응용통계연구 제 21권 3호, 2008년, pp. 547-560

평활된 주기도를 이용한 강수량자료의 군집화

박만식[1] 김희영[2]

요 약

스펙트럼 밀도함수(spectral density function)는 시계열 자료가 정상성(stationarity)을 만족하는 경우에 주파수 영역(frequency domain)에서 시계열 자료의 자기공분산함수(auto-covariance function)을 결정짓는 함수이고, 평활된 주기도(smoothed periodogram)는 스펙트럼 밀도함수의 일치 추정량(consistent estimator)이 됨이 잘 알려져 있다. 본 연구에서는 시계열 자료를 평활된 주기도를 이용하여 군집화하는 방법을 소개한다. 최근 김희영과 박만식 (2007)의 연구에 의하면 이 거리는 정상시계열들을 효율적으로 분류하고 있음을 알 수 있다. 본 연구는 시계열 자료를 분류하는 데 사용된 기존의 거리들을 간략히 소개하고, 우리나라 22개 지역에서 1987년 1월부터 2007년 12월까지 측정한 월별 강수량 자료를 대상으로 평활된 주기도 거리를 이용하여 지역을 군집화한다.

주요용어: 주기도, 평활, 스텍트럼 밀도함수, 군집분석, 강수량.

1. 서론

시계열자료를 군집화하는 일은 경제, 경영, 인구통계, 지리학, 의학, 기상학, 생물학 등의 여러 분야에서 접할 수 있다. 생물학에서는 유전자들의 시간에 따라 발현 수준의 변화를 고려함으로써 발현패턴에 기초한 유전자들의 그룹을 찾아 내는 것이 주요한 연구이다 (Goldstein 등, 2002). 경제 분야에서는 패턴이 유사한 뮤추얼 펀드, 주식거래 종목들을 그룹화하는 일은 투자자, 금융관계자들에게 유용한 정보를 제공할 수 있으며 (Pattarin 등, 2004; Fu 등, 2001), 산업생산지수, 소비지수들에 따라 산업분야를 그룹화할 수 있다 (Piccolo, 1990; Caiado 등, 2006; Maharaj, 2000). 특히, 시계열 자료분석시 모형설정과 예측을 수행해야 할 시계열의 계열수가 상당히 많은 경우에, 주어진 시계열 각각을 예측하는 대신 패턴이 유사한 시계열들을 그룹화한 후에, 각 그룹으로부터 재표현된 것에 기초하여 예측을 수행한다면, 훨씬 효율적일 것이다. 기상학분야에서는 기후를 유형별로 분류하고, 그것에 의해 기후지역을 구분하고자 하는 연구 (Kalpakis 등, 2001; 고정웅 등, 2005)가 있고, 지진학분야에서는 Kakizawa 등 (1998)과 Shumway (2003)은 지진파동, 지뢰폭발파동을 이용하여 군집분석, 판별분석의 연구를 하였다. 보다 많은 적용분야에 관하여는 Liao (2005)의 논문을 참고하기 바란다. Liao (2005)는 기존의 시계열 군집화를 연구한 논문들을 광범위하게 조사하여, 이들 논문들에서 사

1) (136-701) 서울시 성북구 안암동 5가 126-1, 고려대학교 의과대학 의학통계학교실, 연구교수.
E-mail: bayesia@korea.ac.kr

2) (136-701) 교신저자. 서울시 성북구 안암동 5가 126-1, 고려대학교 의과대학 의학통계학교실, 연구교수.
E-mail: starkim@korea.ac.kr

용한 거리, 군집화 알고리즘, 군집결과를 평가하는 방법 그리고 적용한 자료 등을 일목 요연하게 표로 제시하고 있다.

군집분석에서 일반적인 개체들간의 거리로 유클리드 거리(Euclidean distance)를 이용할 수 있다. 그러나, Galeano와 Peña (2000)가 언급한바와 같이 유클리드 거리는 시계열들사이의 거리로는 바람직하지 않다. 왜냐하면 유클리드 거리는 시계열 자료의 특성 중 주요한 자기상관(autocorrelation)을 고려하지 않는다. 즉, 두 개의 시계열 $\mathbf{x} = (x_1, \ldots, x_i \ldots, x_j, \ldots x_n)$와 $\mathbf{y} = (y_1, \ldots, y_i \ldots, y_j, \ldots y_n)$에서의 유클리디드 거리와 $\mathbf{x}^* = (x_1, \ldots, x_j \ldots, x_i, \ldots, x_n)$와 $\mathbf{y}^* = (y_1, \ldots, y_j \ldots, y_i, \ldots, y_n)$의 유클리드 거리가 같기 때문이다.이처럼 시계열 자료들을 군집화하는 연구에서는 시간에 따라 관측된 자료의 속성을 충분히 반영하는 거리를 잘 정의하는 것이 중요한 일이다.

시계열 자료의 군집분석방법은 실제 분석할 자료와 응용분야에 따라서 개체들 간의 다양한 거리가 정의될 수 있다. Corduas와 Piccolo (2008)는 시계열자료 거리를 정의하는 방법을 크게 두 가지로 구분하고 있다. 첫째는, 시계열을 생성하는 확률 모형을 가정한 후(예를 들어, autoregressive and moving-average(ARMA), autoregressive-integrated moving-average(ARIMA), vector-ARMA) 이들 모형의 추론으로부터 거리를 정의하는 방법이다. 둘째는, 자기상관함수(autocorrelation function), 교차상관함수(cross-correlation function) 등 시계열 자료의 특별한 특징(feature)을 정의한 후, 이들로부터 거리를 정의하는 방법이다.

