Journal of the Korean Data & Information Science Society
2009, **20**(1), 57–64

한국데이터정보과학회지

한국 기상자료의 군집분석: 베이지안 모델기반 방법의 응용

주용성[1] · 정형주[2] · 김병준[2]

[12]동국대학교 통계학과

접수 2008년 11월 11일, 수정 2008년 12월 22일, 게재확정 2008년 12월 31일

요 약

이 논문에서는 한국 30개 주요도시를 강수량, 온도, 풍속, 일조량, 습도를 기준으로 군집분석을 행하였다. 군집분석 결과는 지형적 특성에 이 들 기상변수가 큰 영향을 받는 다는 것으로 나타났다. 한국은 비록 작은 나라이기는 하지만, 지형성 영향을 많이 받는 것으로 알려져 있기 때문에 우리의 연구결과는 기상에 관한 기존상식과 일치한다고 이야기 할 수 있다. 풍속을 기준으로 군집분석을 하였을 때, 가장 많은 수의 군집들이 찾아졌고 일조량을 기준으로 했을 때 가장 작은 수의 군집이 찾아졌다. 풍속을 기준으로 했을 때 많은 군집들이 찾아지는 것은 바람은 국소지형에 아주 많은 영향을 받기 때문이라고 여겨진다.

주요용어: 강수량, 군집분석, 스플라인, 프로파일.

1. 서론

세계는 현재 지구 온난화에 따른 지구촌 기상 이변으로 식량 부족, 해수면 상승, 가뭄 및 사막화 등과 같은 심각한 자연재해 현상에 직면하였다. 국제사회는 이에 적극적으로 대처하기 위해 1988년 UN총회 결의에 따라 세계기상기구 (WMO)와 유엔환경계획 (UNEP)에 "기상변화에 관한 정부간 패널 (IPCC)"을 설치하였고, 1992년 6월 유엔환경개발회의 (UNCED)에서 기상변화협약 (UNFCCC)을 채택하였다. 이처럼 국제 사회가 기상변화에 대해 관심을 가지고, 현황과 전망에 관한 연구를 활발하게 진행하고 있는 가운데 우리나라는 1993년 12월에 IPCC에 세계 47번째로 가입하며 기상변화에 관해 국제사회와 공감대를 형성하고 있다. 이는 한국사회가 기상변화 문제는 범세계적인 것이며 지속적인 인류의 공존, 공영에 밀접하게 연관되어 있다는 것을 인식하고 있는 것이다.

최근 미얀마 싸이클론 나르기스나 미국 플로리다지역에 이례적으로 자주 나타나는 허리케인 등 세계 곳곳에서 이상기상에 따른 자연재해가 발생하고 있다. 또한 우리나라의 경우에도 과거보다 현저히 증가한 기습적이고 다발적인 호우와 지표면 온도상승, 해안지역의 적조현상 등 지구 온난화에 따른 이상기상 현상이 지속적으로 발생하고 있다. 이로 인하여, 한반도의 기상이 변화하고 있다는 것을 대부분의 국민들이 동감하고 있는 상태이다. 이러한 사회적 인식과 더불어, 기상에 관한 학문적 관심이 높아지면서, 많은 기상연구 (이용희 등, 2000; Kim 등, 2002; Shon 등, 2002)들이 통계학자들에 의해 진행되었다.

본 연구는 기상자료를 이용하여 지역 (도시)들의 군집을 베이지안 군집분석 방법을 통해 알아보고자 하는 것에 그 목적이 있다. 한국 기상이 국소적 지형에 영향을 많이 받는 경향을 가지기 때문에, 군집들은 지형적 특성과 많은 연관성을 가질 것으로 예상할 수 있다. 이 결과는 한국의 기상과 지역적 특

[1] 교신저자: (100-715) 서울시 중구 필동 3-26, 동국대학교 통계학과, 조교수.
E-mail: yongsungjoo@dongguk.edu
[2] (100-715) 서울시 중구 필동 3-26, 동국대학교 통계학과, 석사과정.

