한국방재학회논문집
제12권 5호 2012년 10월
pp. 233 ~ 240

강수지역 구분을 위한 최적 자료 전처리 기법 분석

The Optimal Analysis of Data Preprocessing Method for Clustering the Region of Precipitation

김억기* · 안원식** · 이채영*** · 엄명진****
Kim, Ug-Gi · Ahn, Won-Sik · Lee, Chae-Young · Um, Myoung-Jin

Abstract

In this study, the data preprocessing methods were analyzed to obtain the optimal clustering solution in South Korea. The geographic data and weather data in 75 stations of Korea Meteorological Administration are applied. The applied data preprocessing methods are general normalization, modified normalization, standardization and factor analysis. After the clustering analysis were conducted by K-mean method with preprocessing data, the efficiency of data preprocessing methods are estimated using the clustering index, such as Dunn index and Silhouette index. The clustering analysis are carried out as the cluster number changes from 3 to 9. Among the data preprocessing methods, the data by factor analysis shows the best efficiency for clustering analysis. However, it is not enough to find the optimal cluster number.

Key words : Data preprocessing method, Clustering analysis, Clusteing index, Factor analysis

요 지

본 연구에서는 우리나라 강수지역 구분을 위한 군집해석시 최적 자료 전처리 기법에 대하여 파악하고자 하였다. 이를 위하여 전국 기상청 관할의 75개 관측소의 지형 및 기상자료를 활용하였다. 적용된 자료 전처리 기법은 4가지로 일반 정규화 방법, 수정 정규화 방법, 표준화 방법 및 요인분석이다. 전처리된 자료를 K-means 군집분석을 통하여 군집을 구분한 후 유효성 측도인 Dunn 지수 및 Silhouette 지수를 통하여 효율성을 분석하였다. 군집수를 3개에서 9개까지 1개씩 늘려가며 분석한 결과 모든 경우에서 요인분석을 통한 자료가 최적의 효율성을 나타내었으나, 최적 군집개수의 산정에는 다소 부족한 것으로 나타났다.

핵심용어 : 자료 전처리 기법, 군집분석, 유효성 측도, 요인분석

1. 서 론

우리나라의 지형은 전국에 걸쳐 다양한 특성을 나타내고 있다. 따라서 지형 특성에 영향을 받는 강수 또한 지역별로 다른 특성을 보이고 있다. 이러한 지역별 강수의 특성을 해석하는 것은 수문학적으로 매우 중요한 개념이다. 지역별 강수 특성에 맞추어 관련된 여러 가지 수문 및 수리 계획이 진행되기 때문이다. 또한, 최근에 지점 빈도해석과 비교되고 있는 지역빈도 해석은 강수의 지역구분이 선행되어야 추후 분석을 수행할 수 있다.

지역빈도해석은 지점 빈도해석의 단점인 자료수의 제한 조건을 어느 정도 해소할 수 있는 개념으로 Hosking and Wallis(1997)에 의해 정립된 후 여러 연구에 적용되고 있는 빈도해석 방법이다(이동진, 허준행, 2001; 허준행 등, 2004; Cunnane, 1989). 이러한 지역 빈도 해석을 수행하기 위해서는 앞서 말한 바와 같이 강수지역의 구분이 선행되어야 하는데 이러한 동질 강수지역을 구분하기 위한 정립된 개념은 아직 미미한 편이다(엄명진 등, 2011). 따라서 동질 강수지역을 구분하기 위한 다양한 방법들이 시도되고 있는데 국내연구 동향은 문영수(1990)가 다양한 군집해석을 통하여 한국의 강수지역을 구분하였으며 그 중에서 Ward 군집해석을 통한 분석이 적합다고 주장하였다. 이순혁 등(2001)은 장기강우특성 및 지리특성을 K-mean 군집해석에 적용하여 강수지역을 구분하였으며, 강우관측지점의 연평균강우량의 변동계수 등을 이용하여 동질성을 판단하였다. 이순혁 등(2003)은 지역빈도 분석을 수행하기 위하여 지역구분을 수행하였으며 대상 관측

*정회원 · 수원대학교 대학원 토목공학과 박사과정, 경기도청 교통건설국장(E-mail : ukkim@gg.go.kr)
**정회원 · 수원대학교 토목공학과 명예교수
***수원대학교 토목공학과 교수
****정회원 · 연세대학교 토목공학과 박사, 일리노이대학교 대기과학과 박사후과정(교신저자)

소로는 57개 강우관측소를 적용하였으며, 강수특성 및 지형특성을 K-mean 군집해석을 통하여 분석한 결과 강수지역을 5개로 구분하였다. 고정웅 등(2005)는 인자분석과 Ward 군집해석을 통하여 한반도 우기의 강수 특성 및 지역구분을 수행하였다. 남우성 등(2008)은 Procrustes 분석, 요인분석 및 Fuzzy-c means 기법을 적용하여 우리나라 강수지역을 6개 지역으로 구분하였으며, 이병주 등(2009)은 미계측유역에 대하여 준분포형 강우-유출모형을 적용하기 위한 방법으로 주성분분석과 계층적 군집분석을 통하여 매개변수의 지역화 방법을 제안하였으며, 박수완 등(2009)은 상수관로 누수위치 자료를 통한 계층적 군집분석을 수행하여 상수관망 유지관리 우선순위를 결정하였으며, 유지영 등(2010)은 가뭄특성 인자(가뭄심도, 지속기간, 강도, 발생빈도 등)를 대상으로 K-means 기법을 적용하여 가뭄지역을 6개 지역으로 구분하였다. 그리고 국외연구 동향을 살펴보면 Mallants and Feyen(1990)은 일강우량 자료를 대상으로 다변량 해석기법인 주성분 분석을 수행하여 강우지역을 구분하였고, Guttman(1993)은 다양한 지형 및 기후 변수들을 활용하여 지역을 구분하였으며, Zhang and Hall(2004)은 여러 가지 군집분석을 통하여 지역을 구분하여 지역홍수빈도해석에 활용하였으며, Dinpashoh et al.(2004)은 대상 자료의 다양성에 대한 대처방법으로 Procrustes analysis 및 요인분석 등을 활용하여 동질 지역 구분을 도모하였다. 하지만 이러한 연구동향 중 군집분석을 위한 대상자료의 전처리에 대하여는 연구를 거의 수행하지 않았다.

