.96	1.01

The similar subjective adjustment approach, as described above for SPI-6, is adopted for SPI 1, 3, 12 and 24 months. In case of SPI-1, the univariate and bivariate discordancy tests applied on drought variables (duration and severity) showed that Boeun station from cluster 3, and Suncheon and Buan stations from cluster 1 are identified as discordant sites, and therefore moved to cluster 2 to make them homogeneous. However, Suncheon station remains discordant even after moving to clusters 2, 3 and 4. Therefore, to improve the homogeneity of cluster 1, Suncheon station was removed from the analysis. In case of SPI-3, the Boeun station from cluster 3, and Suncheon and Gwangju stations from cluster 1 are identified as discordant sites, and therefore moved to cluster 2 to adjust the heterogeneity of the cluster. In case of SPI-12, the Boeun station from cluster 3, Boryeong station from cluster 3 and Jeongeup station from cluster 1 are identified as discordant sites, and therefore moved to cluster 2 to improve the homogeneity of each cluster. In case of SPI-24, the Gunsan station is identified as a discordant site in cluster 3, and moved to cluster 2 to adjust heterogeneity of the region.

Heterogeneity measures of the final adjusted regions for the time scale of 1, 3, 6, 12, and 24 are presented in Table 5. In case of drought duration, with the exception Region I, II of SPI-1, Region III of SPI-3, Region IV of SPI-6 and Region IV of SPI-12 (acceptable homogeneous with $1 \leq H < 2$), all regions satisfy the condition of homogeneity ($H < 1$). In case of drought severity, all regions are homogeneous except Region II of SPI-3. In case of bivariate homogeneity for both duration and severity, except Region IV of SPI-12 all other are homogeneous.

During the process of regional adjustment, it was noticed that the stations having larger number of drought events have more influence on the heterogeneity measure of the region than that of smaller number of drought events. This is because the heterogeneity measures proposed by [48] are calculated in such a way that the weight information from each station is in proportion to its length of records. In case of precipitation regionalization studies, where record lengths are generally short, heterogeneity measure may affect the regionalization process because of the sample size effect on the stability of L-moments. However, since this study is using a fixed length of precipitation record for each station (1980–2015), there are no significant changes in number of drought events from one station to another. Therefore, heterogeneity measure is not influenced too much by the small variation in length of drought events between the stations.

Table 5. Heterogeneity measure for the final adjusted homogeneous regions on different time scales.