성에 관한 더욱 심도 깊은 향후연구를 하는데 큰 도움이 되리라고 생각된다. 자료로는 기상청 홈페이지 (http://www.kma.go.kr/)에 나와 있는 2007년 강수량, 온도, 풍속, 일조량, 습도 측정치의 월별 평균값을 이용하였다. 자료가 수집된 30개 도시들은 표1.1에 나와 있다. 베이지안 모델기반 군집분석 방법은 Booth 등 (2008)에 의해 최초로 개발되었으며, 다차원에 있는 점들의 군집을 찾는 전통적인 K-means나 계층적 군집분석법들과는 달리, 곡선으로 나타나는 함수형 자료의 군집을 찾는다는 데에 그 특성이 있다. 사용된 군집분석 방법은 2절에 자세히 서술되어 있으며, 그 분석결과는 3장에 설명되어 있다.

표 1.1 기상자료가 수집된 30개 도시

광역지역	도시	광역지역	도시
경기도	수원, 동두천, 양평	전라북도	남원, 정읍, 군산
강원도	속초, 원주, 태백, 강릉	전라남도	여수, 목포, 광주
충청북도	청주, 제천, 보은	경상북도	상주, 울진, 포항
충청남도	서산, 천안	경상남도	산청, 밀양
특수행정도시	서울, 인천 (경기), 대전 (충남), 대구 (경북), 부산 (경남), 제주도, 울릉도		

2. 베이지안 모델기반 군집분석

이 장에서는 Booth 등 (2008)에 나와 있는 방법들을 설명하도록 하겠다. 이 방법은 다음과 같은 군집우도함수를 이용해서 군집들을 모형화 하게 된다.

$$f(Y|\theta,\omega)=\prod_{k=1}^{c(\omega)} f(Y_{C_k}|\theta_k) \tag{2.1}$$

여기에서 Y는 자료, ω는 n개의 함수형 관측치들의 분할, $C(\omega)$는 분할들의 개수, $\theta = \{\ \theta_1, \theta_2, \cdots, \theta_{C(\omega)}\ \}$는 각각의 군집의 특성을 나타내는 모수 (θ_k)의 집합, C_k는 군집 k안에 들어가는 함수형 관측치들의 소속 군집을 나타내는 값들의 집합을 나타낸다. 또한, $\bigcup_{k=1}^{C(\omega)} C_k = \{1,\cdots,n\}, i \neq j$일 때에 $C_i \cap C_j = \emptyset$이 된다. 각 군집들은 결합 확률밀도함수 $f(Y_{C_k}|\theta_k)$를 형성하고 있다.

이 논문에서 우리는 군집분석 방법을 연중 강수량, 기온, 풍속, 일조량, 습도 변동에 각기 따로 적용하게 되는데, 각각의 자료를 다음과 같이 표시하도록 하겠다.

$$Y=(Y_1^T,\cdots,Y_i^T,\cdots,Y_n^T\ )^T,\ Y_i=(Y_{i1},\cdots,Y_{it},\cdots,Y_{ip}).$$

여기서 Y_{it}는 i번째 도시에서 t월에 관찰된 값이고, 12개월 동안 30개의 도시에서 관측되었기 때문에 $p=12$, n=30 이다. 연중 기상 변동값들은 모수 함수로 표현하기에는 너무 복잡한 경향선을 가지기 때문에 우리는 프로파일 $Y_i=(Y_{i1},\cdots,Y_{it},\cdots,Y_{ip})$가 이차 스플라인 회귀모형 (Rupper 등, 2003)을 따른다고 가정하였다. 군집 k의 이차 스플라인 회귀모형은 다음과 같다.

$$\begin{aligned} Y_i &= \beta_{0k}+\beta_{1k}t+\beta_{2k}t^2+\sum_{l=1}^{p-2} u_{lk}(t-\tau_l)_+^2 \\ &= X\beta_k + ZU_k + \epsilon_i \end{aligned}$$

여기에서 $X=(1,t,t^2),\ \beta_k=(\beta_{0k},\beta_{1k},\beta_{2k})^T,\ Z=((t-\tau_1)_+^2,\ \cdots,(t-\tau_l)_+^2,\ \cdots,(t-\tau_{p-2})_+^2)$, τ_l은 매듭, $m_+ = \max(0,m)$, $U_k = (u_{ik},\ \cdots,\ u_{p-2k})^T$, $\varepsilon_i \sim N(0,\sigma_k^2 I_p)$, 그리고 $U_k \sim N(0,$