따라서 본 연구에서는 강수자료의 대표적인 전처리 기법으로 대표적인 4가지 방법을 제시 및 적용한 후 강수의 군집분석에 적용하고 군집지수를 산정하여 최적의 전처리 기법을 판단하고자 한다. 대상지역으로는 우리나라 전역을 설정하였으며, 대상자료는 기상청 산하의 75개 관측소 자료를 활용하였다.

2. 강수지역 구분

강수지역을 구분하기 위하여 다양한 기후 및 지형자료를 이용한 군집해석 방법을 이용하고 있다. 하지만 군집해석을 하기 전에 자료의 전처리를 통하여 자료 간의 중요도 및 크기 등을 일치시켜야 한다. 따라서 여러 가지 자료의 전처리 기법을 도입하고 있는 데, 본 연구에서는 그 중 대표적으로 사용하고 있는 4가지 자료 전처리 기법을 적용하여 군집해석을 수행하고자 한다. 군집해석 방법으로는 널리 사용되고 있는 K-means 방법을 선택하였으며, 자료 전처리 기법의 효율성을 판단하기 위한 척도로는 유효성 측도들 중 Dunn 지수 및 Silhouette 지수를 선택하였다. 이러한 기법들을 다음과 같이 정리하였다.

2.1 강수자료 전처리 기법

2.1.1 일반 정규화 방법

자료를 정규화 하기 위해서는 일반적으로 단위 1사이에 자료를 재정렬시킨다. 일부 연구에서는 0.1과 0.9사이에 자료를 재배치시키기도 하지만 본 연구의 대상자료를 0과 1사이에 분포시켰다. 이러한 재배열을 위한 정규화 방법은 Eq. (1)과 같다.

$$X_{i,0to1} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

여기서 X_i는 대상 자료이고, $X_{\min}$는 대상자료의 최소값이며, $X_{\max}$는 대상자료의 최대값이다.

2.1.2 수정된 정규화 방법

일반적으로 정규화방법은 0과 1사이로 자료를 재배치하는 것을 의미하지만, 다양한 연구에서 좀 더 자료를 중앙으로 분포시키고자 자료를 -1과 1사이에 재배열시켰으며, 이 때 중앙 기준점을 0으로 가정하였다. 이러한 수정된 정규화 방법을 수식으로 표현하면 Eq. (2)와 같다.

$$X_{i,-1to1} = 2\frac{X_i - X_{\min}}{X_{\max} - X_{\min}} - 1 \tag{2}$$

여기서 X_i는 대상 자료이고, $X_{\min}$는 대상자료의 최소값이며, $X_{\max}$는 대상자료의 최대값이다.

2.1.3 Z-score 방법

자료를 표준화시키기 위해서 일반적으로 적용되는 방법이 Z-Score 방법이다. 표준화를 통해서 대상자료는 평균이 0, 분산이 1인 자료로 재정렬된다. 이러한 방법은 서로 다른 자료들을 비교하기 위해서 제안된 방법으로 평균으로부터 어느 정도 떨어져 있는지를 쉽게 알 수 있으며 여러 연구에서 널리 사용되는 방법이다. 이러한 자료의 표준화 방법은 Eq. (3)으로 나타낼 수 있다.

$$X_{i,1\sigma} = \frac{X_i - \overline{X}}{\sigma_x} \tag{3}$$

여기서 X_i는 대상 자료이고, $\overline{X}$는 대상자료의 평균이며, σ_x는 대상자료의 표준편차이다.

2.1.4 요인분석

요인분석은 대상 변수들간의 상관관계를 이용하여 서로 유사한 요인들끼리 묶어주는 해석방법이다. 여러개의 변수를 몇개의 공통된 요인으로 묶어줌으로써 대상 자료들을 요약하는데 이용된다. 또한, 중요하지 않은 변인들을 제거함으로써 신뢰도를 높일 수 있다. 요인들을 추출하는 방법에는 주성분 분석과 CFA(Common Factor Analysis)가 있다. 이러한 방법들을 적용 후 적정 요인을 결정하는 방법으로는 주로 아이겐 값(Eigen Value)를 적용하는데 일반적으로 1이상의 요인들만을 추출하여 분석에 이용한다. 요인들을 보다 명확히 하기위하여 요인의 회전을 수행하는데 요인의 회전 방법으로는 직각회전 방식과 비직각회전 방식이 있다. 직각회전으로는 Quartimax, Varimax 및 Equamax 회전 등이 있다. 본 연구에서는 요인추출은 주성분 분석법, 요인회전은 기후자료에 적합하다고 알려진 Varimax rotation(Overall and Klett, 1973; Puvaneswaran, 1990; White et al., 1991)을 적용하였다.

