Journal of the Korean Data & Information Science Society
2017, **28**(5), 1133–1144

http://dx.doi.org/10.7465/jkdi.2017.28.5.1133
한국데이터정보과학회지

기온과 강수량의 수치모델 격자자료를 이용한 기상관측지점의 월별 군집화[†]

김희경[1] · 김광섭[2] · 이재원[3] · 이영섭[4]

[124]동국대학교 통계학과 · [3]기상청 국가기후데이터센터

접수 2017년 4월 17일, 수정 2017년 7월 28일, 게재확정 2017년 8월 14일

요 약

기상자료를 이용한 군집분석은 기상 특성에 근거한 기상 지역의 세분화를 가능하게 하고 군집을 이루는 지형별 기상 특성의 파악을 용이하게 한다. 이때 기상관측자료를 이용한 군집분석은 관측지점의 밀도가 다르기 때문에 우리나라의 기상 특성이 고르게 반영되지 못할 수 있다. 반면 수치모델 격자자료는 5km×5km 간격으로 조밀하고 고른 자료의 생산이 가능하므로 우리나라의 기상 특성을 고르게 반영할 수 있다. 본 연구에서는 기온과 강수량의 수치모델 격자자료를 이용하여 군집분석을 수행하고, 그 결과를 바탕으로 기상관측지점에 대한 군집을 결정하였다. 기상 특성이 월별로 상이할 수 있기 때문에 군집분석은 월별로 수행하였으며, K-Means 군집분석 방법의 단점을 보완하고자 계층적 군집분석 방법인 Ward 방법과 결합하여 적용하였다. 그 결과 우리나라 기상관측지점들에 대해 시 · 공간적으로 세분화된 군집화가 이루어졌다.

주요용어: 강수량, 군집분석, 기상관측지점, 기온, 수치모델 격자자료, K-Means 방법, Ward 방법.

1. 서론

최근 변화되는 기상 특성에 근거하여 기상 지역의 세분화를 위한 지역적 연구가 활발히 진행되고 있다. 특히 기상 요소별 변동 및 특성을 이용하여 다변량 자료의 분류기법 중 하나인 군집분석 (cluster analysis) 기법을 많이 활용하고 있다. 군집분석은 비유사성 (거리)에 근거하여 가까운 순서대로 군집화하는 탐색적인 통계 분석방법 (EDA)이며 대용량 자료를 분석할 때 전체를 유사한 개체들로 군집화 하여 전체 자료를 잘 대표하는 군집들로 나뉘어 관찰함으로서 전체 자료에 대해 효율적인 정보를 얻어 낼 수 있어 다양한 분야에서 적용되고 있다 (Anderberg, 1973; Kim, 2015; Yoon and Choi, 2015). 기상자료를 이용한 군집분석은 우리나라 기상의 특성을 군집별로 파악함으로써 군집을 이루는 지형별로 기상 특성의 파악을 용이하게 한다. 우리나라에는 현재 기상관측지점 (meteorological stations) 중 Figure 1.1과 같이 100여개의 ASOS (automated synoptic observing system)와 500여개의 AWS (automatic weather system)를 합쳐 약 600여개의 기상관측지점 (surface observing stations)이 존재하는데 이러

[†] 이 연구는 기상청 「기상 · 지진See-At기술개발연구」(KMIPA 2015-1020)의 지원으로 수행되었으며, 또한 제 1저자인 김희경의 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2016R1A6A3A11931946).

[1] (04620) 서울시 중구 필동로1길 30, 동국대학교 통계학과, 박사후 연구원.
[2] (04620) 서울시 중구 필동로1길 30, 동국대학교 통계학과, 석사과정.
[3] (07062) 서울시 동작구 여의대방로16길 61, 기상청 국가기후데이터센터, 센터장.
[4] 교신저자: (04620) 서울시 중구 필동로1길 30, 동국대학교 통계학과, 교수. E-mail: yung@dongguk.edu

한 기상관측지점의 관측 지역의 기상 특성이 고르게 반영되지 못할 수 있다. 반면 수치모델 격자자료를 이용하게 되면 5km × 5km 간격으로 조밀하여 위치적으로 고른 자료의 생산이 가능 하므로 남한지역의 기상 특성을 고르게 반영할 수 있다. 기존에 Lee 등 (1999), Ju 등 (2008), Yeo (2011)의 연구와 같이 기상관측 자료를 이용한 군집분석은 계속 연구되어 왔지만 수치모델 격자자료를 이용한 군집분석은 많이 존재하지 않았다. 본 연구에서는 기온과 강수량에 대한 장기간의 수치모델 격자자료를 이용하여 군집분석을 수행하였으며, 그 결과를 바탕으로 기상관측지점에 대한 군집을 할당하였다.

