Journal of the Korean Data & Information Science Society
2011, **22**(5), 941–949

한국데이터정보과학회지

우리나라 기상자료에 대한 군집분석[†]

여인권[1]

[1]숙명여자대학교 통계학과

접수 2011년 8월 25일, 수정 2011년 9월 18일, 게재확정 2011년 9월 22일

요 약

이 논문에서는 1999년 1월 1일부터 2010년 6월 30일까지 전국 72개 관측소에서 측정된 우리나라 기상자료를 평균연결법에 의한 계층적 병합방법을 통해 군집분석을 실시하고 각 기상자료에서 유도된 군집의 특성을 파악해 본다. 이 분석에서 유도된 군집과 2010년 기후변화에 따른 식중독 발생연구에서 사용되었던 산맥을 경계로 구분한 군집을 비교해 본다.

주요용어: 계층적 병합절차, 기상관측소, 수목도.

1. 머리말

지난 20년간 온실가스의 급격한 증가로 인해 지구 표면 온도가 높아지면서 여러 가지 문제가 발생함에 따라 기후변화에 대한 많은 논의가 이루어지고 있으며 기후변화에 따른 생태계 변화 및 관련 질병의 발생 등에 대한 연구가 진행되고 있다. 우리나라에서는 2009년부터 식약청 용역과제로 기후변화에 따른 식중독 발생영향에 대한 연구가 진행되고 있는데 이 연구에서는 기상자료가 식중독 발생에 어떻게 영향을 주는지에 대한 다양한 분석을 진행하고 있다 (정명섭과 오상석, 2009). 2010년도 지속과제에서는 전국에 분포되어 있는 기상관측소 중 60개를 선정하여 일별, 주별, 월별 평균을 기상자료로 사용하였으며 기존 연구보다 세밀한 결과를 얻기 위해 60개 기상관측소를 태백산맥과 소백산맥을 기준으로 세 권역으로 나누고 각 권역에 속해 있는 기상관측소에서 수집된 기상자료의 평균을 이용하여 식중독 발생빈도에 기후가 어떻게 영향을 주는지에 대한 분석을 실시하였다. 이 논문에서는 산맥을 경계로 권역을 나눈 것이 기상자료의 특성을 반영하여 권역을 나누었다고 볼 수 있는지를 기상자료의 군집분석을 통해 확인해 보고 이를 통해 차후에 기상자료를 그룹화 하여 이루어지는 분석에 기본 틀을 제공하고자 한다.

시계열자료에 대한 군집분석 방법은 크게 세 가지로 나누어지고 있다. 첫 번째 방법은 시계열자료들 간의 거리 또는 유사성을 측도로 사용하는 원자료기반방법 (row-data-based approach)이고 두 번째 방법은 시계열 자료를 저차원의 특성벡터 (feature vector)를 추출하고 이 특성을 비교해 군집하는 특성기반방법 (feature-based approach), 세 번째 방법은 모형을 설정하고 모형에 포함된 모수의 추정값 또는 잔차를 비교하여 군집하는 모형기반방법 (model-based approach)이다. 이들 방법의 적절성은 자료의 형태 및 구조에 영향을 받으며 어떤 방법이 더 우수하다고 할 수 없으나 본 논문에서는 특성벡터의 추출 및 모형선택 과정 등에서 발생할 수 있는 오류를 없애기 위해 원자료기반방법을 중심으로 우리나라 기상자료에 대한 군집분석을 실시하였다. 이들 방법에 대한 자세한 내용은 Liao (2005)를 참조하기 바란다.