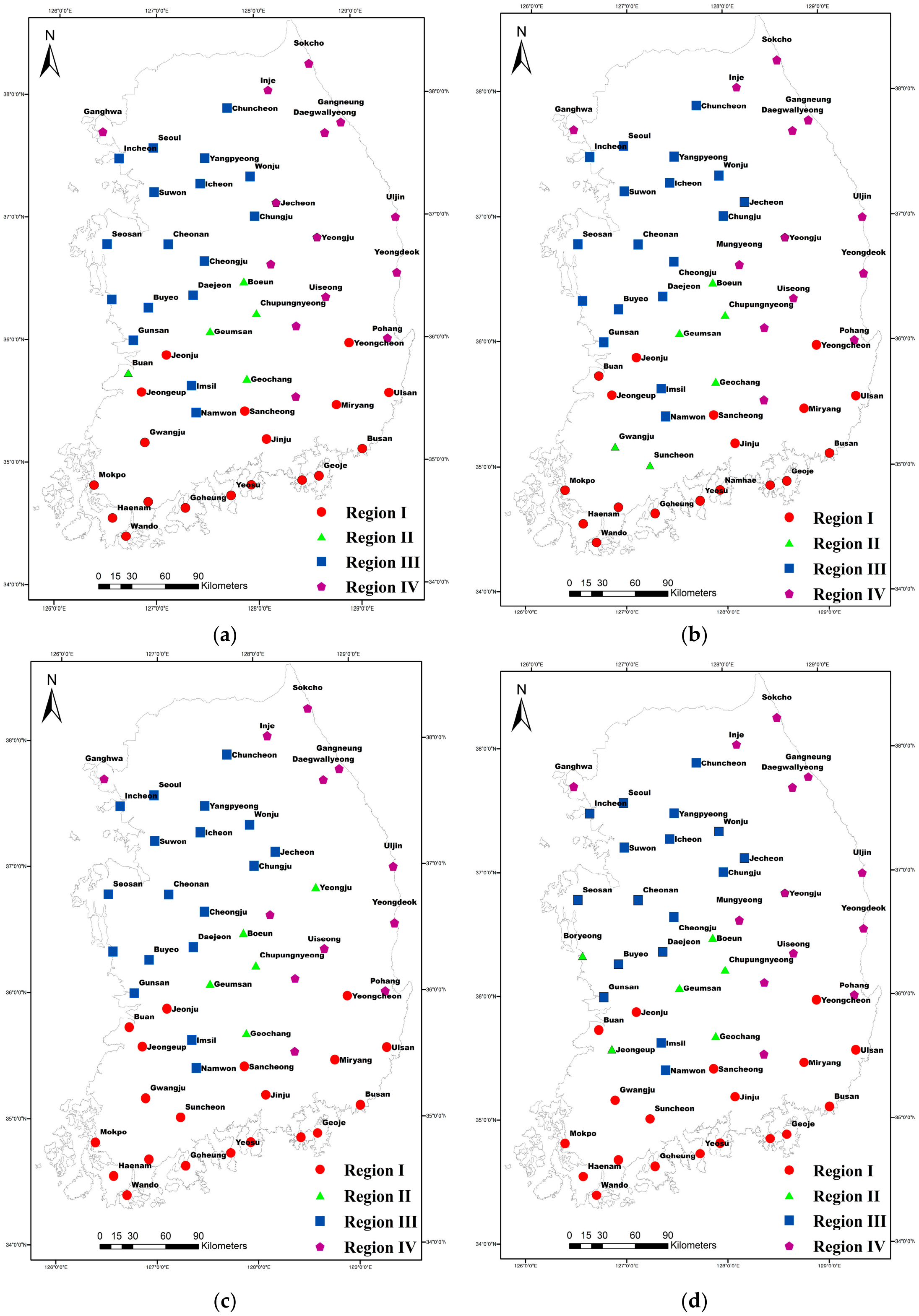
Scale (Months)	Region	Drought Events	Stations	*D*	*S*	*DS*
SPI-1	I	1730	17	1.10(A.H [2])	−0.73(H [1])	0.35(H [1])
	II	484	5	1.49(A.H [2])	−1.39(H [1])	−1.23(H [1])
	III	1970	19	−0.10(H [1])	−1.68(H [1])	−0.29(H [1])
	IV	1365	13	0.71(H [1])	−0.29(H [1])	0.52(H [1])
SPI-3	I	938	16	−2.56(H [1])	−1.00(H [1])	−0.23(H [1])
	II	352	6	0.43(H [1])	1.39(A.H [2])	0.95(H [1])
	III	1221	21	1.21(A.H [2])	−0.13(H [1])	0.58(H [1])
	IV	705	12	−1.13(H [1])	−0.84(H [1])	−1.56(H [1])
SPI-6	I	516	20	−0.70(H [1])	−0.82(H [1])	−0.11(H [1])
	II	139	5	0.85(H [1])	−0.12(H [1])	−0.34(H [1])
	III	492	18	−2.21(H [1])	−2.26(H [1])	−1.45(H [1])
	IV	308	12	1.47(A.H [2])	0.74(H [1])	0.56(H [1])
SPI-12	I	325	17	−2.87(H [1])	−3.33(H [1])	−2.75(H [1])
	II	110	6	−2.39(H [1])	−2.67(H [1])	−2.21(H [1])
	III	420	20	−1.56(H [1])	−1.52(H [1])	−1.20(H [1])
	IV	238	12	1.46(A.H [2])	1.23(A.H [2])	1.13(A.H [2])
SPI-24	I	245	18	−0.86(H [1])	−0.97(H [1])	−0.85(H [1])
	II	41	4	−0.85(H [1])	−1.16(H [1])	−1.02(H [1])
	III	283	21	−2.25(H [1])	−0.33(H [1])	−1.26(H [1])
	IV	158	12	−0.48(H [1])	−0.22(H [1])	−0.36(H [1])

Notes: [1] H = Homogeneous region, [2] A.H = Acceptably homogeneous region.