2.2 K-means 군집분석

군집 분석은 대상 개체들이 지니고 있는 다양한 속성의 유사성을 동질적인 집단으로 군집화하는 방법을 말한다. 군집분석 기본원리는 분석하고자 하는 여러 특성들을 유사성(Similarity) 거리(Distance)로 환산하고 거리가 상대적으로 가까운 개체들을 동질적으로 군집화 하는 것이다(박상우 등, 2003). 여러 군집방법 중 K-means 군집은 가장 가까운 평균을 가진 군집에 속해 있는 각각의 관측치안에 n개의 관측치들을 k개의 군집으로 분할하려는 목적을 가진 군집분석 방법이다. 이러한 K-means 알고리즘(McQueen, 1967)은 주어진 자료에 안에서 군집을 확인할 때 흔히 사용되는 방법이다. 각각의 관측치들이 d차원의 실제 벡터인 관측치들(x_1, x_2,..., x_n)이 주어졌을 경우, K-means 군집은 n개의 관측치들을 k개의 집합($k \le n$)으로 분할하는데 목적이 있다. 또한 군집간 제곱합을 Eq. (4)과 같이 최소하여야 한다.

$$\arg min_s \sum_{i=1}^{k} \sum_{x_j \in S_i} |x_j - \mu_i|^2 \tag{4}$$

여기서, μ_i는 S_i상의 자료들의 평균이다.

2.3 유효성 측도

원 자료로서 군집 결과를 분석하여 자료 본래의 정보를 사용하여 군집화가 잘 되었는지를 판단하는 측도를 일반적으로 내부유효성 측도라 하며, Dunn Index(Dunn, 1974) 및 Silhouette 지수(Rousseeuw, 1987) 등이 있다.

2.3.1 Dunn 지수

Dunn 지수는 같은 군집에 속해 있는 두 개체간의 가장 큰 거리에 대한 서로 다른 군집에 속해 있는 두 개체간의 가장 작은 거리 비(ratio)를 나타내면 Eq. (5)와 같이 산정할 수 있다.

$$V_D = \min_{1 \le i \le k} \left\{ \min_{1 \le j \le k, j \ne i} \left[\frac{\delta(C_i, C_j)}{\max_{1 \le k \le K} \Delta(C_k)} \right] \right\} \tag{5}$$

여기서 $\delta(C_i, C_j)$은 군집 C_i와 C_j의 거리를 나타내며, $\Delta(C_k)$는 C_k에 대한 군집간 거리를 말한다. 같은 군집에 속해 있는 두 개체간의 거리가 작을수록 다른 군집에 속해있는 두 개체간의 거리가 클수록 Dunn 지수는 커지므로 Dunn 지수가 클수록 군집이 잘 이루어졌다고 판단할 수 있다.

2.3.2 Silhouette 지수

Silhouette 지수는 각 군집들의 평균 Silhouette 폭과 전체 자료에 대한 총 평균 Silhouette 폭을 산정하여 구한다. 이는 군집에 대한 군집내 밀집과 군집간 분리 정도를 말하며 최적 군집수를 결정하는 지표로 적용가능하다. 임의의 j번째 군집 $C_j(j = 1,...,k)$에 대하여 군집내 i번째 개체 X_i의 구성요소의 신뢰성 지표로 Silhouette 폭은 각각의 개체에 대하여 Eq. (6)과 같이 산정한다.

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \tag{6}$$

여기에서 $a(i)$는 j번째 군집내의 i번째 개체 X_i와 이를 제외한 모든 개체들 사이의 평균거리이고, b_i는 j번째 군집내의 i번째 개체 X_i와 j번째 군집과 제일 가까운 군집내의 개체들간의 평균거리이다. Silhouette 폭은 -1부터 1사이의 값을 가지며 적절하게 1에 가까울수록 군집화가 잘 이루어 진 것이며 -1에 가까울수록 군집화가 잘 이루어지지 못한 것이다.

3. 대상지역 및 강수특성

본 연구에서는 전국의 75개 기상관측소를 대상으로 수행하였다. 대상관측소의 자료 중 2010년까지의 강수자료를 대상으로 시간적 연속성을 확보한 자료에 대하여서만 분석을 수행하였으며, 최소 관측기간은 9년, 최대 관측기간은 30년이다. 본 연구의 대상관측소는 Fig. 1에 도시하였으며, 각 지점별 관측소 번호를 명시하였다. Table 1에는 각 관측소 명과 위치정보 등을 나타내었다.

이러한 대상관측소들에 대해서 강수지역 구분을 위한 강수특성을 추출하였다. 강수특성은 지형정보를 제외하면 총 8개의 특성을 분석하였다. 분석된 특성들은 관측기간 동안의 일최대값(MMD), 월최대값(MMM), 연최대값(MMY), 우기시 최대값(MMS) 및 연평균 일최대값(AMD), 월최대값(AMM), 연최대값(AMY), 우기시 최대값(AMS)이다. 여기서 우기는 6월에서 9월의 강수량을 뜻한다. 추출된 값들은 Table 2에 정리하였다. 산정된 값들 중 MMY 및 AMY를 보면 MMY는 평균 약 2030 mm의 값을 갖고 최소 1162.9 mm, 최대 3397.4 mm를 나타내었다. AMY는 평균 약 1301 mm를 가지며 최소 779.8 mm, 최대 1895.5 mm의 값을 나타내었다. 따라서 지역에 따라서 MMY는 최대 2235.5 mm, AMY는 최대 1115.7 mm의 강수량의 차이를 보이므로 지역별 강수의 차이가 확연히 존재함을 알 수 있다.