[†] 본 연구는 숙명여자대학교 2010년도 교내연구비 지원에 의해 수행되었음.
[1] (140-742) 서울특별시 용산구 청파동 2가, 숙명여자대학교 통계학과, 교수.
E-mail: inkwon@sookmyung.ac.kr

국내 기상자료를 이용한 기후연구는 장학진과 주용성 (2009), 김동석 등 (2009), 이훈자 (2010)에 의해 실시되었으며 기후자료에 대한 군집분석 연구는 Kim과 Lim (2003), 주용성 등 (2009), 김현구 등 (2010)에 의해 수행되었다. Kim과 Lim (2003)은 1999년 서울에서 측정된 환경오염물질과 기상자료 간의 관계를 알아보기 위한 연구에서 군집분석을 실시하였으며 주용성 등 (2009)은 우리나라 30개 주요 도시의 2007년 월별 강수량, 온도, 풍속, 일조량, 습도 자료에 Booth 등 (2008)의 베이지안 모델기반 군집분석 방법을 적용하여 기상변동추이는 경도와 위치 (내륙 혹은 해안지역)에 의해서도 영향을 많이 받는 것을 보였다. 김현구 등 (2010)은 제주도에서의 풍력발전예보 모형개발을 위해 20개 관측지점의 풍속을 대상으로 군집분석을 실시한 것이 있다. 이들 연구에서 사용된 기상자료는 제한된 기간이나 장소에서 수집된 자료로 한정되고 있다. 이 논문에서 우리나라의 전반적인 결과를 얻기 위해 1999년 1월 1일부터 2010년 6월 30일까지의 전국 72개 기상관측소에서 수집된 4198개 일별 평균기온, 최저기온, 최고기온, 평균습도, 강수량, 신적설량 자료를 이용하였다. 표 1.1은 이들 기상관측소와 기후변화에 따른 식중독 발생 예측 연구에서 설정한 권역을 정리한 것이다.

표 1.1 권역별 72개 기상관측소명

권역	관측소	권역	관측소	권역	관측소	권역	관측소
1권역	강릉	2권역	양평	3권역	구미	3권역	영주
1권역	대관령	2권역	영월	3권역	군산	3권역	영천
1권역	동해	2권역	원주	3권역	남원	3권역	완도
1권역	속초	2권역	이천	3권역	남해	3권역	울산
1권역	영덕	2권역	인제	3권역	대구	3권역	의성
1권역	울릉도	2권역	인천	3권역	마산	3권역	임실
1권역	울진	2권역	제천	3권역	목포	3권역	장수
1권역	태백	2권역	철원	3권역	문경	3권역	장흥
1권역	포항	2권역	청주	3권역	밀양	3권역	전주
2권역	금산	2권역	춘천	3권역	봉화	3권역	정읍
2권역	대전	2권역	충주	3권역	부산	3권역	제주
2권역	동두천	2권역	홍천	3권역	부안	3권역	진주
2권역	보령	3권역	강화	3권역	산청	3권역	천안
2권역	보은	3권역	거제	3권역	서귀포	3권역	추풍령
2권역	부여	3권역	거창	3권역	성산	3권역	통영
2권역	서산	3권역	고산	3권역	순천	3권역	합천
2권역	서울	3권역	고흥	3권역	안동	3권역	해남
2권역	수원	3권역	광주	3권역	여수	3권역	흑산도

2. 기상자료의 군집분석

이 논문에서는 연결법 (linkage method)에 의한 계층적 병합절차 (hierarchical agglomerative procedure)로 평균기온, 최고기온, 최저기온, 평균습도, 강수량, 신적설량 각각에 대해 군집분석을 실시하였으며 수목도 (dendrogram)로 군집분석의 결과를 표시하였다. 관측값 x_{ij}를 i번째 기상관측소의 j번째 시점에서 관측된 기상자료의 값이라고 하고 $\boldsymbol{x_i} = (x_{i1}, \cdots, x_{\in})^T$라고 하면, 임의의 두 기상관측소 a와 b의 기상자료들 간의 거리는 다음과 같은 Minkowski 거리를 이용하여 계산할 수 있다.

$$d(\boldsymbol{x_a}, \boldsymbol{x_b}, p) = \left(\sum_{k=1}^{n} |x_{aj} - x_{bj}|^p \right)^{1/p}$$

여기에서는 각각의 기상자료에 대해 $p = 2$인 Euclidean 거리를 계산하여 72×72 거리행렬을 구했으며 평균연결법 (average linkage method)을 이용하여 군집분석을 실시하였다. 분석은 R 2.13.0에서

cluster package에 있는 agnes함수로 비유사성 및 수목도를 작성하고 cutree함수를 이용하여 5개의 군집을 표시하였다.