The spatial distribution of finally identified homogeneous regions for SPI 1, 3, 6, 12, and 24 are shown in Figure 9a–e, respectively. Region I for all SPI-time scales is spread mainly along the south coast of South Korea and consists of extremely unusual precipitation patterns as compared to other regions. For example, in case of SPI-6, it can be observed from Figure 3b that south coastal areas (surrounding areas of Jangheung and Goheung stations) faced droughts of highest severities because of abrupt changes in precipitation patterns. Overall Region I faced droughts of longest duration and highest severities as compared to other regions because of major contribution of typhoons to the seasonal (particularly summer) precipitation patterns and convective systems within the air mass at south coastal areas. These results match well with the previous literature [29]. A study [17] based on spatial patterns of trends in summer precipitation showed a significant increasing trend in amount and intensity of precipitation at southeast coastal areas.

Region II is the smallest region and delineated for the SPI-time scales (1, 3, 6, 12, and 24 months). The number of stations in this region varies from 4 to 6 and occupies the mid-latitude inland of South Korea. Moderate drought attributes observed in Region II may be because of its location near the coastal areas, thus less affected by summer typhoons, which occur less at the coastal areas.

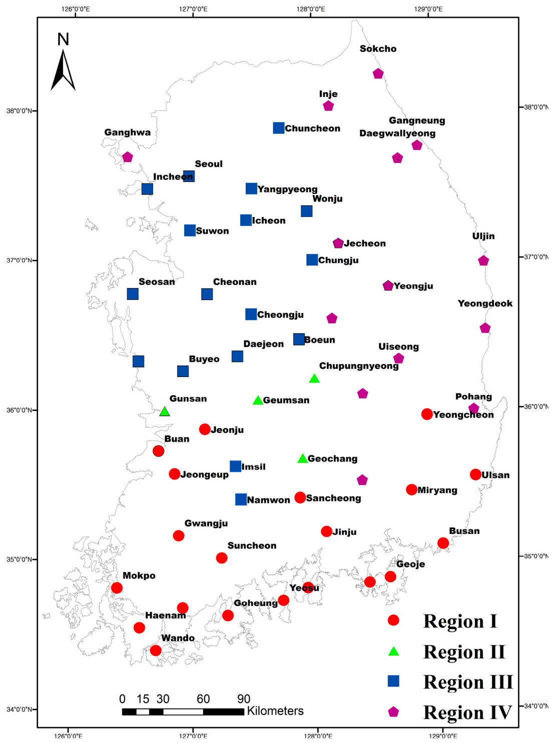
Region III delineated for the SPI-time scales (1, 3, 6, 12, and 24 months) are spread mainly along the north-east side of South Korea. The number of stations varies from 18 to 21. This region has relatively low elevation above sea level as compared to Region I, II and IV, as shown in Figure 1.



(**a**) (**b**)

(**c**) (**d**)

Figure 9. *Cont.*



(**e**)

Figure 9. Spatial distribution of final delineated homogeneous regions for; (**a**) SPI-1; (**b**) SPI-3; (**c**) SPI-6; (**d**) SPI-12; (**e**) SPI-24.

Region IV is spread mainly along the northeast coast of South Korea. Delineation of Region IV for the SPI-time scales (1, 3, 6, 12, and 24 months) indicate that the number of stations varies from 12 to 13. Topographical and hydro-climatic features of this region are extremely complex as compared to other regions. The EL attribute of the region varies from a lowest elevation above sea level to highest elevation above sea level as compared to Region I, II and III, as shown in Figure 1. The greater variation in hydro-climatic attributes in this region can be correlated with an increase in intensity, frequency, and duration of summer typhoons originating in the west Pacific regions.

4. Conclusions

This study has attempted to investigate the regionalization of drought variables together with other physiographic and climatic variables. It aims to cope with complex hydro-climatic and topographical features of South Korea. Therefore, eight important hydrologic, climatic and physiographic characteristics were selected for the process of regionalization. Drought characteristics were extracted using SPI truncation level approach across 55 rainfall stations. HCPC algorithm, which is a blend of Ward's classification method with the K-means algorithm and PCA approach, is investigated and used for the regionalization of drought across the South Korean region. The clusters formed by HCPC algorithm were further validated using four cluster validity indices, connectivity, silhouette width, Dunne index and Calinski and Harabasz index. Statistical homogeneity of the region was tested using a newly extended approach based on L-moment bivariate discordancy and heterogeneity measures. The regionalization of drought is accomplished using the SPI time scale of 1, 3, 6, 12, 24 months. The primary conclusions determined from this study are as follows:

(1) PCA method applied on the selected attributes indicates that almost all selected attributes contributed significantly to the leading principal components because of the existence of a significant correlation between them.