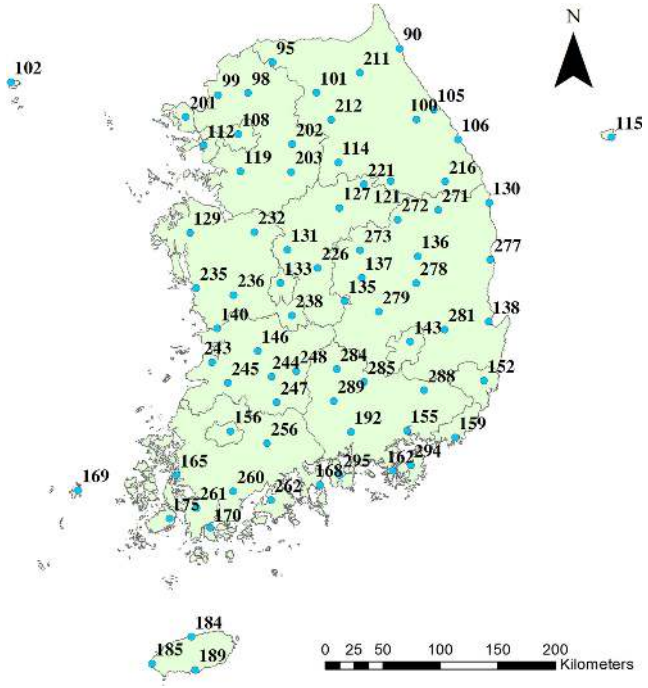


Fig 1. Map of precipitation stations in South Korea

Table 1. Summary of precipitation stations

Station	Name	TMX	TMY	Elevation	Duration
90	Sokcho	336696.5	529000.2	22.9	30
95	Cheorwon	226408.4	516457.7	154.9	23
98	Dongducheon	205083.7	489114.1	112.5	13
99	Munsan	179207.6	487363.1	30	9
100	Daegwallyeong	351314.3	465559.9	772.43	30
101	Chuncheon	264442.2	489443.4	76.8	30
102	Baengnyeongdo	-8265.07	498888.6	145.5	10
105	Gangneung	366383.3	474088.6	26.1	30
106	Donghae	387557.6	447412.6	39.5	18
108	Seoul	196723.8	452434.4	85.5	30
112	Incheon	166570	442091.3	69	30
114	Wonju	283619.5	426894.5	150.7	38
115	Ulleungdo	544621.9	449537.5	220	30
119	Suwon	198644	418986.4	34.5	30
121	Yeongwol	329157.7	410123.1	239.7	16
127	Chungju	284577	386139.7	113.7	30
129	Seosan	154752.7	364033.6	25.2	30
130	Uljin	414534	390830.7	49.4	30
131	Cheongju	239338.2	348757.9	56.4	30
133	Daejeon	233327.1	319073.2	62.6	30
135	Chupungnyeong	289148.9	302933	240.9	30
136	Andong	352544.8	342996.3	140.7	28
137	Sangju	303746.6	323658.3	98	9
138	Pohang	414206.9	284310.8	1.3	30
140	Gunsan	178400.8	278339	26.9	30
143	Daegu	345914	266538.2	57.3	30
146	Jeonju	213744	258275.2	61	30
152	Ulsan	410094.1	231736.1	34.6	30
155	Changwon	343220.8	186798.7	36.8	25
156	Gwangju	189863.1	186304.8	74.5	32
159	Busan	384988.6	180628.9	69.2	30
162	Tongyeong	331024	150928.2	30.8	30
165	Mokpo	143127.3	146995.7	37.4	30
168	Yoesu	267549.7	138459.6	73.3	35
169	Heuksando	57804.55	133528.8	68.5	14
170	Wando	172316.8	100162.6	27.7	38
175	Jindo	136962.2	108188.2	476.4	9
184	Jeju	156042.2	2432.504	19.97	30
185	Gosan	121760.3	-21795.2	70.9	23
189	Seogwipo	159224.4	-27314.5	50.4	34
192	Jinju	294477.3	185793.2	27.1	30
201	Ganghwa	150951.8	467672.4	46.2	30
202	Yangpyeong	243472.4	443358.5	47.4	30
203	Icheon	242688.6	418429	90	30
211	Inje	302152.7	507298.9	198.7	30
212	Hongcheon	277395.3	465247.7	146.2	30
216	Taebaek	376406.8	409781.7	714.2	25
221	Jecheon	305817.1	407366.5	263.1	30
226	Boeun	265513	332417	173	30
232	Cheonan	210568.9	364575	21.3	30
235	Boryeong	160004.2	314463.4	17.9	30

Table 1. Summary of precipitation stations (Continued)