Figure 1.1 Distribution of approximately 600 climatological stations in Korea

2. 연구 방법

2.1. 분석 데이터

우리나라 기상청의 수치모델 격자자료가 생산되는 범위는 Figure 2.1에 나타나있는 것과 같이 왼쪽 상단 위 · 경도 좌표로 (43.3935°, 123.3102°)에서 왼쪽 하단의 위 · 경도 좌표 (31.7944°, 123.7613°)까지이며, 오른쪽 상단 위 · 경도 좌표 (43.2175°, 132.7750°)에서 오른쪽 하단 (31.6518°, 131.6423°)까지이다. 이 범위 안에는 남북한의 육지 및 해상, 중국, 러시아, 일본까지 포함되어 있으며, 가로 149 격자 (745km) × 세로 283 격자 (1,265km)의 좌표 구조로 이루어져 있다. 각 격자점에 대해 1시간 간격의 기상자료가 생산되고 있다. 본 연구에서는 기온과 강수량의 수치모델 격자자료에 대한 군집분석을 위해 2006년 1월 1일부터 2014년 4월 30일까지 KLAPS (Korea local analysis and prediction system) 모델 분석장 자료를 이용 하였다. KLAPS는 초단기 기상분석 및 예측시스템으로서 한반도 및 주위의 영역에대하여 5km × 5km 간격 해상도의 기상 자료를 재분석 및 예측하는 시스템이다 (Kim 등, 2013). 2014년 5월 1일 부터의 기상청 수치모델 격자자료는 모델이 변경되어 기상 자료가 생산되었기 때문에 모델의 효과를 고정시키기 위해 2014년 5월 1일 부터의 수치모델 격자자료는 분석대상에서 제외하고, 2006년 1월 1일부터 2014년 4월 30일까지의 자료만 분석하였다.

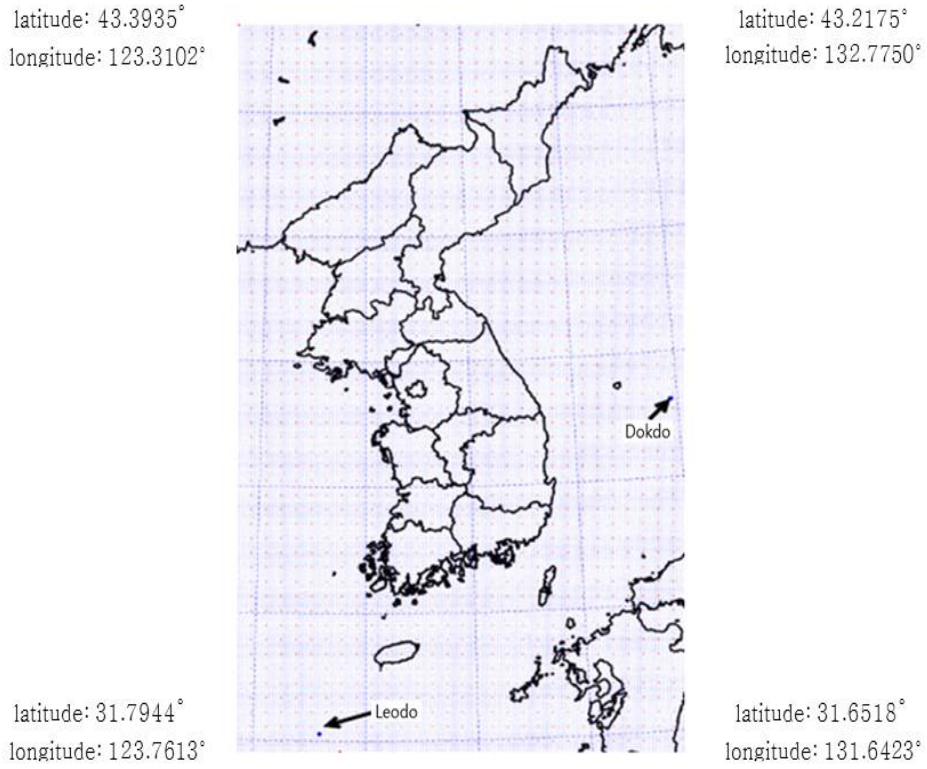

Figure 2.1 Area with 4 edge points including gridded data of numerical model in Korea

본 연구에서는 마스킹 파일의 각 격자점에 대한 위치 정보를 활용하여 남한의 육지와 해안에 해당하는 격자점만을 분석대상으로 하였다. 마스킹 파일에 나타나있는 저장값에 대한 위치정보는 Table 2.1과 같다. 이 정보를 이용하여 저장값이 11부터 63까지인 전체 20,044 격자점에 대해서만 분석을 하였다. 또한 위치 정보를 이용하여 분석 자료를 육지 (3,589개 격자점)와 해안 (16,455개 격자점)으로 구분하여 각각 분석을 실시하였다. 이는 육지와 해안의 구분 없이 전체 자료를 이용하여 군집분석을 실시하면 해안의 특성에 육지의 특성이 가려져 육지에 대한 군집화가 잘 이루어지지 않을 수 있기 때문이다.

KLAPS 모델 분석장 자료를 이용하여 군집분석을 실시하기 위해 먼저 데이터의 차원을 축소시켰다. KLAPS 모델 분석장 자료는 시간 단위의 기온과 강수량 자료이므로, 데이터 차원 축소를 위해 기온에 대해서는 일평균기온을 계산하고, 강수량에 대해서는 일누적강수량을 계산하였다. 계산된 일단위의 기온과 강수량 자료에 대해 각각 군집분석을 실시하였다. 또한 기상의 특성은 월별로 상이할 수 있으므로 해당 월에 따라 군집의 개수나 군집을 이루는 지점들이 달라질 수 있다. 따라서 주어진 수치모델 격자자료에 대해 월별로 군집분석을 실시하였다. 강수량 자료의 경우 변동성이 크다는 특성이 있으므로, 월별로 자료의 최소값과 최대값을 찾아 식 (2.1)과 같이 일누적강수량값이 0과 1사이의 값을 가지도록 변환과정을 거쳤다.