기상자료는 계절성이 있기 때문에 전체 평균과 같은 단순한 대표값으로는 각 지역의 특성을 비교하기 어렵다. 이 논문에서는 군집들 간의 차이를 비교하기 위해 각 기상자료에서 그 날의 전국평균을 뺀 값을 이용하여 군집평균을 계산하였는데 군집 A의 군집평균은 다음과 같이 계산된다.

$$\bar{C_A} = \frac{1}{N_A} \sum_{i \in A} \sum_{j=1}^{n} (x_{ij} - \bar{x_j})$$

여기서 $\bar{x_j}$는 j번째 시점에서의 전국평균이고 N_A는 A군집에 속하는 기상관측소의 전체 자료의 개수를 나타낸다.

2.1. 평균기온에 의한 군집분석

71×72/2=2556개 비유사성 (dissimilarity)을 계산한 결과, 평균거리는 192.27이고 양평과 이천의 거리가 43.27로 최소이었으며 대관령과 서귀포의 거리가 687.28로 최대인 것으로 나타났다. 군집을 5개로 했을 때 각 군집에 해당되는 기상관측소는 아래 표 3.1에 표시 하였는데 평균기온에 의한 군집분석 결과 우리나라 기상관측소는 동해안지역, 태백산맥지역, 중서부남부지역, 중동부지역, 제주지역으로 나누어지는 것으로 나타났으며 각 군집 평균은 $\bar{C}$ =(0.189, -4.033, 0.760, -1.369, 3.384)으로 태백산맥에 위치한 기상관측소의 평균기온이 상대적으로 낮고 제주지역 기상관측소의 평균기온이 높은 것으로 나타났다.

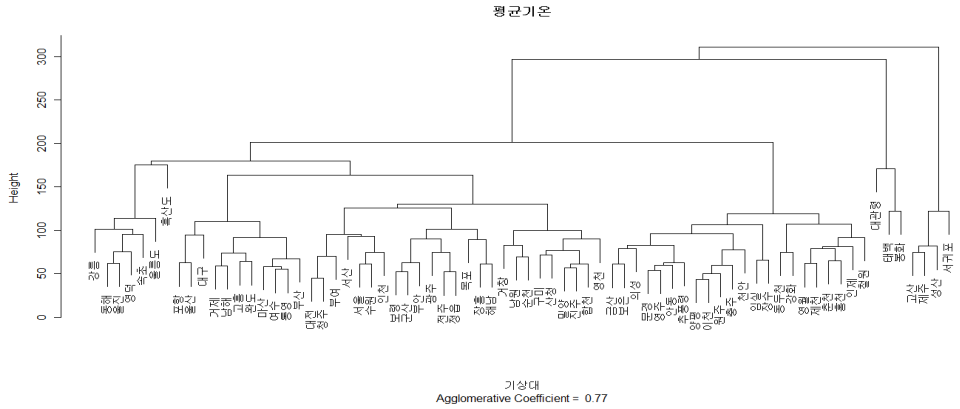

그림 2.1 평균기온에 의한 기상관측소 수목도

2.2. 최고기온에 의한 군집분석

최고기온의 경우 평균거리는 1212.30이고 전주와 정읍의 거리가 61.14로 최소였으며 평균기온과 마찬가지로 대관령과 서귀포의 거리가 616.61로 최대인 것으로 나타나 평균기온보다는 전반적으로 기상관측소 간의 거리가 더 먼 것으로 나타났다. 군집을 5개로 했을 때 동해안지역, 태백산맥지역, 해안지역, 내륙지역, 흑산도로 군집이 나누어지는 것으로 나타났으며 각 군집 평균은 $\bar{C}$ =(-1.166, -4.687, 1.081, -0.449, -1.965)으로 태백산맥에 위치한 기상관측소의 최고기온이 상대적으로 낮은 반면 해안지역 기상관측소의 최고기온이 높은 것으로 나타났다.