Station	Name	TMX	TMY	Elevation	Duration
236	Buyeo	192624.9	308294.2	11	30
238	Geumsan	243108.6	289890.6	170.6	30
243	Buan	174102.2	248095.6	3.6	30
244	Imsil	225602.2	235092.3	248	30
245	Jeongeup	187600.3	229605.9	39.5	30
247	Namwon	229982.2	212141.3	93.5	30
248	Jangsu	246847.4	240137.8	407	23
256	Suncheon	221540.9	175475.6	74.4	30
260	Jangheung	192359.6	132602.2	44.5	30
261	Haenam	160181.6	117687.9	4.6	30
262	Goheung	225016.5	124812.9	53.3	30
271	Bongwhoa	370269.7	384473.9	320.9	23
272	Yeongju	334978.7	375869.4	210.5	30
273	Mungyeong	302484.7	348280.7	170.8	30
277	Yeongdeok	415502.4	339934.2	41.2	30
278	Uiseong	351292.2	318886.3	82.6	30
279	Gumi	318602.4	293362.4	47.4	30
281	Yeongcheon	375724.5	277316.7	93.3	30
284	Geochang	282208.3	241969.8	221.4	30
285	Hapcheon	305786.5	230434.7	33	30
288	Miryang	357984.7	223048.1	10.7	30
289	Sancheong	279575.1	213298.1	138.7	30
294	Geoje	346395.8	155899.1	44.5	30
295	Namhae	284484.4	147175.1	43.2	30

Table 2. Summary of precipitation characteristics

Station	MMD	MMM	MMY	MMS	AMD	AMM	AMY	AMS
90	293	608.9	1894.1	1286.5	146.3	370.63	1301.64	807.89
95	263.1	709.5	1990.1	1538	139.48	471.5	1287.17	938.28
98	290	817.7	1813.7	1404.3	164.25	545.8	1414.83	1073.07
99	202	589.6	1646.5	1310.2	130.06	430.38	1289.27	941.42
100	690	1186.3	2808	1943.8	187.82	500.79	1717.89	1120.21
101	307.5	852.3	1913.3	1440.9	141.53	450.49	1249.16	904.89
102	142	361.9	1161.9	854.6	97.59	248.16	779.78	534.11
105	866	1115.5	1993.9	1466.1	174.23	404	1378.33	840.62
106	313.5	836.6	1825.7	1266.7	136.96	387.8	1199.06	761.5
108	304.1	1115	2169	1792.5	155.51	485.24	1342.83	973.7
112	255	705.4	1867.4	1430.1	136.6	385.66	1135.12	789.69
114	291.5	843.1	1965	1527.7	130.37	397.52	1231.2	865.27
115	257.8	582.2	2119.9	1039.9	98.57	262.27	1310.83	577.05
119	333.2	904.7	1906.7	1413.4	148.89	413.54	1224.88	864.21
121	209.5	727.9	1610.4	1240	119.69	384.4	1143.96	814.42
127	261.5	693	1779.3	1347.2	116.26	361.66	1142.44	797.39
129	274.5	940.6	1887.5	1428.5	125.96	364.84	1190.34	795.22
130	271.6	656.1	1661	1023.4	106.66	283.76	1050.61	635.57
131	247.5	696.3	1563.5	1148.7	112.22	338.92	1148.11	785.14
133	298.6	757.5	1940.8	1472.2	127.48	384.17	1278.45	879.37
135	277	670.4	1696.4	1143.7	111.87	333.63	1113.67	755.86
136	153	538.2	1442.6	944.6	92.97	288.28	998.31	676.97
137	187	587.7	1740.8	1170.2	112.39	368.59	1113.96	785.06
138	494.6	595.4	2035.1	1332.4	121.27	309.42	1087.74	701.06

Table 2. Summary of precipitation characteristics(Continued)