$$Z_i = \frac{\max - X_i}{\max - \min}, \quad i = 1, \cdots, n, \tag{2.1}$$

Table 2.1 Position information for gridded points

Value	Contents	Note		
		Land	Ocean	Exception
0	Exception			√
11	Land : Seoul, Incheon	√		
12	Ocean : Incheon and Gyeonggi and on shore		√	
13	Ocean : Incheon and Gyeonggi and off shore		√	
21	Land : Chungcheong	√		
22	Ocean : Chungcheong on shore		√	
23	Ocean : Chungcheong off shore		√	
31	Land : Gangwon	√		
32	Ocean : Gangwon on shore		√	
33	Ocean : Gangwon off shore		√	
41	Land : Gwangju and Jeolla	√		
42	Ocean : Jeolla on shore		√	
43	Ocean : Jeolla off shore		√	
51	Land : Jeju	√		
52	Ocean : Jeju on shore		√	
53	Ocean : Jeju off shore		√	
61	Land : Busan, Daegu, Ulsan, Gyeongsang	√		
62	Ocean : Busan, Ulsan, Gyeongsang on shore		√	
63	Ocean : Busan, Ulsan, Gyeongsang off shore		√	
71	Land : North Korea			√
72	Ocean : North Korea on shore			√
73	Ocean : North Korea off shore			√
81	Land : Japan			√
91	Land : China , Russia			√

여기서 Z_i는 일누적강수량을 변환한 값이며, X_i는 일누적강수량을 나타낸다. max는 해당 월자료에서의 최대값, min은 최소값을 나타내며, n은 해당월의 관측값의 개수를 의미한다. 일평균기온, 변환된 일누적강수량을 이용하여 각각 월별로 군집분석을 수행하였다.

2.2. 군집분석

군집분석 방법에는 계층적 군집분석 (hierarchical cluster analysis)과 비계층적 군집분석 방법 (non-hierarchical cluster analysis)이 있다. 계층적 군집분석은 처음에 n개의 군집으로부터 시작하여 점차 군집의 개수를 줄여나가는 방법으로 자료에 적합한 군집의 개수를 결정하기가 용이하다. 계층적 군집분석 중 Ward 방법은 군집내 제곱합 증분과 군집간 제곱합을 고려한 방법으로 군집간 정보의 손실을 최소화하도록 군집화를 하는 방법이라 할 수 있다. 여기서 군집간의 정보란 편차제곱합 (error sum of squares; ESS)을 나타낸다 (Ward, 1963; Murtagh와 Legendre, 2014). 비계층적 군집분석은 n개의 개체를 k개의 군집으로 나눌 수 있는 모든 가능한 방법을 고려하여 최적의 군집을 형성하는 방법이다. 대표적으로 K-Means 방법이 있는데, 전체 개체를 k개의 군집으로 나누는 계층적 군집 방법과는 달리 한 개체가 속해있던 군집에서 다른 군집으로 이동하는 재배치 (reallocation)가 가능하다. 또한 거의 모든 형태의 자료에 적용이 가능하고 다른 변환이 필요하지 않아 적용하기 쉬운 장점이 있다. 하지만 K-Means 방법은 군집의 개수 k와 초기값 (seed values)에 의존하는 방법으로 군집의 개수 및 초기값 선택이 최종 군집 선택에 영향을 미치게 된다. 즉, 사전에 정해져야 할 k개의 군집 수에 대한 정보가 주어져야 하며 군집의 중심인 초기값의 설정이 자료에 따라 군집을 형성하는데 있어 큰 영향력을 주는 경향이 있다. 잘못된 초기값의 설정은 올바르지 않은 군집들을 생성하게 된다 (Wagstaff, 2001). 따라서 본

연구에서는 계층적 방법인 Ward 방법을 적용하여 적절한 군집의 개수 k를 결정하고, 선정된 k개의 군집의 중심점을 계산하여 이를 K-Means 방법을 적용시키기 위한 초기값 (seed values)으로 사용하였다. 이렇게 함으로써 적절한 군집 개수와 초기값을 결정하여 초기값에 따라 민감하게 군집분석 결과가 달라질 수 있다는 K-Means 방법의 단점을 보완 하였다.

기온과 강수량의 수치모델 격자자료에 대한 군집분석을 수행한 결과를 바탕으로 지상관측지점에 대해 군집을 할당하였다. 이때 군집 할당의 대상 지점은 비교적 관측기간이 긴 지점들로 ASOS와 AWS를 합쳐 525개의 지점을 대상으로 하였다. 지상관측 지점의 위 · 경도 좌표를 이용하여 기상관측지점과 가장 가까운 격자점을 찾고, 그 격자점의 군집과 동일한 군집으로 할당하였다. 수치모델 격자자료를 이용하여 군집분석을 수행하고 이를 이용하여 기상관측지점에 대해 군집을 할당하는 전반적인 과정을 Figure 2.2에서와 같이 도식화할 수 있다.