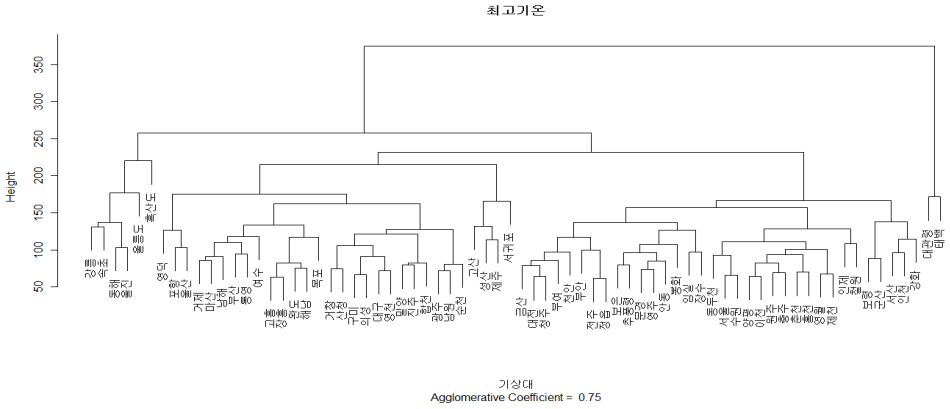

그림 2.2 최고기온에 의한 기상관측소 수목도

2.3. 최저기온에 의한 군집분석

최저기온의 평균거리는 222.03이고 대전과 청주의 거리가 59.01로 최소이었으며 평균기온, 최고기온과 마찬가지로 대관령과 서귀포의 거리가 800.63으로 최대인 것으로 나타났다. 군집을 5개로 나누었을 때 남동해안지역, 태백산맥지역, 내륙지역, 해안지역, 제주지역으로 군집이 나누어지는 것으로 나타났으며 각 군집 평균은 $\bar{C}$ =(2.096, -4.664, -1.965, 0.243, 5.101)으로 태백산맥에 위치한 기상관측소의 최저기온이 상대적으로 낮고 제주권의 최저기온이 높은 것으로 나타났다.