Station	MMD	MMM	MMY	MMS	AMD	AMM	AMY	AMS
140	307.1	714.8	1693.4	1404.1	110.97	326.1	1140.08	748.8
143	219.4	636.6	1639.4	1120.5	107.32	312.35	1007	702.21
146	205.5	656.6	1711.2	1336.8	114.06	360.87	1232.88	817.58
152	389.3	639.1	1946.1	1381.9	135.47	325.82	1197.25	757.78
155	268	780.9	2661.6	1950.4	142.95	404.74	1444.61	921.36
156	333.9	653.9	1909.1	1465.6	134.17	370.66	1313.48	876.47
159	419.2	850.9	2230.9	1658.8	146.53	382.39	1422.39	871.8
162	259.5	794.8	2555.1	1914	132.67	378.79	1427.41	867.91
165	394.7	584.2	1736.8	1179.8	122.01	323.9	1163.54	749.88
168	288	613	2451.4	1516.1	143.31	392.81	1426.96	908.61
169	281.1	624.8	1478.1	1098.6	124.01	299.74	1106.91	715.24
170	414.3	850.6	2646.4	1919.9	162.1	416.74	1507.19	915.58
175	235	618.9	2188.1	1209.5	142.94	423.36	1563.48	924.12
184	420	880	2526	2060.8	161.5	384.18	1496.67	905.42
185	240.3	575.5	1874.5	1368.8	119.5	284.85	1142.79	644.73
189	365.5	918.2	3244.3	2197.2	166.05	439.16	1895.47	1054.28
192	264	760.1	2192.7	1647.8	150.02	428.97	1506.17	1007.19
201	481	1069	2365.4	1920.9	175.5	488.9	1346.37	973.67
202	346	956.6	2254.9	1798.8	161.87	521.58	1417.06	1051.94
203	304	821	2313.3	1791.3	150.81	460.78	1355.38	977.26
211	275.8	886	1739.5	1296.3	136.51	426.27	1210.23	875.74
212	276	1244.5	2375	1906.5	146.13	503.1	1404.96	1032.77
216	338.5	811	1796.8	1285.5	139.47	401.87	1324.32	913.04
221	276.5	1111	2067.3	1596	143.24	456.74	1387.38	982.52
226	407.5	894	2085.1	1642	126.09	399.88	1297.08	905.75
232	273.5	1082.5	1785.8	1384.3	128.79	386.65	1227.69	854.49
235	361.5	996.5	1897.5	1446	133.24	374.66	1244.27	841.72
236	517.6	910.4	2137.5	1699	137.69	408.49	1349.08	921.08
238	202	681.5	1827.3	1250.7	120.92	395.89	1296.64	887.96
243	250.5	677	2074.1	1649	117.47	345.5	1250.3	823.03
244	208.5	731	1974.2	1402.5	117.03	407.69	1354.45	919.47
245	244.5	623.8	1917.3	1374.5	108.35	369.91	1317.16	857.2
247	218	768	2049.6	1408	115.8	410.76	1359.74	938.19
248	182.5	789.5	2208.6	1590.5	117.98	458.18	1459.24	995.03
256	254.5	879.5	2175	1610.7	157.4	446.58	1531.14	1055.17
260	547.4	809	2357.4	1550.4	163.45	434.16	1505.37	1008.13
261	477.5	724.2	2107.9	1500	152.5	373.4	1325.23	859.05
262	487.1	721.5	2484.7	1489.3	169.67	404.32	1453.26	926.46
271	250	748	1735.5	1219	133.09	373.1	1217.87	844.63
272	259.5	751.5	2018.9	1385.5	125.83	384.9	1290.57	894.49
273	173.3	671	1963.1	1354.2	107.57	385.39	1259.59	876.98
277	296	586.5	1841.2	1135.1	113.63	288.89	1072.78	682.33
278	215.5	738.5	1697	1171	100.88	321.88	1031.65	724.85
279	221	651	1749.8	1209	106.66	340.08	1072.42	751.48
281	190.1	642	1724	1199.5	104.03	307.07	1046.75	720.61
284	253	838	1957.5	1383.6	129.03	416.84	1316.64	926.51
285	288.5	814.5	1862.7	1356.5	135.43	398.74	1275.49	908.76
288	216.5	695	1879.6	1375.5	119.82	356.11	1229.27	829.57
289	332.5	854.5	2493.4	1808.8	174.27	475.09	1556.43	1107.9
294	438.3	1115.5	3397.4	2539.3	185.75	495.55	1840.33	1124.21
295	410	849	2843.8	1901	185.69	501.56	1839.32	1153.15

4. 적 용

본 연구에서는 강수지역 구분을 위하여 2절에 서술된 바와 같이 각 관측소별 추출된 강수특성 및 지형정보에 대하여 4가지 강수자료 전처리 기법(일반 정규화 방법, 수정 정규화 방법, Z-Score 방법, 요인분석)을 적용하였으며, 전처리된 자료를 K-means 군집분석을 수행하여 유효성 측도(Dunn 지수, Silhouette 지수)를 산정하여 비교 및 검토하였다.

4.1 강수자료 전처리

본 연구에서는 강수특성 8가지 및 지형정보 3가지(TMX. TMY, Eletevation)을 이용하여 강수자료 전처리를 수행하였다. 각 자료의 최소값, 최대값, 평균 및 표준편차를 이용하여 일반 정규화 자료, 수정 정규화 자료 및 표준화 자료를 산정하였다. 그리고 요인분석을 통하여 대상 자료의 요인을 추출하였다. 요인분석시 적합도 검정으로 Kaiser-Meyer-Olson(KMO) 검정을 수행하였으며, 산정된 검정값은 0.731을 나타내었다. Armstrong and Soelberg(1968)에 의하면 검정값이 0.42이하일 경우에는 요인분석에 적용하기 어렵다고 하였으며, Kaiser and Rice(1974)는 0.5이상의 값을 추천하였다. 따라서 본 연구의 0.731은 적합도 검정을 통과하였다고 판단 할 수 있다. 그리고 적정 요인의 수를 결정하는데 있어서 Eigen Value 1 이상을 적용하였다. Fig. 2와 같이 요인의 수가 4이상일 경우 Eigen Value가 1이상의 값을 갖는 것으로 나타나 본 연구에서는 적정 요인의 수를 4개로 결정하였다. 최종적으로 Varimax 회전을 적용한 후 산정된 요인 값들을 군집분석에 적용하였다.

4.2 군집분석 및 유효성 측도

본 연구에서 적용한 유효성 측도인 Dunn 지수 및 Silhouette 지수의 경우는 2절에서 설명한 바와 같이 산정 값이 클수록 군집이 잘 이루어졌다고 분석된다. 따라서 본 연구에서는 최적의 자료 전처리 방법을 찾기 위한 방법으로 임의로 군집의 수를 3개에서 9개의 범위까지 1개씩 늘려나가면서 K-means 군집분석을 수행하였다. 이렇게 수행된 군집분석 결과를 유효성 측도인 Dunn 지수 및 Silhouette 지수에 적용하여 자료 전처리 방법의 유효성을 측정하였다. 군집분석의 범위에 따라 산정된 유효성 측도는 Fig. 3에 도시하였다. 각각의 전처리 방법에 대하여 일반화 정규화 방법은 Method1, 수정 정규화 방법은 Method2, 표준화 방법은 Method3 그리고 요인분석은 Method4로 표현하였다.