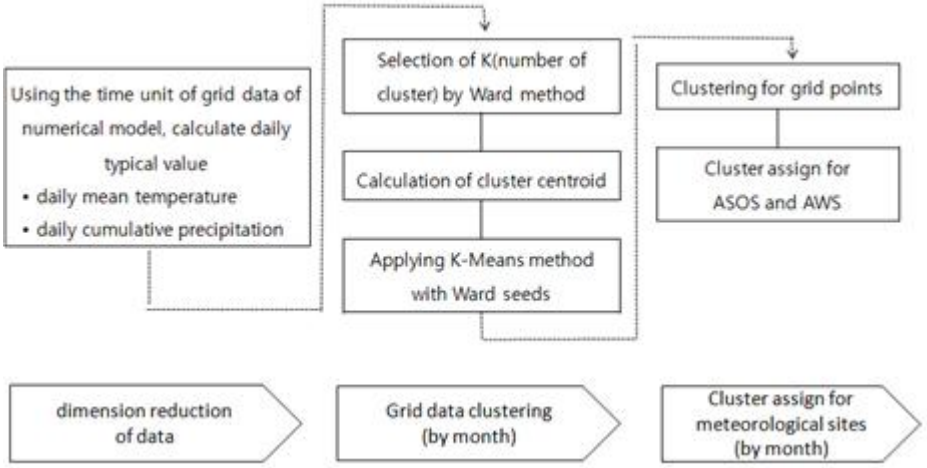

Figure 2.2 Process of cluster analysis for meteorological stations

3. 연구결과

월별로 기온과 강수량의 수치모델 격자자료를 해안과 육지로 구분하고 각각에 대해 군집분석을 실시한 두 결과를 다시 결합하였다. 그 결과 기온에 대한 월별 군집분석 결과 중 봄, 여름, 가을, 겨울의 4계절을 대표할 수 있는 4월, 8월, 10월, 12월의 결과가 Figure 3.1에 나타나있다. 월별 기온의 상이한 특성이 반영되어 각 월마다 군집형성의 결과가 다른 것을 알 수 있다. 기온 자료에 대한 군집분석 결과를 보면 해안가 지역이 내륙과 달리 다른 군집을 형성하는 것을 알 수 있다. 강수량 자료에 대한 월별 군집분석 결과 중 마찬가지로 봄, 여름, 가을, 겨울의 4계절을 대표할 수 있는 4월, 8월, 10월, 12월의 결과가 Figure 3.2에 나타나있다. 월별 강수량의 상이한 특성이 반영되어 각 월마다 군집형성의 결과가 다른 것을 알 수 있다. 강수량에 대한 격자자료의 군집화는 기온자료에 대한 결과와는 달리 격자점의 위치적 영향이 강하게 나타나는 것을 알 수 있다. 즉, 강수량에 대해서는 유사지역의 격자점들이 동일 군집으로 묶이는 경향이 강한 것을 알 수 있다.