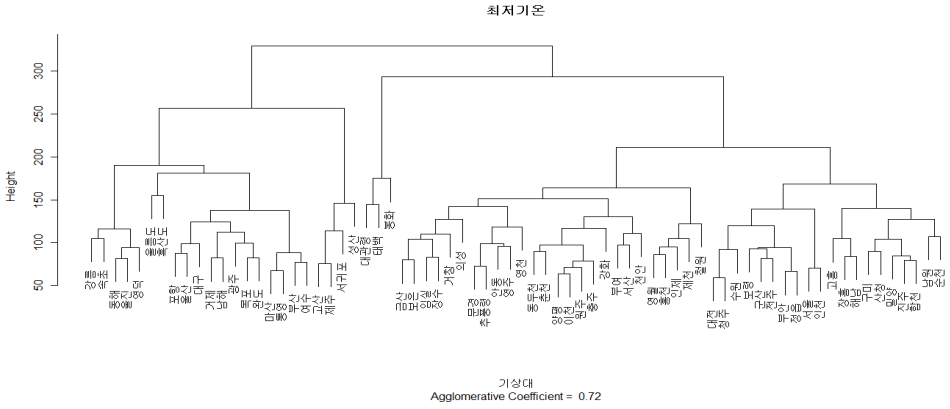

그림 2.3 최저기온에 의한 기상관측소 수목도

2.4. 평균습도에 의한 군집분석

평균습도의 평균거리는 823.26이고 남원과 임실의 거리가 299.16로 최소이었으며 강릉과 흑산도의 거리가 1640.54으로 최대로 조사되었다. 군집을 5개로 했을 때 동해안지역, 대관령울릉도지역, 남해안지역, 제주지역, 기타지역으로 군집이 나누어지는 것으로 나타났으며 각 군집 평균은 $\bar{C}$ =(-4.825,

6.230, -3.459, 4.193, 0.615)으로 대관령울릉도 및 제주지역의 평균습도가 상대적으로 높은 반면 습도가 높을 것으로 예상된 해안지역인 동해안과 남해안 지역이 오히려 낮은 것으로 나타났다.

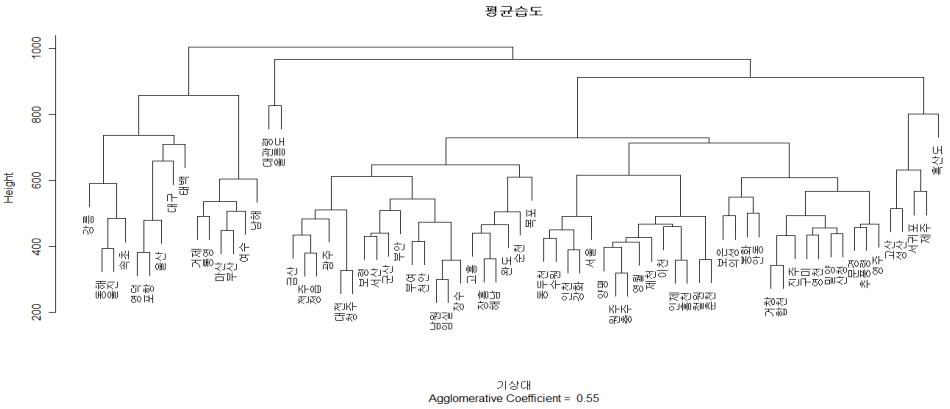

그림 2.4 평균습도에 의한 기상관측소 수목도

2.5. 강수량에 의한 군집분석

강수량의 경우 평균거리는 923.51이고 대구와 영천의 거리가 295.15로 최소였으며 강릉과 서귀포의 거리가 1557.40으로 가장 많은 차이를 보이는 것으로 조사되었다. 군집을 5개로 했을 때 강원동북부지역, 중북부지역, 남해안지역, 제주지역, 기타지역으로 군집이 나누어지는 것으로 나타났으며 각 군집 평균은 $\bar{C}$ =(0.578, -0.067, 0.556, 0.794, -0.283)으로 제주지역 및 해안지역의 강수량이 상대적으로 많은 것으로 나타났다.