유효성 측도의 산정된 결과를 검토해보면, Dunn 지수의 경우 요인분석의 경우 군집수 4개일 경우 최대값을 갖는 것으로 나타났으며, 다른 전처리 방법의 Dunn 지수 최대값 보다 크게 산정되었다. 요인분석 방법이외에서는 수정된 정규화 방법이 군집수 6개에서 최대값을 산정하였고, 일반 정규화 방법에서 군집수 7개에서 최대값을 나타내었으며, 표준화 방법은 군집수 3개의 경우 최대값을 산정하였다. 다만 표준화 방법이나 요인분석의 경우 군집수가 늘어남에 따라 Dunn 지수도 꾸준히 감소하는 것으로 나타나 최적군집수 판별에는 변별력이 없는 것으로 판단된다. 다음으로 Silhouette 지수의 경우를 해석해보면, 요인분석의 경우 군집수 8개에서 최대값을 나타내었으며, Dunn 지수의 경우와 마찬가지로 다른 전처

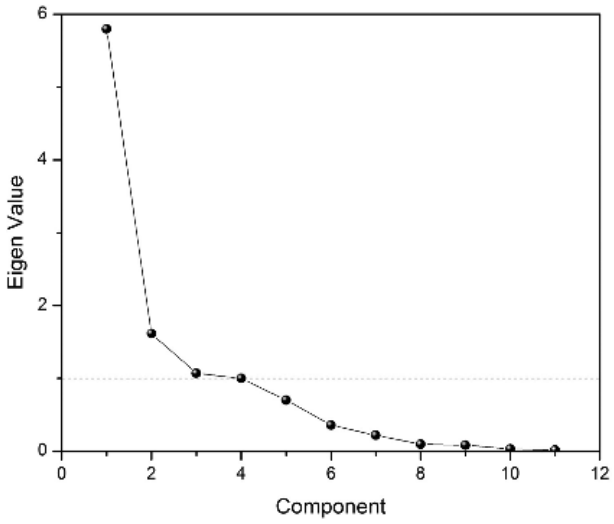
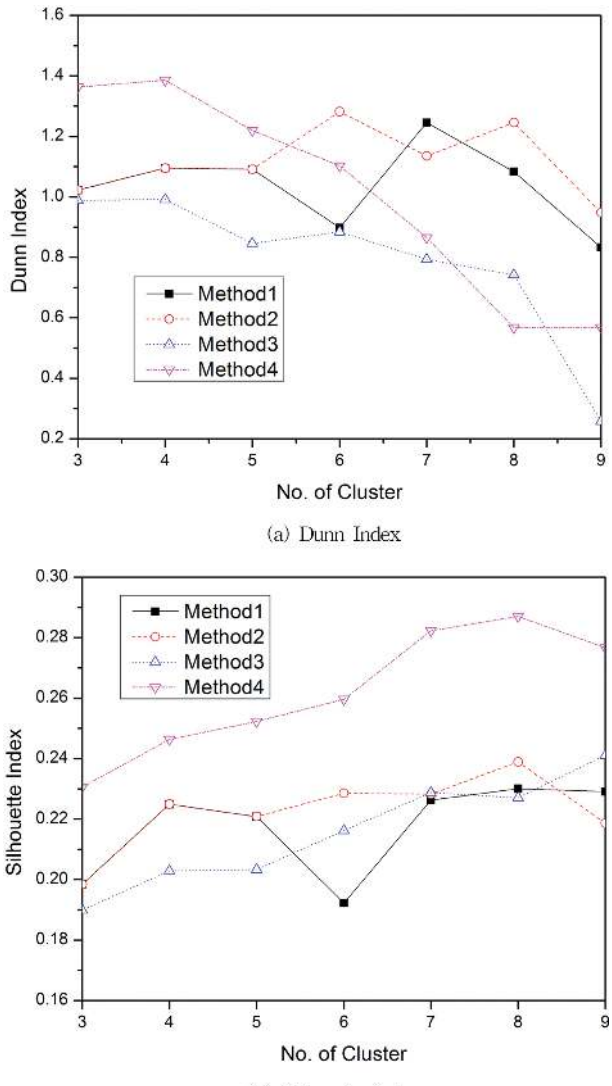


Fig 2. Scree plot



(a) Dunn Index

(b) Silhouette Index

Fig 3. The results of two clustering indexes

리 방법보다 큰 최대값을 산정하였다. 요인분석 방법이외에서는 수정된 정규화 방법이 군집수 8개에서 최대값을 산정하였으며 일반 정규화 방법도 군집수 8개에서 최대값을 나타내었다. 표준화 방법은 군집수 9개에서 최대값을 산정하였다. 다만, Silhouette 지수의 경우에서도 모든 전처리 방법에서 군집수가 늘어날 경우 Silhouette 지수도 증가하는 경향을 나타내어 최적 군집수를 찾아내는 것에는 다소 어려움이 있을 것으로 판단된다.

5. 결 론

본 연구에서는 우리나라의 강수지역을 구분하기 위하여 적합한 자료 전처리 기법을 찾고자 4가지의 기법을 적용하였으며, 전처리된 자료에 대하여 K-means 군집해석을 실시하였으며 기법의 효율성을 분석하기 위하여 유효성 측도를 산정하였다. 이를 위하여 우리나라의 기상청 산하 75개 관측소의 강우자료 및 위치정보를 활용하였다. 본 연구를 통하여 분석한 결과를 요약하면 다음과 같다.