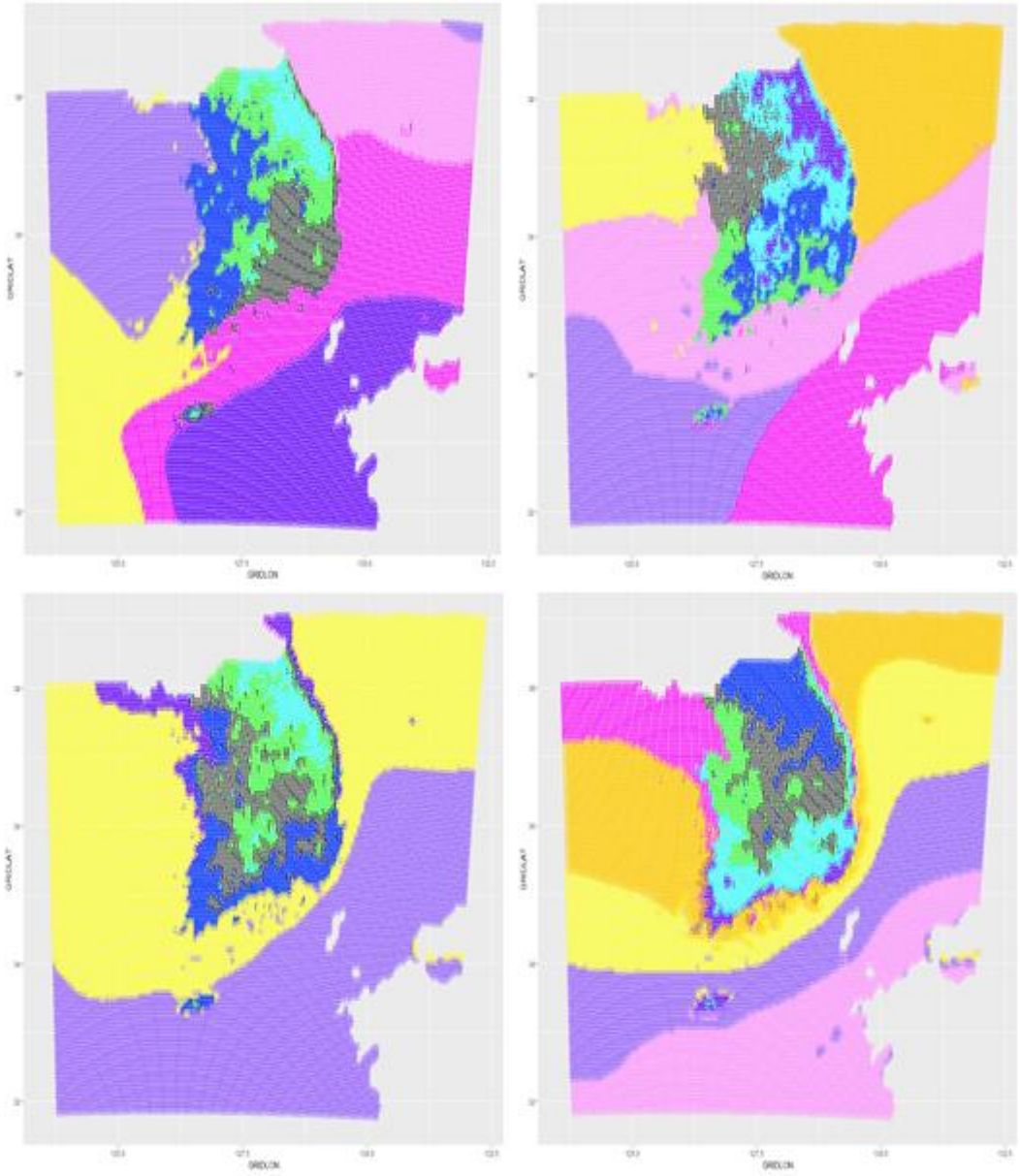

Figure 3.1 Clustering result for temperatures of gridded data of the numerical model (left upper; April, right upper; August, left lower; October , right lower; December)

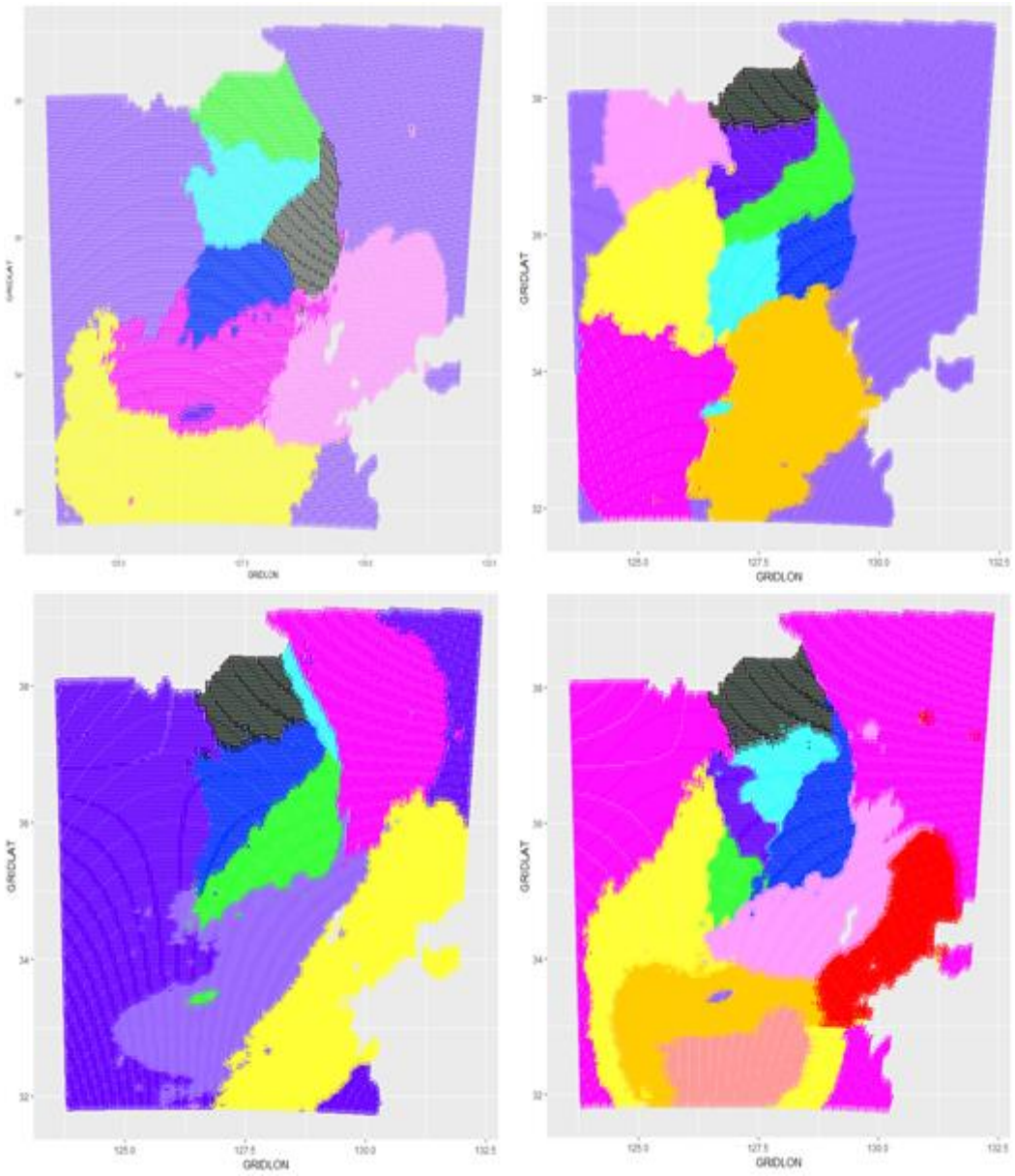

Figure 3.2 Clustering result for temperatures of gridded data of the numerical model (left upper; April, right upper; August, left lower; October , right lower; December)

