다변량 분석을 활용한 강우지역빈도해석의 지역구분인자 선정에 관한 연구

Selection of variables for regional precipitation frequency analysis using multivariate analysis

남우성*, 김태순**, 허준행***
Woosung Nam, Taesoon Kim, Jun-Haeng Heo

요 지

지역빈도해석기법은 수문학적으로 성질이 유사한 지점을 하나의 군으로 구성한 자료를 이용해서 빈도해석을 하는 기법으로, 지점빈도해석이 가질 수 있는 단점들을 보완하기 위한 방안의 하나로 기대되고 있다. 본 논문은, 지역빈도해석기법을 적용하기 위한 단계중의 하나인 군집해석에 사용되는 변수들을 보다 효율적으로 선택하기 위한 연구로서, 다변량 분석방법인 주성분분석과 요인분석, 그리고, 변수선택을 위한 Procrustes Analysis를 통해서 보다 효율적으로 변수를 선택하는 방법을 제안하기 위한 연구이다.

핵심용어 : 다변량 분석, 주성분 분석, 요인 분석, Procrustes Analysis

1. 서론

확률강우량 산정을 위해서 기존에 사용되어 왔던 지점빈도해석은, 우리나라와 같이 강우자료의 기록년수가 빈도해석을 실시하는데 충분하지 못할 경우에는 그 신뢰도가 떨어지는 단점을 가지고 있다. 이와 같이 기록년수의 부족으로 인해서 발생할 수 있는 지점빈도해석의 단점을 해결하기 위해서, 최근에는 해당유역내에 존재하는 강우관측소별로 수문학적인 동질성을 고려한 '군(cluseter)'을 구성한 후에, 해당군의 확률강우량을 구해내는 지역빈도해석기법이 많이 사용되고 있다. 지역빈도해석은 Hosking and Wallis (1997)에 의해서 L-moments를 이용한 기법이 개발된 이후에 많은 연구가 수행되어 왔으며, 최근에는 국내의 여러 연구에도 적용되고 있다.

본 연구는 지역빈도해석을 실시하는데 있어서 필요한 기본적인 단계중의 하나인, 강우자료의 군집해석에 사용되는 변수를 좀더 효율적으로 선택하기 위한 연구로서, 주성분분석과 요인분석, 그리고 Procrustes Analysis를 통해서, 전체 변수를 이용하지 않고서도 충분히 원래 변수의 성질을 설명할 수 있는 최소갯수의 변수를 결정하기 위한 연구이다.

* 정회원・연세대학교 대학원 토목공학과 박사과정 수료・E-mail: nws77@yonsei.ac.kr
** 정회원・연세대학교 대학원 토목공학과 박사후과정 ・E-mail: chaucer@yonsei.ac.kr
*** 정회원・연세대학교 사회환경시스템공학부 토목·환경공학전공 교수・E-mail: jhheo@yonsei.ac.kr

2. 강우자료구축

본 연구에서 사용한 자료는 기상청에서 관리하는 전국의 71개 지점에 관한 일강우자료로서, 자료기간이 15년 미만인 4개 지점은 제외했으며, 1개월이라도 결측치가 있는 해당년도는 분석대상에서 제외했다. 이 자료를 이용해서 주성분분석(principal component analysis)을 실시한 결과, 울릉도, 성산포, 서귀포, 대관령, 제주, 거제, 남해관측소가 모두 다른 자료들과는 상관성이 없는 이상치(outlier)로 판별되었으므로, 이 자료를 제외한 총 60개 지점의 자료를 분석대상으로 해서 연구를 수행했다. 아래 표 1은 이와 같이 구축된 지점별 강우자료중에서 후보군으로 생각할 수 있는 특성치들을 나타낸 것이다.

표 1. 각 지점별 특성치

	Descriptions
MAP	Mean annual precipitation
DayP	Number of days with precipitation in a year
APM	Average precipitation in a month
DP	Number of days with precipitation in a month
MDP	Average maximum daily precipitation in a month
AMaxMDP	Average Maximum of maximum daily precipitation in a month

위의 특성치중 MAP는 년평균강우량(1개), DayP는 년간 강우일수(1개), APM은 월평균강우량(12개), DP는 월간 강우일수(12개), MDP는 매월 강수량중 최대값을 전체 자료기간에 대해서 평균한 값(12개), 그리고, AMaxMDP는 월간 최대 일강우량의 년 최대값을 전체기간에 대해서 평균한 값(1개)이다. 이 특성치들의 각 지점별 개수는 총 39개이고, 여기에 각 지점별 위도, 경도, 고도를 합한 42개가 지점별 특성치가 된다.

