다변량 분석 기법을 활용한 동질 지역 구분

Identification of Homogeneous Regions based on Multivariate Techniques

남우성*, 김태순**, 허준행***

Woo Sung Nam, Tae Soon Kim, Jun-Haeng Heo

요 지

지역빈도해석은 우리나라와 같이 자료 기간이 짧은 경우 지점빈도해석보다 더 정확한 확률강우량을 산정할 수 있는 기법이다. 지역빈도해석을 통한 확률강우량 산정 결과는 수문학적으로 동질한 지역의 구분 결과에 따라 달라진다. 지역을 구분할 때에는 강우에 영향을 미치는 다양한 변수들이 사용될 수 있다. 변수의 유형과 개수가 지역 구분의 효율성을 좌우하기 때문에 활용 가능한 모든 변수들의 정보를 요약할 수 있는 변수들을 선택하는 것이 지역 구분의 효율성 면에서 유리하다고 할 수 있다. 이런 면에서 지역 구분의 효율성을 증대시킬 목적으로 다변량 분석 기법이 활용될 수 있다. 본 연구에서는 주성분 분석, 요인 분석, Procrustes analysis와 같은 다변량 분석 기법을 활용하여 42개의 강우 관련 변수들을 33개의 변수로 줄일 수 있었다. 분석 결과 변수 개수 감소로 인한 정보 손실은 크지 않은 것으로 나타났다. 따라서 이러한 기법에 의한 변수 차원의 축소는 지역 구분의 효율성 향상에 기여할 수 있는 것으로 판단된다. 선정된 변수들을 바탕으로 군집해석을 수행하여 지역을 구분하였고, L-모멘트에 근거한 이질성 척도(H)를 활용하여 구분된 지역의 동질성을 검토하였다. 또한 L-모멘트에 근거한 적합성 척도(Z)를 적용하여 구분된 지역에 적합한 확률분포형을 선정하였고, 선정된 적정 확률분포형을 바탕으로 각 지역에 대한 성장 곡선(growth curve)을 유도하였다.

핵심용어: 주성분 분석, 요인 분석, Procrustes analysis, 지역빈도해석

1. 서 론

우리나라와 같이 강우자료의 기록년수가 100년 미만인 경우에 지점빈도해석을 통해 산정된 재현기간 100년 이상의 확률강우량은 신뢰도가 떨어지는 문제점이 있다. 이러한 지점빈도해석의 단점을 보완하기 위해 지역빈도해석이 확률강우량 산정에 활용되고 있다. 지역빈도해석은 수문학적으로 동질하다고 판정된 지점들을 하나의 지역으로 구분하여 그 지점의 강우자료를 바탕으로 확률강우량을 추정하는 방법이다. 이런 의미에서 지역빈도해석은 충분하지 못한 자료기간을 공간적으로 확장시켜 보완하는 방법이라 할 수 있다. 지역빈도해석은 Hosking and Wallis (1997)에 의해서 L-moments를 이용한 기법이 개발된 이후로 많은 연구가 수행되어 왔으며, 최근 국내 연구에도 적용되고 있다.

지역빈도해석에서 가장 중요한 단계는 동질 지역의 구분이라 할 수 있다. 동질 지역 구분을 위한 다양한 기법과 변수들이 활용되어 왔다. Guttman(1993)은 7개의 지형 및 기후 관련 인자들을

* 정회원 · 연세대학교 대학원 토목공학과 박사과정 · E-mail: nws77@yonsei.ac.kr

** 정회원 · 세종대학교 대학원 토목공학과 박사후과정 · E-mail: chaucer@yonsei.ac.kr

*** 정회원 · 연세대학교 사회환경시스템공학부 토목환경공학전공 교수 · E-mail: jhheo@yonsei.ac.kr

사용하여 지역을 구분했고, Mallants와 Feyen(1993)은 단지 일강우 자료만을 가지고 지역을 구분했다. 최근에는 다변량 분석 기법이 지역 구분에 많이 활용되고 있다. Zhang과 Hall(2004)는 몇 가지 군집해석 기법을 활용해서 지역을 구분했고, Dinpashoh 등(2004)은 주성분 분석, Procrustes analysis, 인자분석을 활용하여 지역을 구분함으로 동질 지역 구분의 효율을 향상에 대해 연구하였다.

본 연구에서는 동질 지역 구분에 활용되는 구분 인자들을 좀 더 효율적으로 선택하기 위해 주성분분석(principal component analysis)과 요인분석(factor analysis)을 적용하고, Procrustes analysis를 통해서 원래 변수의 성질을 충분히 설명할 수 있는 최소 개수 변수를 결정하는데 그 목적이 있다.

2. 강우자료 구축

본 연구에서 활용한 자료는 기상청 관할의 전국 60개 지점의 일강우 자료로서, 자료 기간이 15년 미만인 4개 지점은 제외했으며, 1개월이라도 결측치가 있는 해당년도는 분석 대상에서 제외하였다. 이렇게 선정된 각 지점들의 수문학적 특성을 반영하는 지역구분인자로서 여러 가지를 고려할 수 있지만 본 연구에서는 표 1과 같은 인자들을 선정하였다.