기온과 강수량에 대한 수치모델 격자자료를 이용하여 군집 분석한 앞의 결과를 바탕으로 525개 기상관측지점에 대해 군집을 할당한 결과가 Table 3.1에 나타나 있다. 1월에 대한 결과를 예로 들면, 2006년 1월부터 2014년 1월까지 9년간의 1월 격자자료를 이용하여 군집분석을 실시하고, 그 결과를 바탕으로 기상관측지점을 군집화 한 결과 기온에 대해서는 8개의 군집으로, 강수량에 대해서는 10개의 군집으로 나뉘는 것을 알 수 있다. 각각의 월에 대해서도 해당 월의 기상특성을 반영하여 군집의 개수가 달라지는 것을 알 수 있다. 또한 기온의 경우는 봄과 여름이 다른 계절에 비해 군집이 더 세분화된 것을 알 수 있다.

Figure 3.3은 기온에 대한 525개 기상관측지점에 대한 군집화 결과로 봄, 여름, 가을, 겨울의 각 계절을 대표하여 4월, 8월, 10월, 12월의 결과가 나타나있다. 군집 결과를 살펴보면 각 월에서 해안가를 따라 기온의 특성이 유사하여 하나의 군집을 형성하는 것을 알 수 있다. Figure 3.4는 강수량에 대한 525개 기상관측지점에 대한 군집화 결과로 각 계절을 대표하여 4월, 8월, 10월, 12월의 결과가 나타나있다. 기온과 마찬가지로 4월, 8월, 10월, 12월에서 모두 해안가를 따라 강수량의 특성이 유사하여 하나의 군집을 형성하는 것을 알 수 있으며, 기온에 비해 위치적인 영향이 군집의 형성에 강하게 나타나는 것을 알 수 있다.

Table 3.1 Clusters by month for temperatures and precipitation of surface observing stations

Month	Number of clusters	
	Temperatures	Precipitation
January	8	10
Februry	7	7
March	9	10
April	9	8
May	10	10
June	9	8
July	12	8
August	10	10
September	8	11
October	7	7
November	7	8
December	9	11