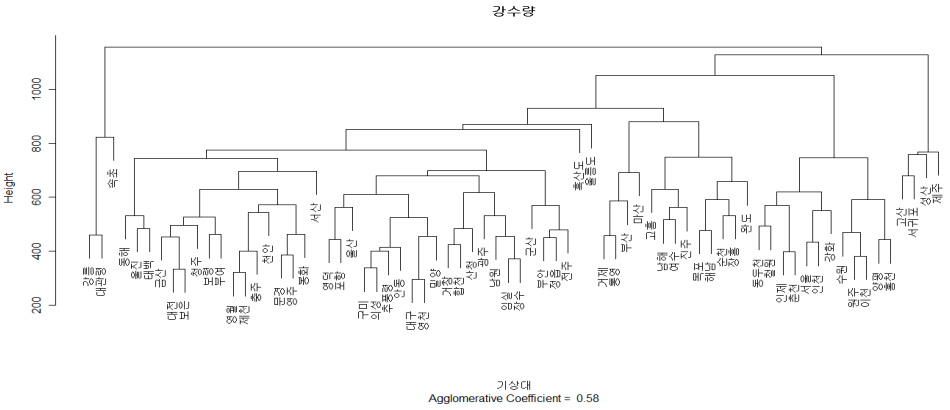

그림 2.5 강수량에 의한 기상관측소 수목도

2.6. 신적설량에 의한 군집분석

신적설량의 평균거리는 73.06이고 눈이 거의 내리지 않는 거제와 마산의 거리가 8.63으로 최소였으며 눈이 많이 내리는 대관령과 울릉도의 거리가 294.57으로 최대인 것으로 나타났는데 이것은 두 지역

은 시차를 두고 눈이 내리는 경향이 있다는 것을 의미한다. 군집을 5개로 했을 때 동해안북부지역, 대관령, 울릉도, 태백, 기타지역으로 군집이 나누어지는 것으로 나타났으며 각 군집 평균은 $\bar{C}$ =(0.057, 0.496, 0.537, 0.142, -0.020)으로 대관령과 울릉도가 상대적으로 많이 내리는 것으로 나타났으며 제주지역의 경우 눈이 내리는 패턴에서 큰 차이가 없어 별개의 군집으로 분류되지 않은 것으로 조사되었다.

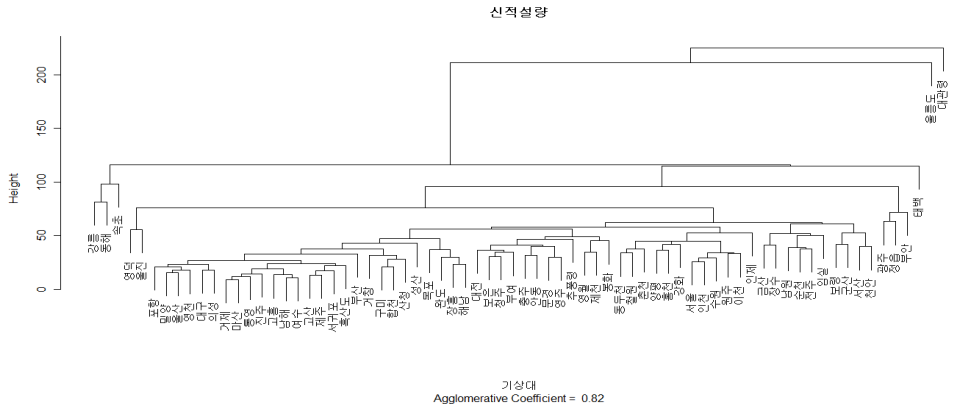

그림 2.6 신적설량에 의한 기상관측소 수목도

3. 군집분석 요약 및 결론

표 3.1은 각 기상자료에 대한 군집분석 결과를 정리한 것으로 평균기온과 최고기온에 의한 군집이 대체로 비슷한 것으로 나타났으며 신적설량의 경우 대관령, 울릉도, 태백에서 눈이 시차를 두고 많이 내려 세 지역 간의 신적설량 거리가 커지면서 각각이 군집이 나누어져 나머지 기상관측소들의 분류가 세분화되지 못하는 문제가 발생했다.

기후변화에 따른 식중독 발생 영향에 대한 2010년 연구에서 적용한 권역과 비교해 보면 다음과 같은 결과를 유도할 수 있다.

- 1권역의 경우 대관령과 태백을 제외하고 평균기온, 최고기온, 최저기온, 평균습도에서 유사한 군집으로 평가할 수 있다.
- 2권역의 경우 최고기온의 관점에서 동일 군집에 속하는 것으로 보이며 평균기온, 최저기온, 평균습도에서도 많은 관측소가 같은 군집에 포함되는 것을 볼 수 있다.
- 3권역의 경우 여러 군집으로 나눠지고 있으며 기상자료에 따라 군집의 형태도 달라지는 것을 볼 수 있으며 특히 제주지역은 평균기온, 최저기온, 평균습도, 강수량에서 별개의 군집으로 나누어질 수 있다.