$\lambda^2\sigma_k^2 I_{p-2}$)이다. λ는 스플라인 곡선의 부드러운 정도를 결정하는데, 이 값에 결과값들이 robust한 것으로 알려져 있다. 이렇게 해서 형성된 각 군집내의 우도함수를

$$\int f(Y_{C_k}, U_k|\beta_k, \sigma_k^2)dU_k$$

라고 하자. 여기에서

$$f(Y_{C_k}, U_k|\beta_k, \sigma_k^2) = \prod_{i\in C_k} \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{\{Y_i - (X\beta_k + ZU_k)\}^2}{2\sigma_k^2}} \times \prod_{j=1}^{p-2} \frac{1}{\sqrt{2\pi}\lambda\sigma_k} e^{-\frac{\{u_{jk}\}^2}{2\lambda\sigma_k^2}}$$

그러면, 식 (2.1)에 들어갈 주변 군집우도함수

$$f(Y_{C_k}|\beta_k, \sigma_k^2) = \int f(Y_{C_k}, U_k|\beta_k, \sigma_k^2)dU_k$$

가 형성 된다. 분산 ω의 사전 분포로는 Crowley (1997) 분포

$$\pi(w) \propto \zeta^{c(w)} \prod_{k=1}^{c(w)} (n-1)!$$

을 사용하였고, 군집의 특성을 나타내는 모수 $\beta = (\beta_1^T, \ldots, \beta_k^T, \ldots, \beta_{C(\omega)}^T)^T$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_k^2, \ldots, \sigma_{C(\omega)}^2)^T$에는 무정보 사전분포인

$$\pi(\beta, \sigma^2|w) \propto \prod_{k=1}^{c} (w)(1/\sigma_k^2)^{2/5}$$

을 사용하였다. 끝으로 분할 ω의 주변사후밀도함수

$$\pi(w|Y) \propto \int\int f(Y|\beta, \sigma^2, w)\pi(\beta, \sigma^2|w)\pi(w)d\beta d\sigma^2$$

가 계산된다. 베이지안 모델기반 군집분석 방법은 이 주변사후밀도함수를 최대화 시키는 분할 ω를 확률적 탐사법 (stochastic search)을 이용하여 찾는다. 확률적 탐사를 시작하기 전에, 초기 partition을 잡게 됩니다. 이러한 partition을 기준으로 전체 cluster들에 공통적으로 적합한 λ를 추정합니다. 이 추정치를 이용해서 확률적 탐사법을 진행합니다. 이러한 추정방법은 Joo 등 (2008)의 supplementary material에 제공되어 있습니다.

3. 분석결과

3.1. 강수량

그림 3.1의 첫 그래프에는 관측된 월별 강수량의 30개 프로파일 (도시)들이 그려져 있다. 이들을 대상으로 군집분석을 한 결과, 강수량에 따른 지역적 특성은 3개의 군집을 형성하는 것으로 나타났다. 그림 3.1의 둘째, 셋째, 넷째 그래프에는 각 군집에 속한 연중 강수량 변동 프로파일들이 얇은 선으로 표시되어 있고, 각 군집 내에서의 평균적 경향을 나타내는 최량선형비편향 추정선 (Best Linear Unbiased Estimator)은 굵은 선으로 표시되어 있다. 군집번호는 각 그래프의 상단에 표시되어 있고, 그 군집에

속하는 도시의 수가 괄호 속에 표시되어 있다. 마지막으로 각 군집 내 도시들의 지역적 분포는 그림 3.1의 둘째 줄에 있는 지도에 표시되어 있다.

동해안과 남해안 일대를 포함하는 군집 1과 3은 9월에 가장 높은 강수량을 가지고 있다. 이는 태풍에 큰 영향을 받은 것으로 추측된다. 군집 3에 속한 제주도와 전라도 지역들은 2007년 9월 13일 11호 태풍"나리"의 영향을 받아 태풍의 이동경로에 따라 9월에 집중적인 호우가 내렸던 곳들이다. 또한 군집 3은 다소 높은 겨울철 강수량 (강설량)을 가지기도 했다. 반면, 주로 경기 ·충북 내륙지역인 군집 2에는 장마철인 8월에 가장 높은 강수량을 보였다.