1) K-means 군집해석결과를 적용한 유효성 측도 산정시 Dunn 지수 및 Silhouette 지수 모두 요인분석을 적용한 자료를 이용할 경우 유효성 측도가 가장 크게 산정되었다. 유효성 측도을 이용한 자료 전처리 기법의 효율성은 요인분석, 수정 정규화 방법, 일반 정규화 방법, 표준화 방법 순으로 나타났다.
2) 군집수를 3개에서 9개로 늘려가면서 K-means 군집해석을 수행하여 최적의 군집수를 결정할 경우 요인분석 및 표준화 방법의 경우 군집수가 증가할 때 일정 경향(감소 또는 증가)을 나타내어 적합한 군집개수를 판별하는데 어려움이 있는 것으로 나타났다.

따라서 본 연구에서는 강수지역의 구분을 위하여 적용한 4가지 자료 전처리 기법 중 요인분석을 통한 자료가 유효성 측도가 가장 높게 나타내어 군집해석에 가장 적합한 것으로 분석 되었지만 최적 군집 개수를 찾는 데에는 미흡한 것으로 판단되어 이에 대한 추후 연구가 필요할 것으로 판단된다.

참고문헌

고정웅, 백희정, 권원태 (2005) 한반도 우기의 강수특성과 지역구분, **한국기상학회지 논문집**, 제41권, 제5호, pp. 101-114.

남우성, 김태순, 신주영, 허준행 (2008) 다변량 분석 기법을 활용한 강우 지역빈도해석, **한국수자원학회논문집**, 제41권, 제5호, pp. 517-525.

문영수 (1990) 클러스터분석에 의한 한국의 강수지역 구분, **한국기상학회지 논문집**, 제26권, 제4호, pp. 203-215.

박상우, 전병호, 장석환 (2003) 다변량 분석기법에 의한 지점강우의 권역화 연구, **한국수자원학회논문집**, 제36권, 제5호, pp. 879-892.

박수완, 임광채, 최창록, 김규리 (2009) 상수관로 누수위치 자료를 이용한 계층적 군집분석, **한국수자원학회논문집**, 제42권, 제3호, pp. 177-190.

엄명진, 정창삼, 남우성, 정영훈, 허준행 (2011) Dunn 지수를 이용한 최적 강수지역 군집수 분석, **한국수자원학회 학술발표회 논문집**, pp. 87-91.

유지영, 최민하, 김태웅 (2010) 군집분석을 이용한 우리나라 가뭄 특성의 공간적 분석, **한국수자원학회논문집**, 제43권, 제1호, pp. 15-24.

이동진, 허준행 (2001) L-모멘트법을 이용한 한강유역 일강우량자료의 지역빈도해석, **한국수자원학회논문집**, 제34권, 제2호, pp. 119-130.

이병주, 정일원, 배덕효 (2009) 다변량 통계분석을 이용한 준분포형 유출모형 매개변수 지역화, **한국수자원학회논문집**, 제42권, 제2호, pp. 149-160.

이순혁, 박종화, 류경식, 지호근, 전택기, 신용희 (2001) 고차확률가중모멘트법에 의한 지역화빈도분석과 GIS기법에 의한 설계강우량 추정(I) -동질성의 지역구분 방법을 중심으로-, **한국농공학회 논문집**, 제43권, 제4호, pp. 57-68.

이순혁, 윤성수, 맹승진, 류경식, 주호길 (2003) L 및 LH-모멘트법과 지역빈도분석에 의한 가뭄우량의 추정(I) - L-모멘트법을 중심으로 -, **한국농공학회 논문집**, 제45권, 제5호, pp. 97-109.

허준행, 이영석, 남우성, 김경덕 (2004) 한강유역에 대한 강우지역빈도해석의 적용성 연구, **한국수자원학회 학술발표회논문집**, pp. 168-172.

Armstrong, J.S., Soelberg, P. (1968) On the interpretation of factor analysis. *Psychol*. Bull. 70, pp. 361-364.

Cunnane, C. (1989) Statistical distributions for flood frequency analysis. *Hydrol. Rep.* No. 33, WMO Publ. No. 718, Geneva.

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

Dunn, J.C. (1974) Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, Vol 4, pp. 95-104.

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

Hosking, J.R. and Wallis, J.R. (1997) *Regional Frequency Analysis: An Approach based on L-Moments.* Cambridge University Press.

Kaiser, H.F., Rice, J. (1974) Little Jiffy Mark. *Educ. Psychol. Measur.* 34, pp. 111-117.

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

McQueen, J.B. (1967) Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, Vol. 1, pp. 281-297.

Overall, J.E. and Klett, C.J. (1972) *Applied multivariate analysis*. McGraw-Hill, New York.

Puvaneswaran, M. (1990) Climatic classfication for Queensland using multivariate statistical techniques. *Int. J. Climatol*. Vol. 10, pp. 591-608.

Rousseeuw, P.J. (1987) Silhouettes: Graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53-65.

White, D., Richman, M., and Yanel, B. (1991) Climate regionalization and rotation of principal components. *Int. J. Climatorl.* Vol. 11, pp. 1-25.

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

◎ 논문접수일 : 2012년 10월 08일
◎ 심사의뢰일 : 2012년 10월 11일
◎ 심사완료일 : 2012년 10월 16일