표 1. 지역구분을 위한 수문학적 인자

인자	설명
MAP	연평균강우량(Mean Annual Precipitation)
DayP	연간 강우일수(number of Days with Precipitation in a year)
APM_i, i=1, 2, …, 12	월평균강우량(Average Precipitation in a Month)
DP_i, i=1, 2, …, 12	월간 강우일수(number of Days with Precipitation in a month)
MDP_i, i=1, 2, …, 12	매월 강수량 중 최대값을 전체기간에 대해서 평균한 값(average Maximum Daily Precipitation in a month)
AMaxMDP	월간 최대 일강우량의 연최대값을 전체 기간에 대해서 평균한 값 (Average Maximum of Maximum Daily Precipitation in a month)

표 1에서 볼 수 있는 것처럼 수문학적 인자 39개에 지형학적 인자인 위도, 경도, 고도를 포함시켜 42개의 인자를 지역구분을 위한 인자로 선정하였다.

3. 주성분분석과 Procrustes Analysis

주성분분석과 요인분석은 변수가 많은 경우에 군집해석의 효율성 저하 문제를 해결하기 위해 주로 사용되는 방법이다. 두 방법 모두 기본이론은 비슷하지만 주성분분석은 상관성이 있는 변수들이 있을 경우 이 변수들이 보유한 정보의 손실을 최소화 하면서 주성분(principal component)이라 불리는 새로운 변수를 기존 변수의 선형조합으로 만들어내서 활용하는 것이고, 요인분석은 기존 변수들의 상관성을 이용하여 요인(factor)라 불리는 변수 내의 공통적인 새로운 변수를 도출하

여 이 요인들이 가지고 있는 특성으로 전체 자료의 특성을 최대한 설명하는 기법이다.

Procrustes analysis (Krzanowski, 1987)는 다변량 해석을 위해서 변수를 선택하는 과정을 모의하는 방법의 하나로서 원래 설정된 p개의 변수를 이용해서 구한 주성분분석 점수(principal component score) 중에서 k차원에 해당하는 점수와 설정된 변수 중에서 적절한 개수를 제외한 q개의 변수를 이용해서 구한 주성분분석 점수 중 k차원에 해당하는 점수를 비교해서, 두 개 차원의 점수의 차이를 최소화시키는 q개의 변수를 찾아내는 기법이다.

그림 1은 Procrustes analysis의 절차를 나타낸 것으로서 모든 변수를 가지고 있는 원래의 행렬을 $X_{(n\times p)}$라고 할 때, 원래 변수의 성질을 최대한 반영하는 최소한의 변수를 가지는 행렬을 $X_{(n\times q)}$라 하고, 원래 변수로부터 구한 주성분분석 점수 행렬 중 k차원의 행렬 $Z_{(n\times k)}$와 $Y_{(n\times k)}$을 비교해서 그 차이가 최소가 되도록 만드는 것을 의미한다.

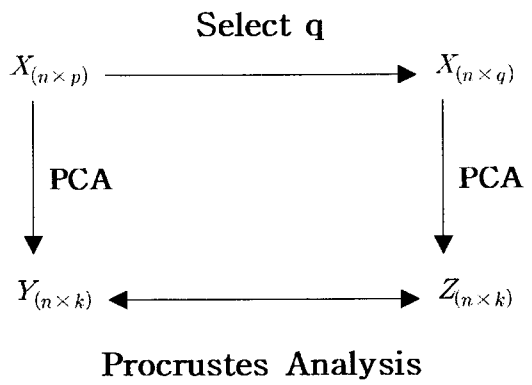

그림 1. Procrustes Analysis를 활용한 변수 선정

$Z_{(n\times k)}$와 $Y_{(n\times k)}$을 비교하기 위해서는 아래의 M^2으로 정의된 제곱오차합(sum of squared differences)이 최소가 되도록 하는 차원수를 구하면 된다.

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

여기서, '은 전치(transpose)를 의미하고, Σ는 행렬 $Z'Y$의 SVD (singular value decomposition)를 통해서 구한 대각행렬(diagonal matrix)을 가리킨다. $Z'Y$는 아래와 같이 정의된다.

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

여기서, $UU' = I_k$이고, $V'V = VV' = I_k$이다.