후자의 방법 중 Caiado 등 (2006)은 시계열의 주기도(periodogram)를 이용하여 몇 가지 거리들을 제안하였다. 일반적으로 시계열분석은 시간영역(time domain)과 주파수영역(frequency domain)에서의 분석으로 나누어진다. 시계열자료의 복잡한 연관성은 주파수 영역에서 거의 독립적(asymptotic independent)인 통계량인 주기도로 변환될 수 있다. 그리고, 주기도는 시계열이 정상성(stationary)을 따를 때 자기 공분산함수(auto-covariance function)를 고유하게 결정짓는 스펙트럼 밀도함수(spectral density function)의 추정량으로 이용된다. 그러나, 주기도는 스펙트럼 밀도함수의 점근적인 불편추정량(asymptotic unbiased estimator)은 되지만, 일치추정량이 되지는 못함이 잘 알려진 사실이다. 시계열자료분석에서 모수의 추정, 미래시점의 예측 등 거의 모든 추론은 자기 공분산함수를 이용한 식으로부터 유도된다. 따라서, 이론적으로 자기 공분산함수를 고유하게 결정짓는 스펙트럼 밀도함수의 좋은 추정량을 이용하여 두 시계열들 사이의 거리로 정의하는 것이 보다 바람직할 수 있다. 최근 김희영과 박만식 (2007)의 연구는 Caiado 등 (2006)의 연구의 단점을 보완하여 평활된 주기도(smoothed periodogram)를 시계열들 사이의 거리로 새로이 정의하고, 모의실험을 통하여 다양한 시나리오에서 그 성능을 Caiado 등 (2006)의 거리와 비교하였다. 그 결과 평활된 주기도는 Caiado 등 (2006)의 거리보다 정확히 정상시계열들을 autoregressive(AR)모형과 moving-average(MA)모형으로, 또한 ARMA모형과 ARIMA모형으로 군집화함을 보여주고 있다.

본 논문에서는 김희영과 박만식 (2007)의 거리를 이용하여 기상학분야 자료에 촛점을 맞추고자 한다. 본 논문의 구성은 다음과 같다. 2절에서는 시간영역에서의 거리, 주파수 영역에서의 거리들을 간단히 소개한다. 3절에서는 실제 자료로 서울을 비롯한 22개 지역의 1987년 1월부터 2007년 12월까지의 월별 강수량자료를 이용하여 지역을 군집화한 결과에 대해 살펴

본다. 4절에서는 결론과 시계열 자료의 군집화에 대한 추후 연구과제에 대해서 언급한다.

2. 거리 측도들(Distance Measures)

관측된 두 개의 시계열 $\mathbf{x} = \{x_1, \ldots, x_n\}$, $\mathbf{y} = \{y_1, \ldots, y_n\}$ 사이의 거리를 정의하는 몇 가지 방법들을 소개한다.

2.1. 시간영역에서의 거리(Distances on the Time Domain)

2.1.1. 자기상관함수 거리(autocorrelation function distance)

$\mathbf{x}$와 $\mathbf{y}$의 추정된 자기 상관함수 $\hat{\boldsymbol{\rho}}^{\mathbf{x}} = (\hat{\rho}_1^{\mathbf{x}}, \ldots, \hat{\rho}_L^{\mathbf{x}})'$, $\hat{\boldsymbol{\rho}}^{\mathbf{y}} = (\hat{\rho}_1^{\mathbf{y}}, \ldots, \hat{\rho}_L^{\mathbf{y}})'$를 이용하여 다음과 같은 3가지 거리들을 고려할 수 있다. 첫째는, 자기상관함수들 사이의 유클리드 거리로써

$$d_{ACF}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{L} (\hat{\rho}_i^{\mathbf{x}} - \hat{\rho}_i^{\mathbf{y}})^2}. \tag{2.1}$$

따라서, 선택된 모든 차수의 자기 상관함수에 대하여 균일한 가중치를 두는 방법이다. 둘째는, d_{ACF}와는 달리 차수에 따라 상이한 가중치를 부여하는 방법으로,

$$d_{ACFG}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{L} m_i \, (\hat{\rho}_i^{\mathbf{x}} - \hat{\rho}_i^{\mathbf{y}})^2}. \tag{2.2}$$

Caiado 등 (2006)은 $m_i = p(1-p)^i$, $i = 1, \ldots, L$, $0 < p < 1$로 선택하여 차수가 커짐에 따라 작은 가중치를 부여하였다. 세번째는, 자기 상관함수들의 마할라노비스(Mahalanobis)거리로써

$$d_{ACFM}(\mathbf{x}, \mathbf{y}) = \sqrt{(\hat{\boldsymbol{\rho}}^{\mathbf{x}} - \hat{\boldsymbol{\rho}}^{\mathbf{y}})' \, \boldsymbol{\Omega}^{-1} \, (\hat{\boldsymbol{\rho}}^{\mathbf{x}} - \hat{\boldsymbol{\rho}}^{\mathbf{y}})}, \tag{2.3}$$

여기서 $\boldsymbol{\Omega}$는 $\hat{\boldsymbol{\rho}}^{\mathbf{x}} - \hat{\boldsymbol{\rho}}^{\mathbf{y}} = (\hat{\rho}_1^{\mathbf{x}} - \hat{\rho}_1^{\mathbf{y}}, \ldots, \hat{\rho}_L^{\mathbf{x}} - \hat{\rho}_L^{\mathbf{y}})'$의 분산 공분산 행렬(variance-covariance matrix)이다. 따라서, $\boldsymbol{\Omega}$의 (i, j) 원소는

$$\mathrm{Cov}\left(\hat{\rho}_i^{\mathbf{x}} - \hat{\rho}_i^{\mathbf{y}}, \, \hat{\rho}_j^{\mathbf{x}} - \hat{\rho}_j^{\mathbf{y}}\right) = \mathrm{Cov}\left(\hat{\rho}_i^{\mathbf{x}}, \hat{\rho}_j^{\mathbf{x}}\right) + \mathrm{Cov}\left(\hat{\rho}_i^{\mathbf{y}}, \hat{\rho}_j^{\mathbf{y}}\right) - 2\mathrm{Cov}\left(\hat{\rho}_i^{\mathbf{x}}, \hat{\rho}_j^{\mathbf{y}}\right) \tag{2.4}$$

이다. 지금까지 거의 대부분의 연구들은 두 시계열 $\mathbf{x}$와 $\mathbf{y}$의 연관성을 고려하지 않고 논의를 전개하였다. 따라서, 식 (2.4)에서 세 번째 항은 0으로 간주하고, 첫 번째 항과 두 번째 항은 Bartlett (1946)의 근사를 이용하면 다음과 같다.