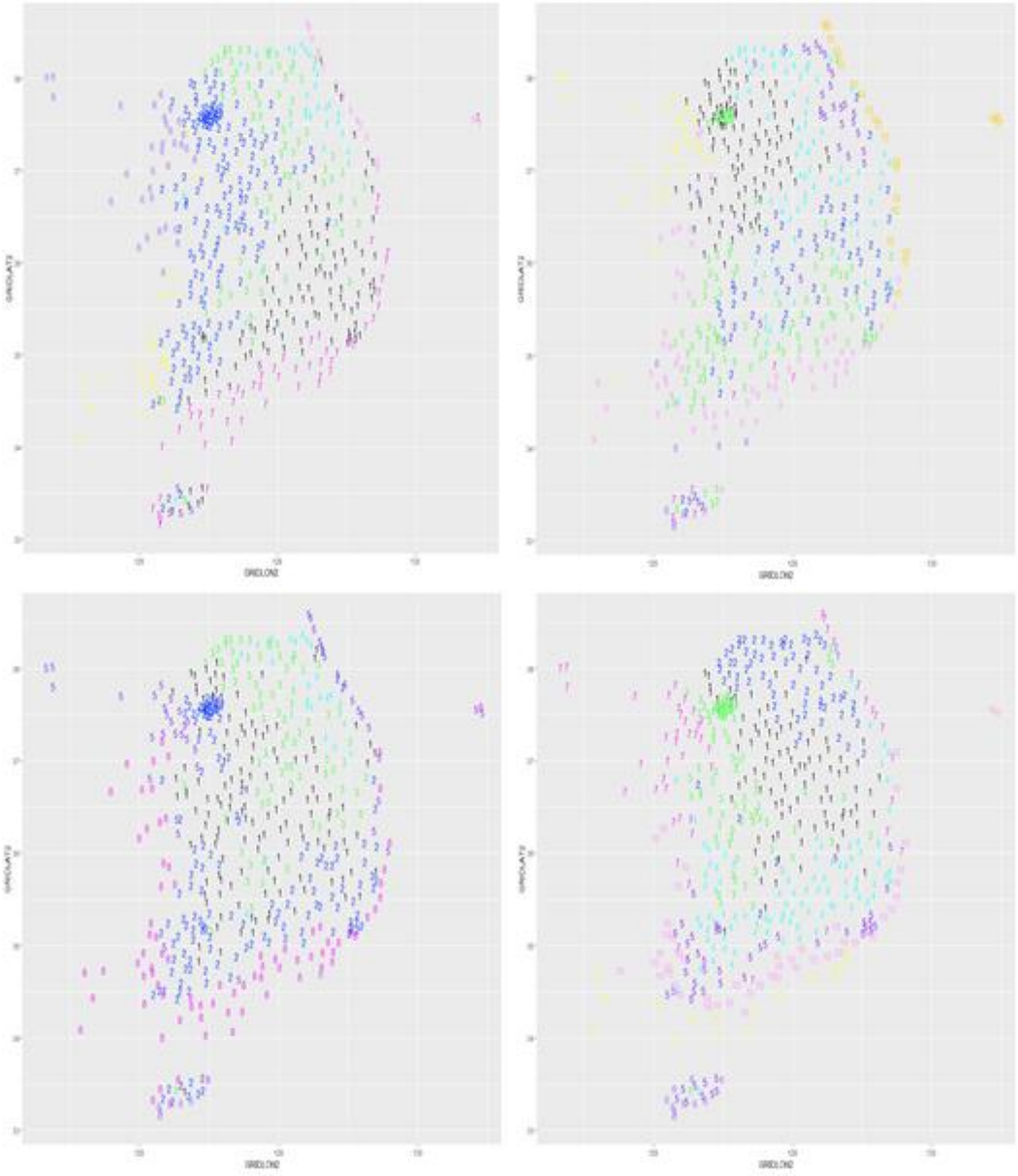

Figure 3.3 Clustering result for temperatures of surface observing stations (left upper; April, right upper; August, left lower; October, right lower; December)

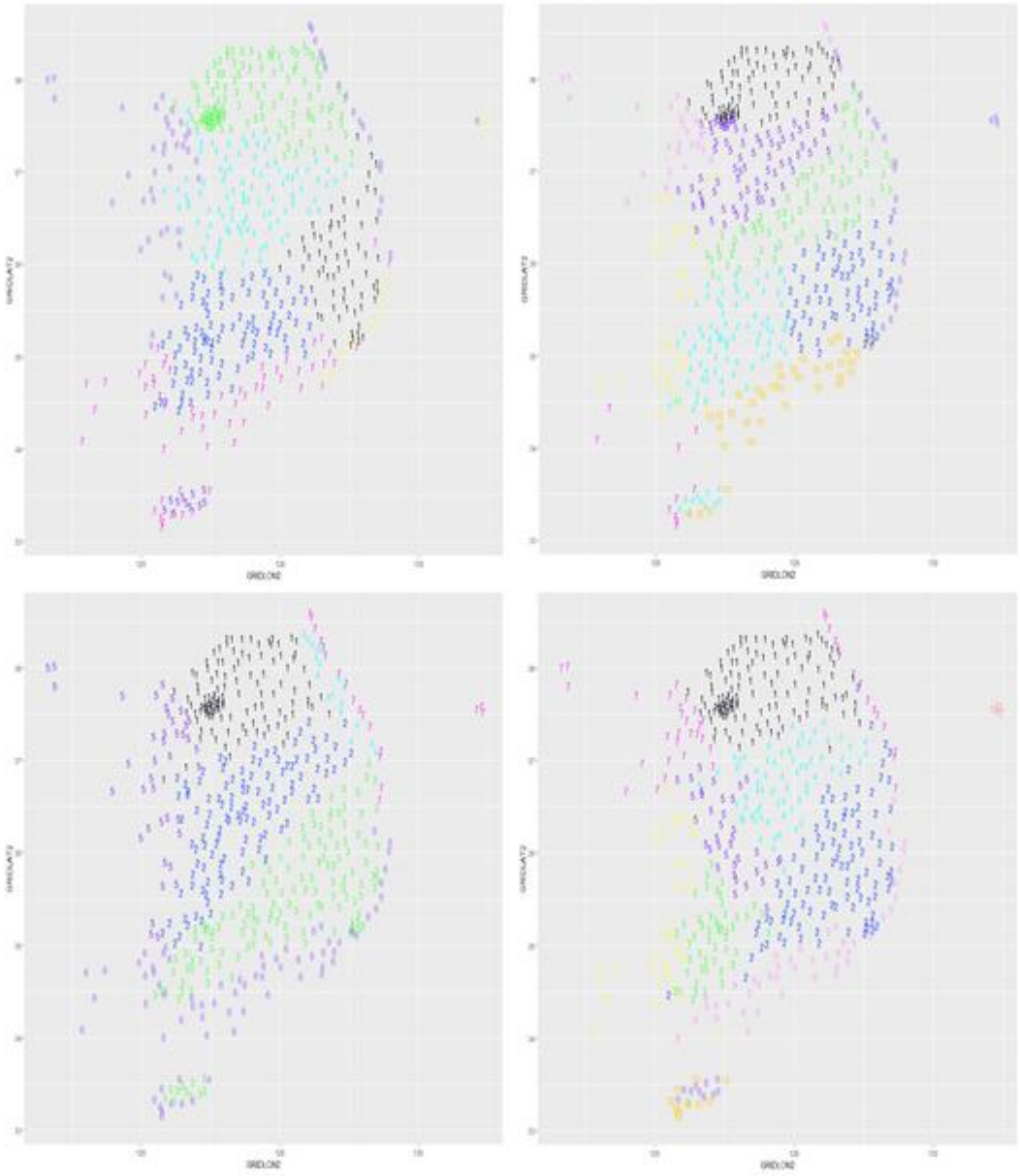

Figure 3.4 Clustering result for temperatures of surface observing stations (left upper; April, right upper; August, left lower; October, right lower; December)