기상자료별로 군집을 이루는 형태에 약간의 차이가 있는 것을 볼 수 있으며 기존 식중독 발생 예측연구에서 사용한 산맥을 경계로 나눈 권역에서 3권역이 상이한 형태의 군집으로 묶여지는 경향이 있는 것으로 조사되었다. 본 연구에서는 개별 기상자료에 대한 군집분석으로 기상자료에 따라 약간씩 다른 결과가 도출되어 통합적 군집을 유도하지 못한 한계점이 있다. 현재 기상관측소에서는 이 논문에서 고려한 6개의 자료 외에 풍향, 최대풍속, 일조량, 일사량, 운(구름)량 등 다양한 자료가 수집되고 있으며 이

들 자료는 분석 또는 사용목적에 따라 필요여부가 결정되고 이에 따른 군집유도가 필요하기 때문에 모든 변수를 고려해 군집을 정하는 것은 적절하지 않는 것을 생각된다. 앞에서 언급한 식중독 발생 예측연구에서는 식중독 발생에는 기온과 습도가 중요한 설명변수로 파악되었는데 차후 연구에서는 Kakizawa 등(1998)과 Singhal과 Seborg (2005) 등에 의해 연구된 다변량시계열 군집분석 방법을 통해 이들 변수들을 통합적으로 분석할 예정이다.

표 3.1 각 기상자료에 의한 기상관측소 군집분석 결과

권역	관측소명	평균기온	최고기온	최저기온	평균습도	강수량	신적설량	권역	관측소명	평균기온	최고기온	최저기온	평균습도	강수량	신적설량
1권역	강릉	1	1	1	1	1	1	3권역	구미	3	3	4	3	2	3
1권역	대관령	2	2	2	2	1	2	3권역	군산	3	4	4	3	2	3
1권역	동해	1	1	1	1	2	1	3권역	남원	3	3	4	3	2	3
1권역	속초	1	1	1	1	1	1	3권역	남해	3	3	1	4	4	3
1권역	영덕	1	3	1	1	2	3	3권역	대구	3	3	1	1	2	3
1권역	울릉도	1	1	1	2	2	4	3권역	마산	3	3	1	4	4	3
1권역	울진	1	1	1	1	2	3	3권역	목포	3	3	1	3	4	3
1권역	태백	2	2	2	1	2	5	3권역	문경	4	4	3	3	2	3
1권역	포항	3	3	1	1	2	3	3권역	밀양	3	3	4	3	2	3
2권역	금산	4	4	3	3	2	3	3권역	봉화	2	4	2	3	2	3
2권역	대전	3	4	4	3	2	3	3권역	부산	3	3	1	4	4	3
2권역	동두천	4	4	3	3	3	3	3권역	부안	3	4	4	3	2	3
2권역	보령	3	4	4	3	2	3	3권역	산청	3	3	4	3	2	3
2권역	보은	4	4	3	3	2	3	3권역	서귀포	5	3	5	5	5	3
2권역	부여	3	4	3	3	2	3	3권역	성산	5	3	5	5	5	3
2권역	서산	3	4	3	3	2	3	3권역	순천	3	3	4	3	4	3
2권역	서울	3	4	4	3	3	3	3권역	안동	4	4	3	3	2	3
2권역	수원	3	4	4	3	3	3	3권역	여수	3	3	1	4	4	3
2권역	양평	4	4	3	3	3	3	3권역	영주	4	4	3	3	2	3
2권역	영월	4	4	3	3	2	3	3권역	영천	3	3	3	3	2	3
2권역	원주	4	4	3	3	3	3	3권역	완도	3	3	1	3	4	3
2권역	이천	4	4	3	3	3	3	3권역	울산	3	3	1	1	2	3
2권역	인제	4	4	3	3	3	3	3권역	의성	4	3	3	3	2	3
2권역	인천	3	4	4	3	3	3	3권역	임실	4	4	3	3	2	3
2권역	제천	4	4	3	3	2	3	3권역	장수	4	4	3	3	2	3
2권역	철원	4	4	3	3	3	3	3권역	장흥	3	3	4	3	4	3
2권역	청주	3	4	4	3	2	3	3권역	전주	3	4	4	3	2	3
2권역	춘천	4	4	3	3	3	3	3권역	정읍	3	4	4	3	2	3
2권역	충주	4	4	3	3	2	3	3권역	제주	5	3	5	5	5	3
2권역	홍천	4	4	3	3	3	3	3권역	진주	3	3	4	3	4	3
3권역	강화	4	4	3	3	3	3	3권역	천안	4	4	3	3	2	3
3권역	거제	3	3	1	4	4	3	3권역	추풍령	4	4	3	3	2	3
3권역	거창	3	3	3	3	2	3	3권역	통영	3	3	1	4	4	3
3권역	고산	5	3	5	5	5	3	3권역	합천	3	3	4	3	2	3
3권역	고흥	3	3	4	3	4	3	3권역	해남	3	3	4	3	4	3
3권역	광주	3	3	1	3	2	3	3권역	흑산도	1	5	1	5	2	3