$$\begin{aligned}\mathrm{Cov}\left(\hat{\rho}_i^{\mathbf{x}} - \hat{\rho}_i^{\mathbf{y}}, \, \hat{\rho}_j^{\mathbf{x}} - \hat{\rho}_j^{\mathbf{y}}\right) = &\sum_{k=1}^{L} \left[\left\{\hat{\rho}_{(k+i)}^{\mathbf{x}} \hat{\rho}_{(k-i)}^{\mathbf{x}} - 2\hat{\rho}_{(i)}^{\mathbf{x}} \hat{\rho}_{(k)}^{\mathbf{x}}\right\}\left\{\hat{\rho}_{(k+j)}^{\mathbf{x}} \hat{\rho}_{(k-j)}^{\mathbf{x}} - 2\hat{\rho}_{(j)}^{\mathbf{x}} \hat{\rho}_{(k)}^{\mathbf{x}}\right\}\right] \\ &+ \sum_{k=1}^{L} \left[\left\{\hat{\rho}_{(k+i)}^{\mathbf{y}} \hat{\rho}_{(k-i)}^{\mathbf{y}} - 2\hat{\rho}_{(i)}^{\mathbf{y}} \hat{\rho}_{(k)}^{\mathbf{y}}\right\}\left\{\hat{\rho}_{(k+j)}^{\mathbf{y}} \hat{\rho}_{(k-j)}^{\mathbf{y}} - 2\hat{\rho}_{(j)}^{\mathbf{y}} \hat{\rho}_{(k)}^{\mathbf{y}}\right\}\right].\end{aligned}$$

2.1.2. 편자기 상관함수 거리(partial autocorrelation function distance)

x와 y의 추정된 편자기 상관함수 $\hat{\boldsymbol{\phi}}^{\mathbf{x}} = \left(\hat{\phi}_1^{\mathbf{x}}, \ldots, \hat{\phi}_L^{\mathbf{x}}\right)'$, $\hat{\boldsymbol{\phi}}^{\mathbf{y}} = \left(\hat{\phi}_1^{\mathbf{y}}, \ldots, \hat{\phi}_L^{\mathbf{y}}\right)'$를 이용하여 다음과 같은 편자기 상관함수들 간의 유클리드 거리를 고려할 수 있다.

$$d_{PACF}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{L} \left(\hat{\phi}_i^{\mathbf{x}} - \hat{\phi}_i^{\mathbf{y}}\right)^2}. \tag{2.5}$$

2.2. 주파수 영역에서의 거리(Distances on the Frequency Domain)

시계열 $\{X_t : t \in Z\}$은 실수값 만을 취하는 정상시계열(stationary time-series)로서, 평균 $E(X_t) = \mu$, 자기 공분산함수 $\gamma(h) = E(X_{t+h}X_t) - E(X_{t+h})E(X_t)$을 갖고, $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$를 만족한다고 가정하자. 그러면, 다음과 같은 스펙트럼 밀도함수가 정의된다.

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) \exp(-i\lambda h), \quad \lambda \in [-\pi, \pi]. \tag{2.6}$$

스펙트럼 밀도함수의 특징을 살펴보면 다음과 같다.

(1) $f(\lambda)$는 항상 0 이상의 값을 취하게 되고

(2) $f(\lambda)$는 우함수, 즉 $f(\lambda) = f(-\lambda)$가 되며

(3) 자기 공분산함수 $\gamma(h)$와 스펙트럼 밀도함수 $f(\lambda)$는 아래의 관계를 가진다.

$$\gamma(h) = 2\int_0^{\pi} f(\nu)\cos(\nu h)\, d\nu, \qquad f(\lambda) = \frac{1}{2\pi}\left[\gamma(0) + 2\sum_{h=1}^{\infty}\gamma(h)\cos(\lambda h)\right].$$

따라서, 주파수 영역에서 스펙트럼 밀도함수는 자기공분산함수를 고유하게 결정짓는 함수이다.

관측된 시계열 자료 $\{x_1, \ldots, x_n\}$를 이용하여 스펙트럼 밀도함수를 추정하는 방법은, 첫째로, 자료에 특별한 확률모형을 가정한 후, 추정치들로부터 스펙트럼 밀도함수를 추정하는 방법과 둘째로, 다음과 같이 정의되는 주기도(periodogram)를 이용하는 방법이다.

$$I_n(\omega_j) = \frac{1}{n}\left|\sum_{t=1}^{n} X_t \exp(-it\omega_j)\right|^2. \tag{2.7}$$

여기서, $\mathbb{F}_n = \{j \in \mathbb{Z} | -\pi < \omega_j \equiv (2\pi j)/n \leq \pi\} = \{-\lfloor (n-1)/2 \rfloor, \ldots, \lfloor n/2 \rfloor\}$이고, $\lfloor x \rfloor$는 x을 넘지않는 최대 정수이다. 주기도의 특징은 다음과 같다.

(1) $I_n(\omega_j) = I_n(-\omega_j)$이고,

(2) 식 (2.5)에서 스펙트럼 밀도함수가 자기공분산함수 $\hat{\gamma}$를 이용한 식으로 표현되는 것과 유사하게, 주기도는 표본 자기공분산함수를 이용하여 다음과 같이 표현된다.

$$I_n(\omega_j) = \begin{cases} n\left|\frac{1}{n}\sum_{t=1}^{n} X_t\right|^2, & \omega_j = 0\text{이면,} \\ \sum_{|k|<n} \hat{\gamma}(k)\exp\{-ik\omega_j\}, & \text{그 이외.} \end{cases} \tag{2.8}$$

2.2.1. 주기도에 의한 거리(periodogram-based distance)

Caiado 등 (2006)은 두 시계열 $\mathbf{x} = \{x_1, \ldots, x_n\}$, $\mathbf{y} = \{y_1, \ldots, y_n\}$의 주기도

$$I_n^{\mathbf{x}}(\omega_j) = \frac{1}{n}\left|\sum_{t=1}^{n} x_t \exp(-it\omega_j)\right|^2, \qquad I_n^{\mathbf{y}}(\omega_j) = \frac{1}{n}\left|\sum_{t=1}^{n} y_t \exp(-it\omega_j)\right|^2$$

를 이용하여 $\mathbf{x}$와 $\mathbf{y}$의 거리를 다음과 같이 정의하였다.

$$d_I(\mathbf{x},\mathbf{y}) = \left[\sum_{j=1}^{\lfloor n/2\rfloor}\left\{I_n^{\mathbf{x}}(\omega_j) - I_n^{\mathbf{y}}(\omega_j)\right\}^2\right]^{\frac{1}{2}},$$

$$d_{NI}(\mathbf{x},\mathbf{y}) = \left[\sum_{j=1}^{\lfloor n/2\rfloor}\left\{NI_n^{\mathbf{x}}(\omega_j) - NI_n^{\mathbf{y}}(\omega_j)\right\}^2\right]^{\frac{1}{2}},$$

$$d_{LNI}(\mathbf{x},\mathbf{y}) = \left[\sum_{j=1}^{\lfloor n/2\rfloor}\left\{\ln NI_n^{\mathbf{x}}(\omega_j) - \ln NI_n^{\mathbf{y}}(\omega_j)\right\}^2\right]^{\frac{1}{2}},$$