(2) Although the mixed nature of HCPC clustering algorithm helps to increase the robustness in the partitioning of the region, a comprehensive understanding of hydro-climatic and topological characteristics across the region is necessary for deciding number of clusters.

(3) HCPC cluster validation indices such as connectivity, silhouette width, Dunne index and Calinski and Harabasz index are found effective in identifying the optimal partitions and helped to reduce the subjective adjustment in the formation of clusters. Validation indices indicate that the 55 stations across South Korea should be divided into four regions that are closer to being homogeneous.

(4) The clusters formed by the HCPC algorithm is found very useful in adjusting the regions to improve their homogeneity. It is suggested that the clusters formed by the HCPC algorithm should be considered as primary clusters, which could be easily modified with relatively less effort to make them homogeneous.

(5) Univariate and bivariate homogeneity and discordancy tests applied on drought variables (duration and severity) showed the significant difference in their results. This is because of the reason that the univariate homogeneity and discordancy measure can use only one variable (either duration or severity) at a time to modify clusters, and unable to take account of the correlation between drought variables. Compared to univariate L-moment, bivariate L-comoment is better able to model drought events described by their duration and severity. This fact is particularly important when bivariate drought frequency analysis is a matter of interest.

(6) Regionalization of drought for SPI time scale of 1, 3, 6, 12, and 24 months showed that the variation in SPI time scale has a direct impact on the discordancy and homogeneity of the region because of the variation in length and frequency of dry and humid periods.

Acknowledgments: This research was supported by a grant (17RDRP-B134213-05) from Regional Development Research Program funded by Ministry of Land, Infrastructure and Transport of Korean government.

Author Contributions: Muhammad Azam designed, carried out the analysis and wrote the paper. Hyung Keun Park reviewed and edited the manuscript. Hyung San Kim and Seung Jin Maeng provided assistance in the calculations.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chung, Y.-S.; Yoon, M.-B.; Kim, H.-S. On Climate Variations and Changes Observed in South Korea. *Clim. Chang.* **2004**, *66*, 151–161. [CrossRef]
2. Azam, M.; Kim, H.S.; Maeng, S.J. Development of flood alert application in Mushim stream watershed Korea. *Int. J. Disaster Risk Reduct.* **2017**, *21*, 11–26. [CrossRef]
3. Kim, H.; Muhammad, A.; Maeng, S.-J. Hydrologic Modeling for Simulation of Rainfall-Runoff at Major Control Points of Geum River Watershed. *Procedia Eng.* **2016**, *154*, 504–512. [CrossRef]
4. Dai, A.; Trenberth, K.E.; Karl, T.R. Global variations in droughts and wet spells: 1900–1995. *Geophys. Res. Lett.* **1998**, *25*, 3367–3370. [CrossRef]
5. Min, S.K.; Kwon, W.T.; Park, E.H.; Choi, Y. Spatial and temporal comparisons of droughts over Korea with East Asia. *Int. J. Climatol.* **2003**, *23*, 223–233. [CrossRef]
6. Smith, A.B.; Katz, R.W. US billion-dollar weather and climate disasters: Data sources, trends, accuracy and biases. *Nat. Hazards* **2013**, *67*, 387–410. [CrossRef]
7. Um, M.-J.; Yun, H.; Jeong, C.-S.; Heo, J.-H. Factor analysis and multiple regression between topography and precipitation on Jeju Island, Korea. *J. Hydrol.* **2011**, *410*, 189–203. [CrossRef]
8. Singh, P.K.; Kumar, V.; Purohit, R.C.; Kothari, M.; Dashora, P.K. Application of Principal Component Analysis in Grouping Geomorphic Parameters for Hydrologic Modeling. *Water Resour. Manag.* **2009**, *23*, 325–339. [CrossRef]
9. Richman, M.B. Rotation of principal components. *J. Climatol.* **1986**, *6*, 293–335. [CrossRef]
10. Lee, J.Y.; Yang, J.S.; Kim, D.K.; Han, M.Y. Relationship between land use and water quality in a small watershed in South Korea. *Water Sci. Technol.* **2010**, *62*, 2607–2615. [CrossRef] [PubMed]