참고문헌

김동석, 홍수진, 박준표 (2009). 대구지역 지중온도의 변화예측. <한국데이터정보과학회지>, **20**, 649-654.

김현구, 이영섭, 장문석 (2010). 제주도 일단위 풍력발전예보 모형개발을 위한 군집분석 및 기상통계모형 실험. <한국환경과학회지>, **10**, 1229-1235.

이훈자 (2010). 경기도 수원시 미세먼지 농도의 시계열모형 연구. <한국데이터정보과학회지>, **21**, 1117-1124.

장학진, 주용성 (2009). 서울의 온도 패턴 변화. <한국데이터정보과학회지>, **20**, 89-96.

정명섭, 오상석 (2009). <기후변화에 따른 식중독 발생 영향분석 및 관리 체계 연구>, 식품의약품안전청 최종보고서 정책-식품-2009- 09, 서울.

주용성, 정형주, 김병준 (2009). 한국 기상자료의 군집분석: 베이지안 모델기반 방법의 응용. <한국데이터정보과학회지>, **20**, 57-64.

Booth, J., Casella, G. and Hobert, J. (2008). Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society B*, **70**, 119-140.

Kakizawa, Y., Shumway, R. H. and Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, **93**, 328-340.

Kim, J. and Lim, J. (2003). Cluster analysis with air pollutants and meteorological factors in Seoul. *Journal of the Korean Data & Information Science Society*, **14**, 773-787.
Liao, T. W. (2005). Clustering of time series data-A survey. *Pattern Recognition*, **38**, 1857-1874.
Singhal, A. and Seborg, D. E. (2005). Clustering multivariate time-series data. *Journal of Chemometrics*, **19**, 427-438.

Journal of the Korean Data & Information Science Society
2011, **22**(5), 941–949

한국데이터정보과학회지

Clustering analysis of Korea's meteorological data[†]

In-Kwon Yeo[1]

[1]Department of Statistics, Sookmyung Women's University

Received 25 August 2011, revised 18 September 2011, accepted 22 September 2011

Abstract

In this paper, 72 weather stations in Korea are clustered by the hierarchical agglomerative procedure based on the average linkage method. We compare our clusters and stations divided by mountain chains which are applied to study on the impact analysis of foodborne disease outbreak due to climate change.

Keywords: Dendrogram, hierarchical agglomerative procedure, weather station.

[†] This research was supported by the Sookmyung Women's University Research Grants 2010.
[1] Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.