여기서, $NI_n(\omega_j)$는 주기도를 표본분산(sample variance)으로 나눈 것으로 $NI_n(\omega_j) = I_n(\omega_j)/\hat{\gamma}(0)$이다. 위의 세가지 거리들 중 d_{NI}와 시간영역에서의 거리들 중 d_{ACF}는 서로 같음을 아래에 식에 의해 알 수 있다.

$$d_{NI}(\mathbf{x},\mathbf{y}) = 2\sqrt{n}\sqrt{\sum_{k=1}^{n-1}(\hat{\rho}_i^{\mathbf{x}} - \hat{\rho}_i^{\mathbf{y}})^2} = 2\sqrt{n}d_{ACF}(\mathbf{x},\mathbf{y}).$$

2.2.2. 평활된 주기도에 의한 거리(smoothed periodogram-based distance)

Caiado 등 (2006)이 제시한 거리는 $(1/2\pi)I_n(\omega_j)$가 $f(\omega_j)$의 일치추정량이 되지 못한다는 것을 염두에 두지 않았다. 따라서, 일반적으로 일치추정량을 만들기 위하여 다음과 같이 평활을 이용하여 스펙트럼 밀도함수의 추정량으로 사용한다.

$$\hat{f}(\omega_j) = \frac{1}{2\pi}\sum_{|s|\le m} W_n(s) I(\omega_{j+s}). \tag{2.9}$$

여기서 $W_n(\cdot)$는 윈도우(window), $2m+1$은 윈도우의 폭이라고 부른다. m과 $W_n(\cdot)$이 만족할 조건과 통계적 성질은 Brockwell과 Davis (1991)을 참고하기 바란다. 최근 김희영과 박만식 (2007)은 평활된 주기도를 이용하여 다음과 같은 거리를 새로이 제안하였다.

$$d_I^*(\mathbf{x},\mathbf{y}) = \left[\sum_{j=1}^{\lfloor n/2\rfloor}\left\{\widetilde{I}_n^{\mathbf{x}}(\omega_j) - \widetilde{I}_n^{\mathbf{y}}(\omega_j)\right\}^2\right]^{\frac{1}{2}},$$

$$d_{NI}^*(\mathbf{x},\mathbf{y}) = \left[\sum_{j=1}^{\lfloor n/2\rfloor}\left\{N\widetilde{I}_n^{\mathbf{x}}(\omega_j) - N\widetilde{I}_n^{\mathbf{y}}(\omega_j)\right\}^2\right]^{\frac{1}{2}},$$

$$d_{LNI}^*(\mathbf{x},\mathbf{y}) = \left[\sum_{j=1}^{\lfloor n/2\rfloor}\left\{\ln N\widetilde{I}_n^{\mathbf{x}}(\omega_j) - \ln N\widetilde{I}_n^{\mathbf{y}}(\omega_j)\right\}^2\right]^{\frac{1}{2}},$$

여기서 $\widetilde{I}_n(\omega_j)$와 $N\widetilde{I}_n(\omega_j)$는

$$\widetilde{I}_n(\omega_j) = \sum_{|s|\le m} W_n(s) I_n(\omega_{j+s}), \qquad N\widetilde{I}_n(\omega_j) = \frac{\widetilde{I}_n(\omega_j)}{\widehat{\gamma}(0)}$$

이다. 서론에서 언급한 바와 같이, 김희영과 박만식 (2007)은 다양한 모의실험의 결과를 통해 평활된 주기도에 의한 거리가 주기도에 의한 거리와 비교하여 보다 정확히 정상시계열들을 AR모형과 MA모형으로 그리고 정상시계열(ARMA)모형과 비정상시계열(ARIMA)모형으로 군집화함을 보였다. 김희영과 박만식 (2007)은 다음 식 (2.10)의 modified Daniell 윈도우를, 윈도우의 폭 $2m+1$은 3을 사용하였다.

$$W_n(s) = \begin{cases} (2m(\mathbf{1}[|s|=m]+1))^{-1}, & |s|\le m\text{이면,} \\ 0, & \text{그 이외.} \end{cases} \tag{2.10}$$

3. 실제 응용(Real Application)

이 절에서는 1987년 1월부터 2007년 12월까지의 252개월 동안의 우리나라 22개 지역에 대한 월별 강수량(mm)자료를 이용하여 (편)자기상관함수를 이용한 거리, 주기도를 이용한 거리, 평활된 주기도를 이용한 거리로 군집분석을 한다. 자료는 기상청 홈페이지(http://www.kma.go.kr)로부터 구한 자료로써, 일별 강수량을 합한 월별 강수량자료이고, 그림 3.1은 관측된 22개 지점을 나타낸다(서울, 인천, 수원, 춘천, 속초, 강릉, 원주, 충주, 청주, 대전, 전주, 남원, 광주, 목포, 구미, 대구, 포항, 울산, 부산, 거제, 남해, 여수).

자료분석에 앞서, 기상학분야에서 강수량을 이용한 기존의 군집분석 연구를 살펴보자. 기상학분야에서는 지역에 따라 다양하게 나타나는 기후를 유형별로 분류하고, 그것에 의하여 기후지역을 구분하는 연구가 필요하며, 특히 강수는 다른 기후 요소에 비하여 국지성이 강하고, 지역성을 잘 반영하기 때문에 강수량를 이용한 기후구분 연구가 활발히 이루어졌다. 군집분석을 이용하여 기후지역을 구분한 연구로는 문영수 (1990), 이동규와 박정균 (1999) 등의 논문

그림 3.1: 우리나라의 22개 월별 강수량 측정지역

이 있다. 문영수 (1990)는 1976년에서 1985년까지 연간 36개의 순별(10일단위) 평균강수량을 변수로 사용하였고, 이동규와 박정균 (1999)는 1973년에서 1995년까지 6·8월의 3개월에 걸쳐 각 관측지점에 대해 총강수량, 순별(10일단위) 강수량, 일강수강도, 강수일 비율, 평균 강수지속일, 평균 무강수 지속일을 구하고, 다시 전체 관측기간에 대해 평균하여 군집분석의 변수로 사용하였다. 또한, 전통적으로 인자분석을 먼저 한 후에 강수특성을 결정하는 주요 인자를 추출하고, 추출된 인자들을 변수로 하여 군집분석을 한 연구들도 다수 있다 (고정웅 등, 2005; 이승호, 1993; 김성렬과 양진석, 1995). 위에 열거한 연구들은 전형적인 다변량 자료를 군집분석하는 연구라고 할 수 있으며, 본 논문에서 제안한 방법과는 확연히 다름을 알 수 있다.