4. 변수 선정

앞서 언급한 Procrustes analysis를 통해서 총 42개의 지점별 특성치 중에서 9개의 변수를 제거한 33개의 변수를 이용한 결과가 원래의 주성분 점수를 가장 잘 나타내는 것으로 나타났다. 제

거된 변수는 DP4, 고도, 경도, DP3, DP9, MDP9, APM9, MDP5, DP8의 9개이다. 변수의 개수를 조정하고 나면 고유치(eigenvalue)도 변하는데 42개 변수를 이용했을 때는 7개의 고유치가 1을 넘는 값을 나타냈지만, Procrustes analysis를 거친 후에는 6개의 변수가 1을 넘는 것으로 나타났다. 가장 큰 고유치를 갖는 주성분의 분산 설명력 역시 32.7%에서 35.6%로 증가했으며, 7개의 주성분 모두에 걸친 설명력은 89.9%였는데 반해 6개의 주성분을 이용한 결과는 91.34%로 증가하는 경향을 보였다. 표 2는 42개의 변수를 이용한 경우와 33개의 변수를 이용한 경우에 대한 고유치를 나타낸 것이다.

주성분분석과 Procrustes analysis를 이용해서 변수를 선택한 후에는 요인분석을 적용해서 각 변수를 대표해서 설명할 수 있는 요인을 선택하는 과정을 거치게 된다. 요인분석을 위해서는 우선 사용할 요인 개수를 결정해야 하는데 이를 위해서 Scree Plot을 도시하여 각 요인별로 고유치가 어떻게 변화하는지 검토한 후 결정하게 된다. 본 연구에서는 Scree Plot을 그려본 결과 요인이 6개인 지점부터 고유치가 급격히 감소하는 양상을 보였기 때문에 요인의 개수를 5개로 결정하였다.

요인의 개수를 확인한 후 각 요인별로 어떤 변수들을 대표할 수 있는지 결정하기 위해서 인자패턴(factor pattern)이 높은 것들을 기준으로 각 변수와 요인별 상관관계를 살펴보게 된다. 표 3은 이런 과정을 거쳐서 구한 요인과 변수의 관계를 나타낸 것으로, Factor 1은 주로 늦가을부터 초봄까지의 변수들에 영향을 받는 요인이라 할 수 있고, Factor 2는 주로 봄과 초여름에 관련된 변수라고 할 수 있다. Factor 3은 대부분의 변수들이 여름의 집중호우기에 관련된 변수들과 관련이 있고, Factor 4는 주로 강우일수와 관련된 요인이라 할 수 있으며, Factor 5는 5월과 6월의 강우일수와 관계가 있는 요인이라 할 수 있다.

표 2. 변수 개수 조정에 따른 고유치

	변수 42개				변수 33개			
	고유치	차이	비율	누가비율	고유치	차이	비율	누가비율
1	13.74	6.17	0.327	0.327	11.75	4.83	0.356	0.356
2	7.58	0.96	0.180	0.507	6.92	1.25	0.210	0.566
3	6.62	2.23	0.158	0.665	5.67	2.51	0.172	0.738
4	4.38	1.38	0.104	0.769	3.16	1.55	0.096	0.834
5	3.00	1.63	0.07	0.839	1.61	0.58	0.049	0.883
6	1.36	0.23	0.03	0.869	1.03	0.32	0.031	0.914
7	1.13	0.38	0.03	0.899	0.71	0.36	0.022	0.936

표 3. 요인별 변수분포

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

5. 결 론

본 연구는 지역빈도해석의 지역구분을 위해 군집해석을 적용할 경우 사용되는 인자를 선택하기 위한 방법의 하나로 Procrustes analysis를 활용하였다. 주성분분석과 요인분석을 통해서 기존의 42개 변수를 33개 변수로 줄이면서도 주성분의 설명력은 높이고 주성분의 개수는 줄이는 결과를 얻을 수 있었다. 변수 개수의 감소에도 불구하고 전체 자료의 정보는 보존되는 것으로 나타났다. 이와 같이 주성분분석과 Procrustes analysis, 요인분석을 이용하면 더욱 효율적인 군집해석을 수행할 수 있는 변수 선택이 가능할 것이다. 본 연구 결과를 이용하여 군집해석 결과의 신뢰성을 향상시킬 수 있을 것이라고 판단된다.

감 사 의 글

이 논문은 2005년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2005-041-D00808).

참 고 문 헌

1. Dinpashoh, Y., Fakheri-Fard, A., Moghaddam, M., Jahanbakhsh, S., Mirnia, M. (2004). "Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods" Journal of Hydrology, 297, 109-123.
2. Guttman, N.B. (1993). "The use of L-Moments in the determination of regional precipitation climates" Journal of Climatology, 6, 2309-2325.
3. Hosking, J.R.M. and Wallis, J.R. (1997). Regional frequency analysis, Cambridge University Press.
4. Krzanowski, W.J. (1987). Selection of variables to preserve multivariate data structure using principal components. Applied Statistics, 36(1), pp. 22-33.
5. Mallants, D., Feyen, J. (1990). "Defining homogeneous precipitation regions by means of principal component analysis" Journal of Applied Meteorology, 29, 892-901.
6. Zhang Jingyi, M.J. Hall (2004). "Regional flood frequency analysis for the Gan-Ming River basin in China" Journal of Hydrology, 296, 98-117.