4. 결론

우리나라에는 현재 ASOS와 AWS를 합쳐 600여개의 기상관측지점이 존재하는데 이러한 기상관측지점의 관측자료를 이용하여 군집분석을 수행하게 되면 관측지점의 밀도가 다르기 때문에 남한 지역의 기상 특성이 고르게 반영되지 못할 수 있다. 반면 수치모델 격자자료를 이용하게 되면 5km × 5km 간격으로 조밀하여 위치적으로 고른 자료의 생산이 가능 하므로 남한지역의 기상 특성을 고르게 반영할 수 있다. 본 연구에서는 2006년 1월 1일 부터 2014년 4월 30일까지의 KLAPS 모델 분석장 자료를 이용하여 군집분석을 실시하고 그 결과를 바탕으로 기상관측지점을 군집화 하였다. 군집분석은 바다와 육지로 각각 나누어 실시하였으며, 월별로 기상특성이 상이하므로 군집화 과정도 월별로 각각 진행하였다. 월별로 각각 다른 수치모델 격자자료에 대한 군집결과를 바탕으로 기상관측지점에 대해 위 · 경도 좌표를 이용하여 가장 가까운 격자점을 찾고, 그 격자점과 동일한 군집을 할당하였다. 수치모델 격자자료를 이용하여 군집분석 후 기상관측지점에 군집을 할당하는 것은 격자자료의 조밀한 자료를 바탕으로 이루어졌기 때문에 남한지역의 기상 특성을 고르게 반영하였다고 볼 수 있다. 또한 관측 지점을 직접 이용한 군집화가 아니고 수치모델 격자자료를 이용한 군집화이기 때문에 관측지점의 신설이나 이전이 발생하였을 경우 새로운 관측지점에 대해 재분석 없이 쉽게 군집 지정이 가능할 것이다.

본 연구의 결과를 바탕으로 남한지역의 기상 특성을 고르게 반영한 다양한 군집별 기상 특성 분석이 가능할 것으로 기대된다. 또한 월별 기상특성을 반영한 기상관측지점의 세분화는 각 군집별 관측자료의 세분화된 품질관리에 활용 될 수 있을 것이다. 나아가 기상특성에 따라 시 · 공간적으로 세분화된 품질관리는 고품질 기상관측자료의 생산을 가능하게 할 것이다.

References

Anderberg, M. R. (1973). *Cluster analysis for applications*, Academic Press.

Ju, Y., Jung, H. and Kim, B. (2008). Cluster analysis with Korean weather data: Application of model-based Bayesian clustering method. *Journal of Korean Data & Information Science Society*, **20**, 57-64.

Kim, H. M., Oh, S. K. and Lee, Y. H. (2013). Design of heavy rain advisory decision model based on optimized RBFNNs using KLAPS reanalysis data. *Journal of Korean Institute of Intelligent Systems*, **23**, 473-478.

Kim, J. (2015). Cluster analysis for Seoul apartment price using symbolic data. *Journal of the Korean Data & Information Science Society*, **26**, 1239-1247.

Lee, D. K. and Park, J. G. (1999). Regionalization of summer rainfall in South Korea using cluster analysis. *Journal of Atmospheric Sciences*, **35**, 511-518.

Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S. (2001). Constrained K-means clustering with background knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning*, **18**, 577-584.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236-244.

Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification*, **31**, 247-295.

Yeo, I. K. (2011). Clustering analysis of Korea's meterological data. *Journal of the Korean Data & Information Science Society*, **22**, 941-949.

Yoon, S. and Choi, Y. (2015). Functional clustering for electricity demand data: A case study. *Journal of the Korean Data & Information Science Society*, **26**, 885-894.

Journal of the Korean Data & Information Science Society
2017, **28**(5), 1133–1144

http://dx.doi.org/10.7465/jkdi.2017.28.5.1133
한국데이터정보과학회지

Cluster analysis by month for meteorological stations using a gridded data of numerical model with temperatures and precipitation[†]

Hee-Kyung Kim[1] · Kwang-Sub Kim[2] · Jae-Won Lee[3] · Yung-Seop Lee[4]

[124]Department of Statistics, Dongguk University
[3]KMA National Climate Data Center

Received 17 April 2017, revised 28 July 2017, accepted 14 August 2017

Abstract

Cluster analysis with meteorological data allows to segment meteorological region based on meteorological characteristics. By the way, meteorological observed data are not adequate for cluster analysis because meteorological stations which observe the data are located not uniformly. Therefore the clustering of meteorological observed data cannot reflect the climate characteristic of South Korea properly. The clustering of 5km×5km gridded data derived from a numerical model, on the other hand, reflect it evenly. In this study, we analyzed long-term grid data for temperatures and precipitation using cluster analysis. Due to the monthly difference of climate characteristics, clustering was performed by month. As the result of K-Means cluster analysis is so sensitive to initial values, we used initial values with Ward method which is hierarchical cluster analysis method. Based on clustering of gridded data, cluster of meteorological stations were determined. As a result, clustering of meteorological stations in South Korea has been made spatio-temporal segmentation.

Keywords: Cluster analysis, gridded data of numerical model, K-Means method, meteorological stations, precipitation, temperatures, Ward method.

[†] This work was funded by the Korea Meteorological Administration Research and Development Program under Grant KMIPA 2015-1020 and the first author, Hee-Kyung Kim, was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1A6A3A11931946).
[1] Post Doc. Researcher, Department of Statistics, Dongguk University-Seoul, Seoul 04620, Korea.
[2] Graduate student, Department of Statistics, Dongguk University-Seoul, Seoul 04620, Korea.
[3] Director, KMA National climate data center, Seoul 07062, Korea.
[4] Corresponding author: Professor, Department of Statistics, Dongguk University-Seoul, Seoul 04620, Korea. Email: yung@dongguk.edu