분석에 고려된 22개 지역의 252개월 동안의 시계열자료의 시도표를 살펴보면 지역에 관계없이 6 9월에는 많은 강수량을, 1, 2월에는 적은 강수량을 나타내는 전형적인 계절 변동을 보이고 있다. 먼저 자료를 시간영역의 관점에서 분석해 보자. 계절성을 제거하기 위한 12차 차분(differentiation) 후의 표본 자기 상관함수($\hat{\rho}$)를 서울과 인천, 목포와 광주의 4개지역에 대해서만 시차 24까지 표 3.1에 제시하였다.

표 3.1을 자세히 살펴보면, 지역적으로 가까운 서울과 인천은 시차 10, 12, 23에서 모두 유의하며 동일부호를 갖는다. 인천은 시차 22에서 유의한 음의 값을 가지며, 서울 또한 유의하지는 않으나 시차 22에서 음의 값을 가진다. 그리고 목포와 광주는 시차 1에서는 유의한 양의 값을, 시차 12에서는 유의한 음의 값을 가진다.

본 논문에서 전체 22개 지역의 자기 상관함수를 모두 기술할 수는 없으나 몇 가지 특징을 살펴보면 다음과 같다. 먼저, 모든 지역이 시차 12에서 상당히 유의한 음의 상관관계를 보이고 있고, 지역에 따라 조금씩 다르긴 하지만 시차 48까지 중에서 유의한 시차들은 1, 2, 9, 10,

표 3.1: 서울, 인천, 목포, 광주의 표본 자기상관함수

지역	자기상관함수							
	$\widehat{\rho}_1$	$\widehat{\rho}_2$	$\widehat{\rho}_3$	$\widehat{\rho}_4$	$\widehat{\rho}_5$	$\widehat{\rho}_6$	$\widehat{\rho}_7$	$\widehat{\rho}_8$
	$\widehat{\rho}_9$	$\widehat{\rho}_{10}$	$\widehat{\rho}_{11}$	$\widehat{\rho}_{12}$	$\widehat{\rho}_{13}$	$\widehat{\rho}_{14}$	$\widehat{\rho}_{15}$	$\widehat{\rho}_{16}$
	$\widehat{\rho}_{17}$	$\widehat{\rho}_{18}$	$\widehat{\rho}_{19}$	$\widehat{\rho}_{20}$	$\widehat{\rho}_{21}$	$\widehat{\rho}_{22}$	$\widehat{\rho}_{23}$	$\widehat{\rho}_{24}$
	0.0150	−.0676	−.0045	0.0214	0.0201	0.0040	0.0106	0.0077
서울	−.0227	*0.1470	0.0484	*−.4906	−.0147	−.0677	0.0558	−.0261
	−.0168	−.0091	−.0358	−.0288	0.0203	−.1343	*−.1995	−.0851
	0.0103	−.1149	0.0124	−.0128	0.0282	−.0059	−.0049	0.0001
인천	−.0444	*0.2120	0.0536	*−.5013	−.0188	−.0565	0.0579	0.0062
	−.0069	−.0009	−.0343	0.0023	0.0197	*−.2016	*−.0189	0.0004
	*0.1714	−.0235	0.0335	−.0406	−.0092	0.0370	0.0471	0.0149
광주	−.0629	0.0157	−.1035	*−.5445	−.0217	0.0343	−.0014	0.0776
	0.0209	−.0452	−.0664	−.0232	0.0552	−.0217	0.0539	0.1598
	*0.2221	−.0726	0.0152	−.0169	−.0089	0.0273	−.0286	−.0340
목포	−.0463	0.0171	−.1132	*−.3919	−.0683	0.0764	0.0171	0.1061
	0.1000	−.0178	−.0547	0.0084	0.0387	−.0003	0.0713	−.0186

Notes: significant value under 5% significance level is with $*$.

11, 12, 13, 14, 21, 22, 23, 24, 35, 36, 47, 48이다. 특히, 시차 10에서 유의한 지역들은 강릉, 서울, 속초, 수원, 원주, 인천, 청주, 춘천, 충주이고, 모두 양의 값을 갖는다. 이들 지역은 그림 3.1을 참고로 하면, 청주보다 지리적으로 북쪽에 위치한 지역이다. 또한, 시차 1에서 유의한 지역은 광주, 목포, 울산, 부산이고, 양의 값을 갖는다. 남해와 거제는 시차 1에서 유의수준 10%에서 유의한 양의 값을 갖는다. 이들 지역은 지리적으로 남쪽에 위치하고 있다.

다음으로 주파수영역에서 살펴보자. 그림 3.2은 서울, 인천, 광주, 목포의 12차 차분한 자료에 대한 주기도(periodogram)를 보여준다. 252개월 자료를 12차 차분후 분석에 사용하였으므로, 식 (2.7)에서 $n = 240$이고, 빈도는

$$\omega_j \equiv \frac{2\pi j}{n} \in \left\{-\omega_{\left[\frac{(n-1)}{2}\right]} = -\omega_{119}, \ldots, \omega_{120} = \omega_{\left[\frac{(n-1)}{2}\right]}\right\}$$