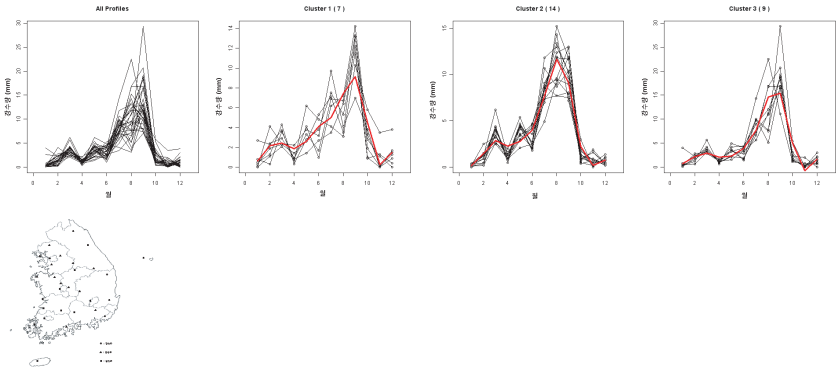

그림 3.1 강수량 변동의 군집분석 결과

3.2. 기온

기온에 따른 지역적 특성은 4개의 군집으로 나타났다. 결과는 그림 3.2에 요약되어 있다. 군집 1은 비슷한 경도상의 내륙지방들이며 군집 2는 군집 1의 내륙지역을 둘러싸는 것과 같은 형상의 지역들로써 서해안 지역을 포함한다. 군집 3과 4는 주로 동 ·남해안지역으로서 특히 군집 3은 울릉도, 제주도를 포함하는 지역들이다. 내륙지역인 군집 1은 겨울철에 상대적으로 낮은 온도를 보이고 있고, 동·남해안 지역인 군집 3과 4에서는 겨울철에 높은 온도를 보여주고 있다. 또한 군집 3과 4에서는 6월과 7월에 비해서 8월에 온도가 급하게 상승한 반면에, 서해안과 내륙지역인 군집 1과 2에서는 6월과 7월에 온도가 이미 많이 상승하기 때문에 8월에 상대적으로 완만한 상승을 하는 것으로 나타났다.

3.3. 풍속

풍속에 따른 지역적 특성은 12개의 군집으로 나타났다. 결과는 그림 3.3에 요약되어 있다. 이는 풍속이 국소적 지형특성에 영향을 많이 받아, 특이한 양상을 쉽게 보일 수 있기 때문이다. 이들 중 7개의 군집은 각각 하나의 지역만을 포함하고 있다. 내륙지역인 군집 1, 2, 3, 5, 7 그리고 8의 경우에는 3월 4월에 가장 강한 바람이 부는 것으로 나타났다. 제주도인 군집 10과 동해안지역인 군집 4 (울진), 9 (포항), 그리고 11 (강릉)인 경우에는 겨울에 가장 강한 바람이 부는 것으로 나타났다. 주로 남·서해 지역인 군집 6과 12의 경우에는 풍속이 연중 고르게 나타났다.