11. Kadiolu, M. Regional variability of seasonal precipitation over Turkey. *Int. J. Climatol.* **2000**, *20*, 1743–1760. [CrossRef]
12. Basu, B.; Srinivas, V.V. Regional flood frequency analysis using kernel-based fuzzy clustering approach. *Water Resour. Res.* **2014**, *50*, 3295–3316. [CrossRef]
13. Rossi, G.; Benedini, M.; Tsakiris, G.; Giakoumakis, S. On regional drought estimation and analysis. *Water Resour. Manag.* **1992**, *6*, 249–277. [CrossRef]
14. Basist, A.; Bell, G.D.; Meentemeyer, V. Statistical Relationships between Topography and Precipitation Patterns. *J. Clim.* **1994**, *7*, 1305–1315. [CrossRef]
15. Yoo, J.; Kwon, H.-H.; Kim, T.-W.; Ahn, J.-H. Drought frequency analysis using cluster analysis and bivariate probability distribution. *J. Hydrol.* **2012**, *420–421*, 102–111. [CrossRef]
16. Lee, J.J.; Kwon, H.H.; Kim, T.W. Spatio-temporal analysis of extreme precipitation regimes across South Korea and its application to regionalization. *J. Hydro-Environ. Res.* **2012**, *6*, 101–110. [CrossRef]
17. Chang, H.; Kwon, W.-T. Spatial variations of summer precipitation trends in South Korea, 1973–2005. *Environ. Res. Lett.* **2007**, *2*, 45012. [CrossRef]
18. Bae, D.H.; Jung, I.W.; Chang, H. Long-term trend of precipitation and runoff in Korean river basins. *Hydrol. Process.* **2008**, *22*, 2644–2656. [CrossRef]
19. Hosking, J.R.M.; Wallis, J.R. *Regional Frequency Analysis: An Approach Based on L-Moments*; Cambridge University Press: New York, NY, USA, 1997; ISBN 9780521019408.
20. Serfling, R.; Xiao, P. A contribution to multivariate L-moments: L-comoment matrices. *J. Multivar. Anal.* **2007**. [CrossRef]
21. Chebana, F.; Ouarda, T. Multivariate L-moment homogeneity test. *Water Resour. Res.* **2007**, *43*. [CrossRef]
22. Jung, I.W.; Bae, D.H.; Kim, G. Recent trends of mean and extreme precipitation in Korea. *Int. J. Climatol.* **2011**, *31*, 359–370. [CrossRef]
23. Helsel, D.R.; Hirsch, R.M. *Statistical Methods in Water Resources*; Elsevier: Amsterdam, The Netherlands, 1992; Volume 49, pp. 323–352.
24. McLeod, A.A.I. Kendall: Kendall Rank Correlation and Mann-Kendall Trend Test. R Package Version 2.2. 2011. Available online: https://CRAN.R-project.org/package=Kendall (accessed on 2 October 2017).
25. Pohlert, T. Trend: Non-Parametric Trend Tests and Change-Point Detection. R package Version 1.0.1. 2017. Available online: https://CRAN.R-project.org/package=trend. (accessed on 2 October 2017).
26. Hipel, K.W.; McLeod, A.I. *Time Series Modelling of Water Resources and Environmental Systems*; Elsevier: Amsterdam, The Netherlands, 1994; Volume 45.
27. Choi, G.; Kwon, W.-T.; Boo, K.-O.; Cha, Y.-M. Recent Spatial and Temporal Changes in Means and Extreme Events of Temperature and Precipitation across the Republic of Korea. *J. Korean Geogr. Soc.* **2008**, *43*, 681–700.
28. Im, E.S.; Jung, I.W.; Bae, D.H. The temporal and spatial structures of recent and future trends in extreme indices over Korea from a regional climate projection. *Int. J. Climatol.* **2011**, *31*, 72–86. [CrossRef]
29. Kim, J.-S.; Jain, S. Precipitation trends over the Korean peninsula: Typhoon-induced changes and a typology for characterizing climate-related risk. *Environ. Res. Lett.* **2011**, *6*, 34033. [CrossRef]
30. Vogel, R.M.; Stedinger, J.R. Minimum variance streamflow record augmentation procedures. *Water Resour. Res.* **1985**, *21*, 715–723. [CrossRef]
31. Mckee, T.B.; Doesken, N.J.; Kleist, J. The relationship of drought frequency and duration to time scales. In Proceedings of the AMS 8th Conference Applied Climatology, Anaheim, CA, USA, 17–22 January 1993; Volume 17, pp. 179–184.
32. Maeng, S.J.; Azam, M.; Kim, H.S.; Hwang, J.H. Analysis of Changes in Spatio-Temporal Patterns of Drought across South Korea. *Water* **2017**, *9*, 679. [CrossRef]
33. Edwards, D.C.; McKee, T.B. Characteristics of 20th Century Drought in the United States at Multiple Time Scales. Master's Thesis, Colorado State University, Fort Collins, CO, USA, May 1997.
34. Guttman, N.B. Comparing the Palmer Drought Index and the Standardize Precipitation Index. *J. Am. Water Resour. Assoc.* **1998**, *34*, 113–121. [CrossRef]
35. Seiler, R.A.; Hayes, M.; Bressan, L. Using the standardized precipitation index for flood risk monitoring. *Int. J. Climatol.* **2002**, *22*, 1365–1376. [CrossRef]
36. Lee, J.-H.; Seo, J.-W.; Kim, C.-J. Analysis on Trends, Periodicities and Frequencies of Korean Drought Using Drought Indices. *J. Korea Water Resour. Assoc.* **2012**, *45*, 75–89. [CrossRef]
37. Thorn, H.C.S. Some methods of climatological analysis. *WMO Tech.* **1966**, *81*, 16–22.