이다. 주기도는 우함수이므로 $\{\omega_1, \ldots, \omega_{120}\}$에 대해서 나타내었다. 그리고 빈도 $\omega_j \equiv 2\pi j/n$에 대응하는 주기(period)는 $2\pi/\omega_j = n/j$이다. 그림 3.2를 통해 알 수 있듯이, 서울과 인천의 주기도 중에서 큰 순서대로 5개를 나열하면, 서울은 $I(\omega_{48})$, $I(\omega_{68})$, $I(\omega_7)$, $I(\omega_{28})$, $I(\omega_{91})$이고, 인천은 $I(\omega_{48})$, $I(\omega_{92})$, $I(\omega_{28})$, $I(\omega_{68})$, $I(\omega_{72})$이다. 두지역 모두 첫번째 봉우리를 제외하고는 나머지 4개의 봉우리의 크기가 비슷하고, 주기도의 전반적인 함수 형태는 서울과 인천이 유사함을 알 수 있다. 마찬가지로 광주와 목포에서 큰 순서대로 5개의 주기도는, 광주는 $I(\omega_{31})$, $I(\omega_{90})$, $I(\omega_{11})$, $I(\omega_9)$, $I(\omega_{33})$이고, 목포는 $I(\omega_{31})$, $I(\omega_{14})$, $I(\omega_{12})$, $I(\omega_{32})$, $I(\omega_{50})$이다. 광주와 목포 역시 첫번째 봉우리가 가장 크며, 나머지 4개의 봉우리의 크기는 유사함을 나타낸다. 주기도 $I(\omega_j)$가 나타내는 주기가 유의수준 5%에서 유의한지 조사하면, 서울은 $I(\omega_7)$, $I(\omega_{15})$, $I(\omega_{27})$, $I(\omega_{28})$, $I(\omega_{33})$, $I(\omega_{48})$, $I(\omega_{68})$, $I(\omega_{72})$, $I(\omega_{91})$, $I(\omega_{92})$, $I(\omega_{111})$이고, 인천은 $I(\omega_{15})$, $I(\omega_{28})$, $I(\omega_{48})$, $I(\omega_{68})$, $I(\omega_{72})$, $I(\omega_{92})$, $I(\omega_{112})$으로써, 유의한 주기도들이 거의 일치함을 나타낸다. 또한, 광주는 $I(\omega_9)$, $I(\omega_{11})$, $I(\omega_{31})$, $I(\omega_{32})$, $I(\omega_{33})$, $I(\omega_{51})$, $I(\omega_{79})$, $I(\omega_{90})$,

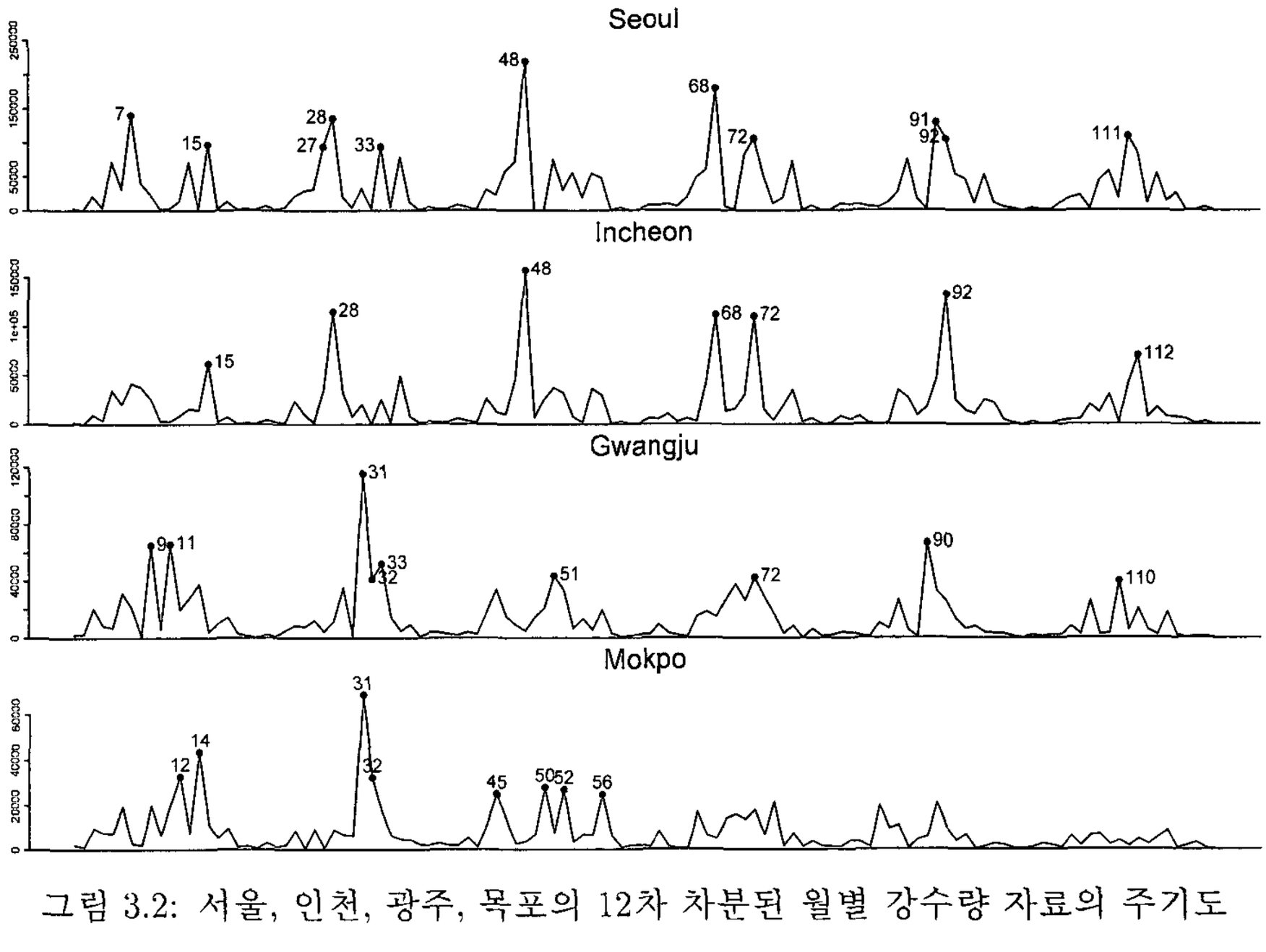


그림 3.2: 서울, 인천, 광주, 목포의 12차 차분된 월별 강수량 자료의 주기도

$I(\omega_{110})$이고, 목포는 $I(\omega_{12})$, $I(\omega_{31})$, $I(\omega_{32})$, $I(\omega_{45})$, $I(\omega_{52})$, $I(\omega_{56})$이다. 각 지역별 가장 큰 값을 가지는 주기도를 살펴보면 다음과 같다. 서울, 인천, 수원, 충주는 $I(\omega_{48})$, 속초, 강릉은 $I(\omega_{49})$, 춘천, 청주, 대전은 $I(\omega_{68})$, 원주, 구미, 포항은 $I(\omega_{71})$, 광주, 목포, 남원, 남해, 여수, 울산은 $I(\omega_{31})$, 부산은 $I(\omega_{30})$, 전주는 $I(\omega_{28})$, 대구는 $I(\omega_{107})$, 거제는 $I(\omega_{50})$이다. 대구와 거제는 첫번째 크기의 주기도가 남쪽에 위치한 지역들과 차이가 있으나, 두번째로 큰 주기도는 각각 $I(\omega_{31})$, $I(\omega_{30})$이다. 따라서, 광주, 목포, 남원, 남해, 여수, 거제, 부산, 울산, 포항, 대구는 12차 차분 후의 월별 강수량이 대략 8년$(= 240/31 = 7.74)$의 주기를 가지는 것으로 판단할 수 있다. 그리고 서울, 인천, 수원, 충주는 월별 강수량의 12차 차분한 후에 5년$(= 240/48)$의 주기를 보인다고 할 수 있다.