3. 주성분분석과 Procrustes Analysis

주성분분석과 요인분석은 대상 자료의 특성치가 많을 경우에 발생할 수 있는 군집해석의 비효율성을 해결하기 위해서 주로 사용하는 방법으로, 두 방법 모두 비슷한 기본이론을 가지고 있지만, 주성분 분석은 서로 연관이 있는 변수들이 관측되었을 때, 이 변수들이 가지고 있는 정보들을 최대한 포함하는 새로운 변수(principal component)를 관측된 변수들의 선형조합을 이용해서 만들어내는 것이고, 요인분석은 관측된 변수들의 상호관련성을 이용하여 변수속에 내재된 요인(factor)이라는 소수의 공통적인 새로운 변수를 찾아내서 이 요인들이 가지고 있는 특성으로 전체 자료가 가지는 특성을 최대한 설명하고자 하는 기법이다. 두 기법 모두 군집해석을 위한 성분분석에 사용되지만, 주성분분석은 주로 이상치를 검토하는 단계에 사용되고, 요인분석이 실제로 관측된 자료들 중에서 공통되는 요인들을 추출해내는데 사용된다.

앞서 언급한 것과 같이, 주성분분석을 이용해서 이상치로 판별된 7개 지점을 제외시킨 총 60개 지점에 대해서 Procrutes Analysis를 실시했다. Procrutes 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 k-차원에 해당하는 점수와 설정된 변수

중에서 적절한 개수를 제외한 q개의 변수를 이용해서 구한 주성분분석 점수중 k-차원에 해당하는 점수를 비교해서, 두 개 차원의 점수의 차이를 최소화시키는 q개의 변수를 찾아내는 기법이다.

아래 그림 1은 Procrutes Analysis의 절차를 나타낸 것으로, 모든 변수를 가지고 있는 원래의 행렬을 $X_{(n\times p)}$라고 할 때, 원래 변수의 성질을 가장 잘 나타내면서도 최소한의 변수를 가지고 있는 행렬을 $X_{(n\times q)}$라 하고, 원래 변수로부터 구한 주성분분석 점수 행렬중 k-차원의 행렬을 $Y_{(n\times k)}$라고 하면, $X_{(n\times q)}$를 이용해서 구한 주성분분석 점수 행렬중 k-차원의 행렬을 $Z_{(n\times k)}$와 $Y_{(n\times k)}$를 비교해서 가장 최소한의 차이가 나도록 만드는 것이 Procrustes Analysis이다.

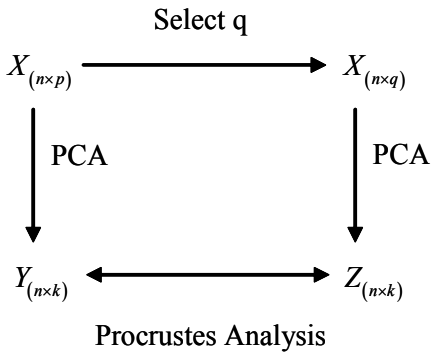

그림 1. Procrustes Analysis를 이용한 변수 선택과정

$Z_{(n\times k)}$와 $Y_{(n\times k)}$의 차이를 비교하기 위해서는, 아래의 M^2로 정의된 제곱오차합(sum of squared differences)을 최소화시키는 차원수를 구하면된다.