38. Kim, C.J.; Park, M.J.; Lee, J.H. Analysis of climate change impacts on the spatial and frequency patterns of drought using a potential drought hazard mapping approach. *Int. J. Climatol.* **2014**, *34*, 61–80. [CrossRef]
39. Lee, J.H.; Kim, C.J. A multimodel assessment of the climate change effect on the drought severity-duration-frequency relationship. *Hydrol. Process.* **2013**, *27*, 2800–2813. [CrossRef]
40. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417–441. [CrossRef]
41. Argüelles, M.; Benavides, C.; Fernández, I. A new approach to the identification of regional clusters: Hierarchical clustering on principal components. *Appl. Econ.* **2014**, *46*, 2511–2519. [CrossRef]
42. Husson, F.; Josse, J.; Pages, J. *Principal Component Methods-Hierarchical Clustering-Partitional Clustering: Why Would We Need to Choose for Visualizing Data*; Agrocampus Ouest: Rennes, France, 2010.
43. Handl, J.; Knowles, J.; Kell, D.B. Computational cluster validation in post-genomic data analysis. *Bioinformatics* **2005**, *21*, 3201–3212. [CrossRef] [PubMed]
44. Rousseeuw, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
45. Dunn, J.C. Well-Separated Clusters and Optimal Fuzzy Partitions. *J. Cybern.* **1974**, *4*, 95–104. [CrossRef]
46. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **1974**, *3*, 1–27.
47. Hosking, J.R.M. L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *J. R. Stat. Soc. B* **1990**, *52*, 105–124.
48. Hosking, J.R.M.; Wallis, J.R. Some statistics useful in regional frequency analysis. *Water Resour. Res.* **1993**, *29*, 271–281. [CrossRef]
49. Dai, A. Global precipitation and thunderstorm frequencies. Part II: Diurnal variations. *J. Clim.* **2001**, *14*, 1112–1128. [CrossRef]
50. Dinpashoh, Y.; Fakheri-Fard, A.; Moghaddam, M.; Jahanbakhsh, S.; Mirnia, M. Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods. *J. Hydrol.* **2004**, *297*, 109–123. [CrossRef]
51. Burn, D.H. Delineation of groups for regional flood frequency analysis. *J. Hydrol.* **1988**, *104*, 345–361. [CrossRef]
52. Burn, D.H. Cluster analysis as applied to regional flood frequency. *J. Water Resour. Plan. Manag.* **1989**, *115*, 567–582. [CrossRef]
53. Hayes, M.J.; Svoboda, M.D.; Wilhite, D.A.; Vanyarkho, O.V. Monitoring the 1996 Drought Using the Standardized Precipitation Index. *Bull. Am. Meteorol. Soc.* **1999**, *80*, 429–438. [CrossRef]
54. Sims, A.P. Adopting drought indices for estimating soil moisture: A North Carolina case study. *Geophys. Res. Lett.* **2002**, *29*, 13343. [CrossRef]
55. Ji, L.; Peters, A.J. Assessing vegetation response to drought in the northern Great Plains using vegetation and drought indices. *Remote Sens. Environ.* **2003**, *87*, 85–98. [CrossRef]
56. Ramachandra Rao, A.; Srinivas, V.V. Some problems in regionalization of watersheds. *Water Availab. Glob. Chang.* **2003**, *280*, 301–308.