이제 김희영과 박만식 (2007)에서 제안된 평활된 주기도를 기반으로 한 거리와 (편)자기상관함수거리를 이용하여 군집분석한 결과들에 대해 살펴보자. 먼저, 전체 지역에서의 계절 조정된 자료를 이용한 평활된 주기도 $N\widetilde{I}_n(\cdot)$는 그림 3.3에 나와 있다. 그림 3.4은 2.1절에서 소개한 시간영역에서의 거리들 중 $d_{ACF}(\cdot)$, $d_{PACF}(\cdot)$ 그리고 2.2절에서 소개한 주파수 영역에서의 거리들 중 $d_{NI}(\cdot)$, $d^*_{NI}(\cdot)$, $d_{LNI}(\cdot)$, $d^*_{LNI}(\cdot)$을 지역들 사이의 거리로 이용하여, 완전연결법(complete-linkage)에 의한 군집화 결과를 덴드로그램(dendrogram)으로 나타낸 것이다. 여기서 식 (2.1)과 (2.5)에서의 $L = 60$은 계절차분 후의 자료수를 4로 나눈 값을 이용하였다 (Brockwell과 Davis, 1991, p.221). 각각의 거리에서 적절한 군집의 갯수를 정하는 기준은 몇 가지 통계를 참고로 할 수 있다. 하지만 본 연구에서는 관측지역의 수가 많지 않아서 22개 지역을 대략 3개의 그룹으로 군집화하는 것을 고려하였다. 그림 3.4 (c)는 $d_{NI}(\cdot)$를 이용한 경

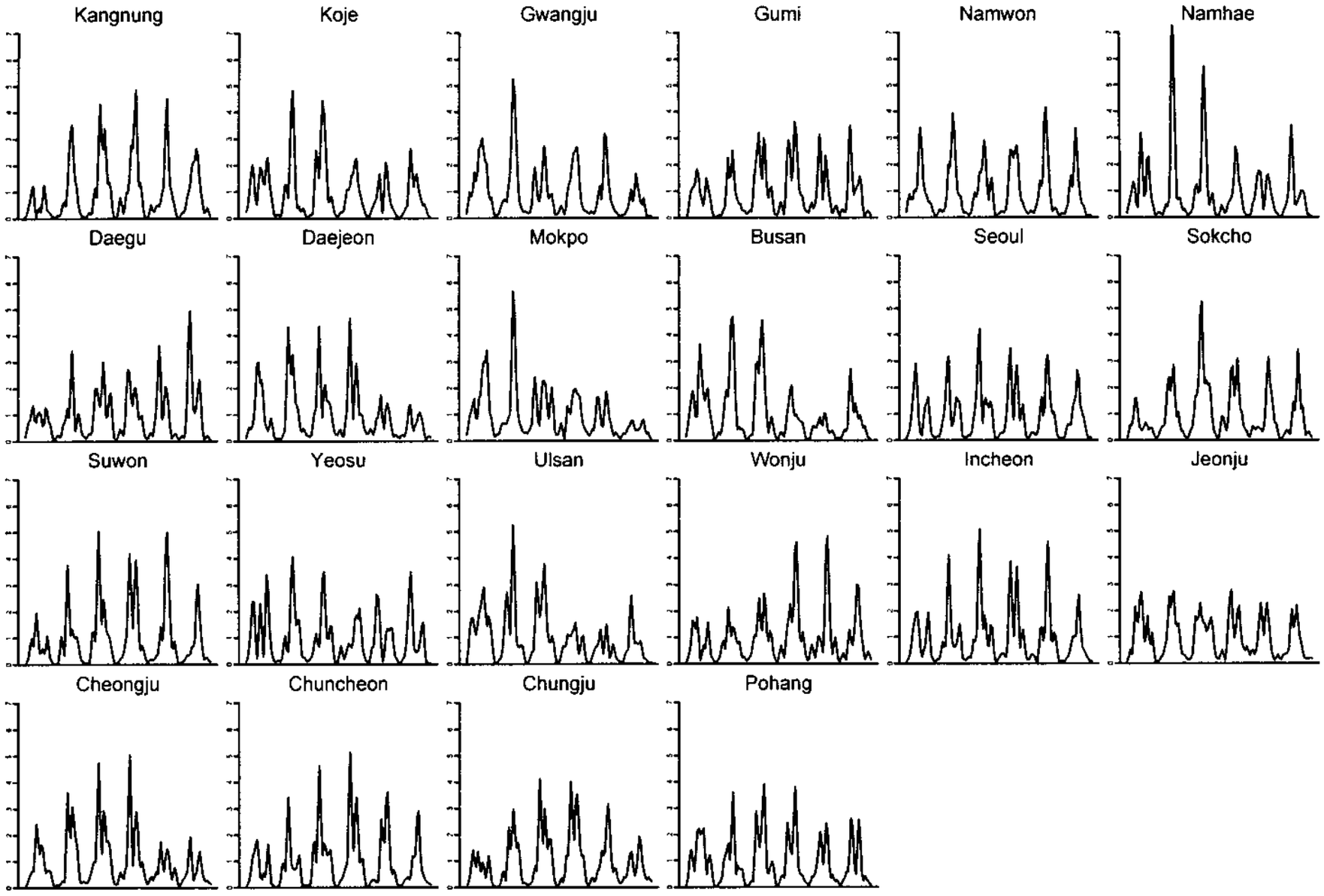
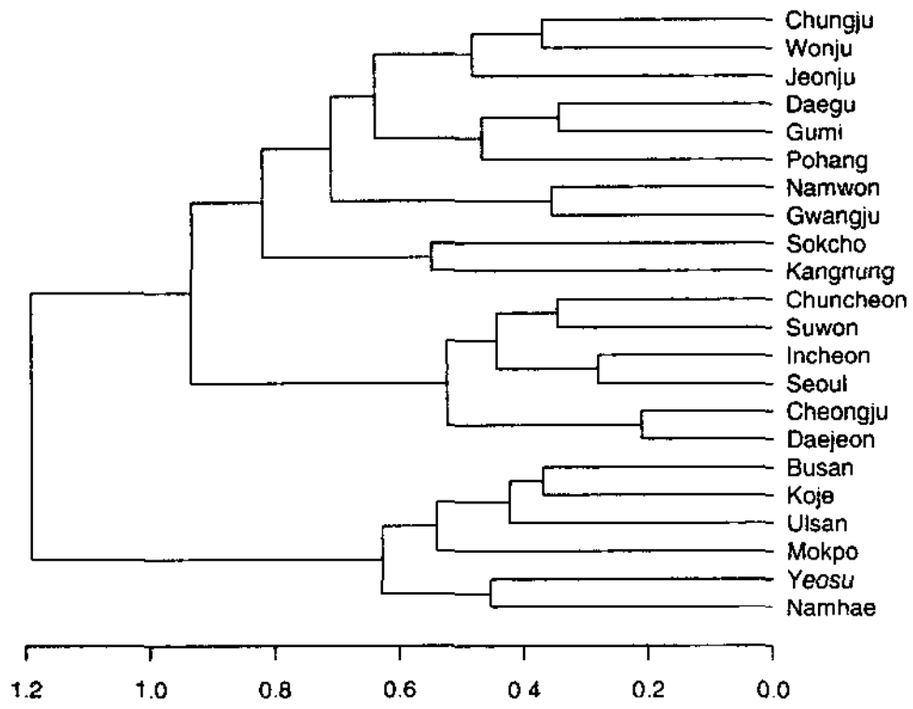