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

여기서, '은 trasnpose를 의미하고, Σ는 $Z'Y$행렬의 singular value decompostion을 통해서 얻어진 대각행렬(diagonal matrix)을 의미한다. $Z'Y$는 아래 식으로 정의된다.

$$Z'Y = U\Sigma V' \quad (2)$$

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

4. 군집해석을 위한 변수선택

위에 언급한 Procrustes Analysis를 통해서, 총 42개의 지점별 특성치중에서 9개의 변수를 제거한 33개의 변수를 이용한 결과값이 원래의 주성분점수를 가장 잘 나타내는 변수의 집합으로 나타났다. 제거된 변수는 dp04, alti, long, dp03, dp09, mdp09, apm09, mdp05, dp08의 9개이다. 변수의 개수를 조정하고 나면, 고유치값에도 변화가 생기는데, 원래의 42개 변수를 이용했을때는 7개의 고유치가 1.0을 넘는 값이 나왔지만, Procrustes Analysis를 거친후에는 6개의 변수가 고유치 1.0을 넘는 것으로 나타났다. 가장 큰 고유치값을 갖는 주성분의 variance 설명력 역시 32.7%에서 35.6%로 증가했으며, 7개의 주성분 모두에 걸친 설명력은 89.9%였는데 반해서, 6개의 주성분을 이용한 결과는 91.34%로 증가하는 경향을 보였다. 아래의 표 1은 42개의 변수를 이용한 경우와 33개의 변수를 이용한 경우에 대한 고유치를 나타낸 것이다.

주성분분석과 Procrustes Analysis를 이용해서 변수를 선택한 후에는, 요인분석을 이용해서 각각의 변수를 대표해서 설명할 수 있는 요인을 선택하는 과정을 거치게 된다. 요인분석을 위해서는 우선 사용할 요인을 몇 개나 사용할 것인지를 결정해야하는데, 이를 위해서 Scree Plot을 그려서 각

요인별로 고유치가 어떻게 변화하는지를 살펴본 후에 요인의 개수를 결정하게 된다. 본 연구에서는 Scree Plot을 그려본 결과 요인이 6개인 지점부터 고유치가 급격히 감소하는 형태로 나타나는 것을 확인했기에, 요인의 개수는 5개로 결정했다.

요인의 개수를 확인한 후에는, 각 요인별로 어떤 변수들을 대표할 수 있는지를 결정하기 위해서 Factor Pattern이 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 변수값들에 영향을 받는 요인이라고 말할 수 있고, Factor 2는 주로 봄과 초여름에 관련된 변수라고 말할 수 있다. Factor 3은 대부분의 변수들이 여름의 집중호우기에 관련된 변수들에 관련된 요인값을 가지고 있고, Factor 4는 주로 강우일수와 관련된 요인이라고 할 수 있고, Factor 5는 5월과 6월의 강우일수와 관계가 있는 요인이라고 할 수 있다.

표 2. 변수 42개와 33개인 경우의 고유치

	변수 42개				변수 33개			
	Eigenvalue	Difference	Proportion	Cumulative	Eigenvalue	Difference	Proportion	Cumulative
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. 5개 요인별 변수분포표

	Descriptions
Factor 1	APM01, 02, 10, 11, 12 / MDP01, 02, 03, 10, 11, 12
Factor 2	APM03, 04, 05, 06 / MDP04, 06 / Latitude
Factor 3	APM07, 08/ MDP07, 08 / MAP / DP07 / AMaxMDP
Factor 4	DayP / DP10, 11, 12/ DP01 ,02
Factor 5	DP05, 06

5. 결론

본 연구는 지역빈도해석의 군집해석을 위해서 사용되는 변수를 선택하기 위한 방법의 하나로 Procrustes Analysis를 사용한 연구로서, 주성분분석과 요인분석을 통해서 기존의 42개의 변수를 33개의 변수로 줄이면서도, 주성분의 설명력은 높이고 주성분의 개수는 줄이는 결과를 보여주었다. 이와 같이 주성분분석과 Procrustes Analysis, 그리고 요인분석을 이용하면 보다 효율적으로 군집해석을 수행할 수 있는 변수선택이 가능해지게 된다. 본 연구의 결과를 이용하면, 지역빈도해석의 기본적인 수행절차인 군집해석의 결과의 신뢰성을 개선시킬 수 있을 것이라고 판단된다.

참고문헌

1. Hosking, J.R.M. and Wallis, J.R. (1997). Regional frequency analysis, Cambridge University Press.
2. Krzanowski, W.J. (1987). "Selection of variables to preserve multivariate data structure, using principal components." *Applied Statistics*, 36(1), 22-33.