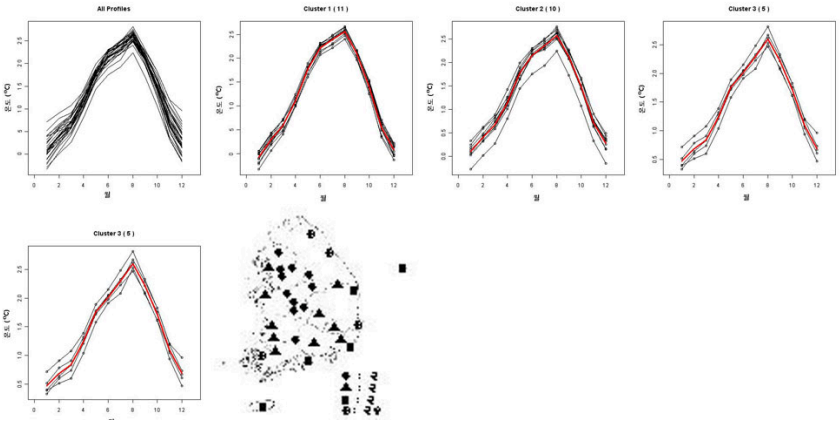

그림 **3.2** 기온 변동의 군집분석 결과

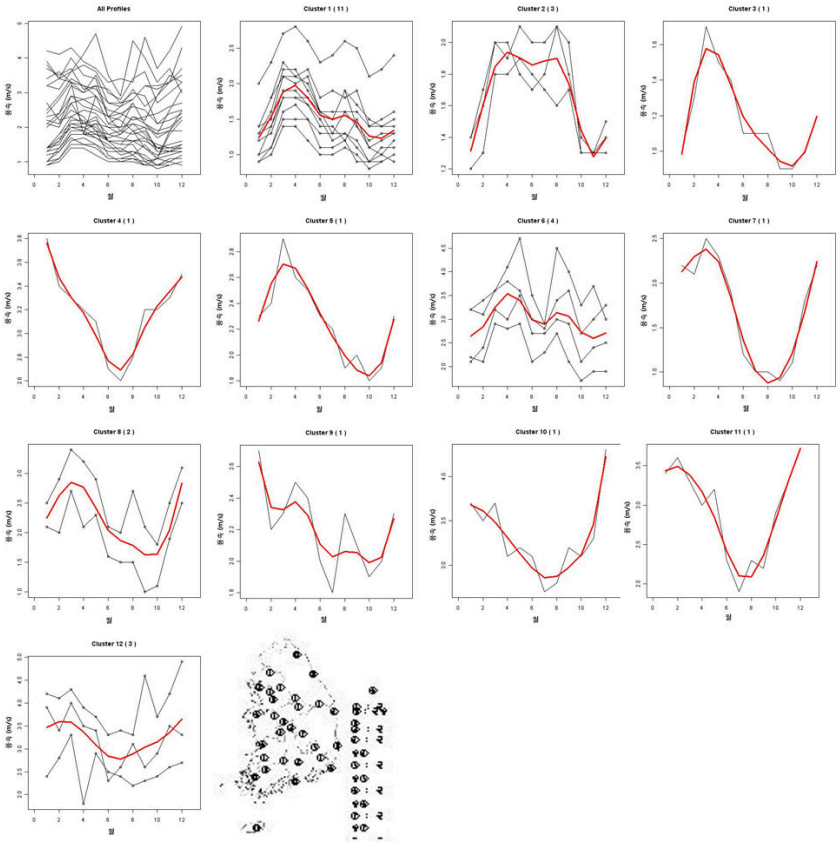

그림 **3.3** 풍속 변동의 군집분석 결과

3.4. 일조량

제주만이 상대적으로 다른 일조량 연중곡선을 가지고 나머지 도시들은 하나의 군집을 형성하는 것으로 나타났다. 모든 관측 위치들이 비산악 지역에 위치해 있어서 지형적 특성이 일조량에 큰 영향을 미치지 않았던 것으로 생각된다. 결과는 그림 3.4에 요약되어 있다. 제주도는 다른 지역에 비해서 겨울에는 상대적으로 적은 일조량을, 여름철인 7월 ∼ 9월에는 많은 일조량을 보이는 특이한 경향을 가진 것으로 나타났다.

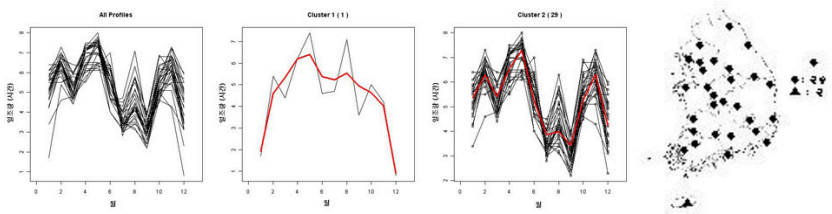

그림 3.4 일조량 변동의 군집분석 결과

3.5. 습도

습도에 따른 지역적 특성은 4개의 군집으로 나타났다. 결과는 그림 3.5에 요약되어 있다. 이 4개의 군집들은 경도선을 따라서 비교적 뚜렷하게 구별되는 경향을 보여주고 있다. 군집 4는 동해안 지역, 군집 3은 동해안과 인접한 내륙지역, 군집 1인 서해안 내륙지역, 군집 2는 서해안 지역과 울릉도를 포함한다. 해안지역과 내륙지역의 차이가 뚜렷하게 나타났다. 내륙지역인 군집 1의 경우 겨울철에 높은 습도를 보이는 반면, 동해안 지역인 군집 4는 겨울철에 낮은 습도를 보이고 있다. 또한 서해안 지역인 군집 2가 상대적으로 일정한 습도를 유지하는 반면 동해안 지역인 군집 4는 가장 큰 연중 습도차를 보여주고 있다.

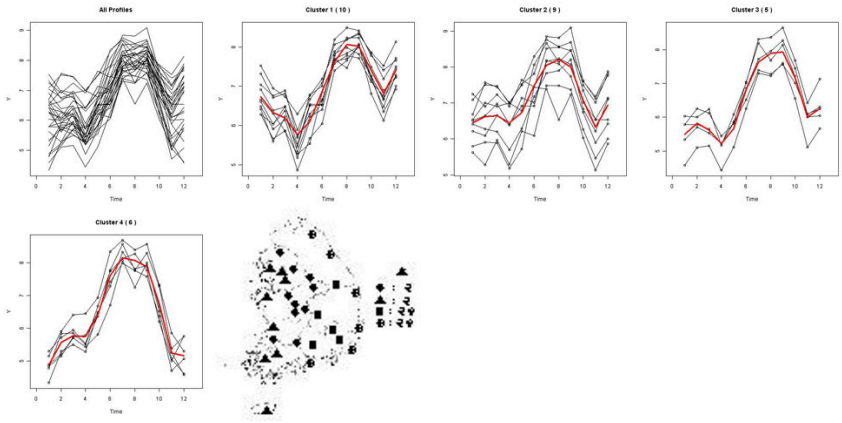

그림 3.5 습도 변동의 군집분석 결과

4. 결론

이 논문에서는 기상관측의 기본적인 요소인 강수량, 기온, 풍속, 일조량, 습도를 이용해서 지역 (도시)들 간의 기상적 유사성을 가지는 군집들을 찾아냈다. 위도별로 주로 비슷한 기상변동추이를 가질 것이라는 통념과는 달리, 기상변동추이는 경도와 위치 (내륙 혹은 해안지역)에 의해서도 많은 영향을 받는 것으로 나타났다. 이는 남한의 산맥들이 주로 남북으로 뻗어 있기 때문에 기인한 것으로 생각이 된다.

미래에는 특정 기상요소를 대상으로 단순히 통계분석만을 행하는 것이 아니라, 종합적인 안목으로 한반도가 어떠한 지역적 특성을 가지고 있는가에 대한 광범위한 연구를 기상학자와 같이 행하면 더욱 의미있는 결과가 나올 것으로 예상한다.

참고문헌

이용희, 최정희, 오재호 (2000). 대기조성변화에 따른 지역기상 변화의 통계적 예측. <한국데이터정보과학회지>, **2**, 333-344.

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.

Crowley, E. M. (1997). Product partition models for normal means. *Journal of the American Statistical Association*, **92**, 192-198.

Joo, Y., Booth, J., Namkoong, Y. and Casella, G. (2008). Model-Based Bayesian Cluster Analysis. *Bioinformatics*, **24**, 874-875.

Kim, B., Cho, C., Chung, H., Park, J., Shin, S. and Lee, Y. (2002). Impact of the Additional Observation Data on the Weather Analysis. *Proceedings of Joint Conference of Korean Data And Information Science Society and Korean Data Analysis Society 2002,* 1-4.

Sohn, K. T., Hong, C., Kwon, H. J. and Park, J. K. (2002). Prediction of the number of Tropical Cyclones over Western North Pacific in TC season. *Proceedings of Joint Conference of Korean Data And Information Science Society and Korean Data Analysis Society 2002,* 5-15.

Ruppert, D., Wand, M. P. and Caroll, R. J. (2003). *Semiparametric regression*, Cambridge University Press, New York, 2003.

Journal of the Korean Data & Information Science Society
2009, **20**(1), 57–64

한국데이터정보과학회지

Cluster analysis with Korean weather data: Application of model-based Bayesian clustering method

Yongsung Joo[1] · Hyungjoo Jung[2] · Byungjun Kim[2]

[12]Statistics Department, Dongguk Unoversity

Received 11 November 2008, revised 22 December 2008, accepted 31 December 2008

Abstract

In this paper, 30 main cities are clustered based on precipitation, temperature, wind speed, photo period, and humidity. We found that the resulting clusters has strong relationships with geographical locations. These results make sense because, although Korea is a small country, Korean weather is known to have strong locality. The largest number of clusters is found when wind speed is used as an interested variable for clustering and the smallest number of clusters is found when photo period is used. The large number of clusters based on wind speed indicates that wind speed is affected easily by local geography.

Keywords: Clustering, precipitation, profile, spline.

[1] Corresponding Author: Assistant Professor, Statistics Department, Dongguk Unoversity, Pildong 3-26, Joonggu, Seoul 100-715, Korea. E-mail: yongsungjoo@dongguk.edu

[2] Graduate student, Statistics Department, Dongguk Unoversity, Pildong 3-26, Joonggu, Seoul 100-715, Korea.