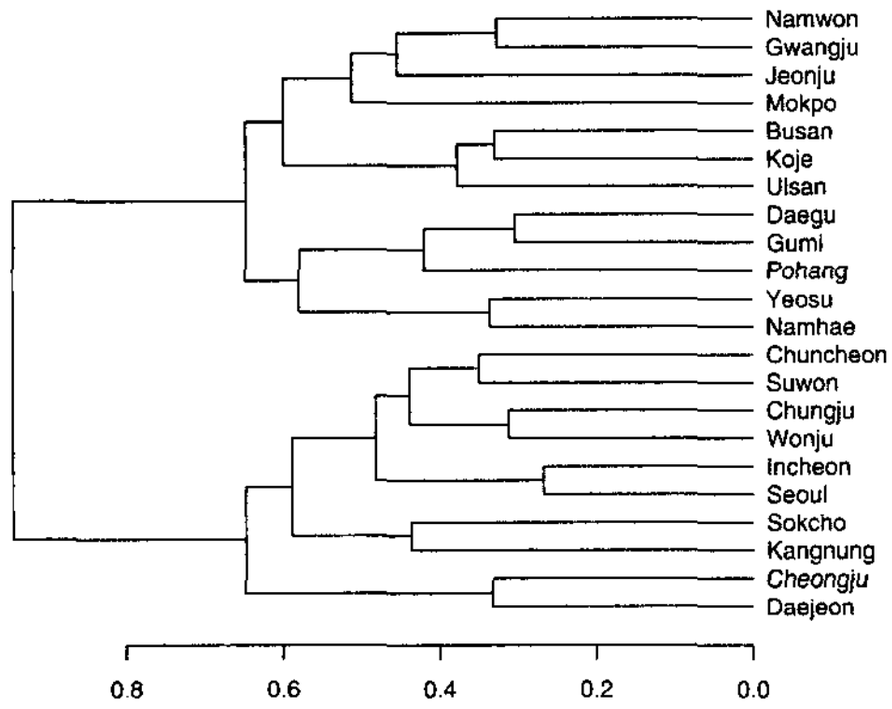
그림 3.3: 우리나라 22개 지역의 12차 차분된 월별 강수량 자료의 주기도

우에 G_1^{NI} ={대전, 청주, 서울, 인천, 춘천, 원주, 수원, 충주}, G_2^{NI} ={강릉, 속초}, G_3^{NI} ={거제, 부산, 남해, 여수, 목포, 광주, 남원, 울산, 포항, 전주, 구미, 대구}임을 나타낸다. 그림 3.4 (d)는 $d_{NI}^*(\cdot)$를 사용한 결과로, G_1^{*NI}={광주, 목포, 울산, 거제, 부산, 남해, 여수}, G_2^{*NI}={포항, 구미, 대구, 대전, 청주, 남원, 전주} 그리고 G_3^{*NI}={강릉, 속초, 서울, 수원, 인천, 춘천, 원주, 충주}로 나누어 진다. 군집분석의 평가 방법 또한 다양한 측도가 있으나, 덴드로그램의 가로축을 나타내는 비유사성측도(dissimilarity measure)의 수치를 주목하면 평활된 주기도를 이용한 거리($d_{NI}^*(\cdot)$)가 주기도를 이용한 거리($d_{NI}(\cdot)$)보다는 더 적절하다고 판단된다. 그리고 $d_{NI}(\cdot)$에 의해서는 강릉과 속초 만을 하나의 군집(G_2^{NI})으로 형성한다. 그러나, 그림 3.4 (d)의 $d_{NI}^*(\cdot)$에 의한 덴드로그램에서는 강릉과 속초는 군집 G_3^{*NI}에 포함되고, 군집화 과정을 살펴보면 G_3^{*NI} 내에서 서울, 수원, 인천, 춘천, 원주, 충주에 비하여 맨 마지막에 G_3^{*NI}에 포함됨을 나타낸다. 이는 최근의 김희영과 박만식 (2007)의 연구에서도 나타낸 바와 같이 시계열 자료를 군집화하는 데 $d_{NI}^*(\cdot)$이 $d_{NI}(\cdot)$보다는 좀더 좋은 거리라는 사실과도 일치한다. 그림 3.4 (a)는 자기상관함수의 군집분석 결과로 d_{NI}^*의 군집분석결과와 다소 차이가 있다. 예를 들어, 광주와 속초(혹은 강릉)은 지리적으로 멀리 떨어져 있으나 자기상관함수에 의한 거리에서는 같은 군집으로 그룹화되었다. 편자기상관함수(그림 3.4 (b))에 의한 거리는 2개의 군집으로 구분하고 있으며 이는 군집분석으로 얻을 수 있는 정보가 d_{NI}^*에 의한 거리에 비해 충분하지 않음을 알 수 있다.

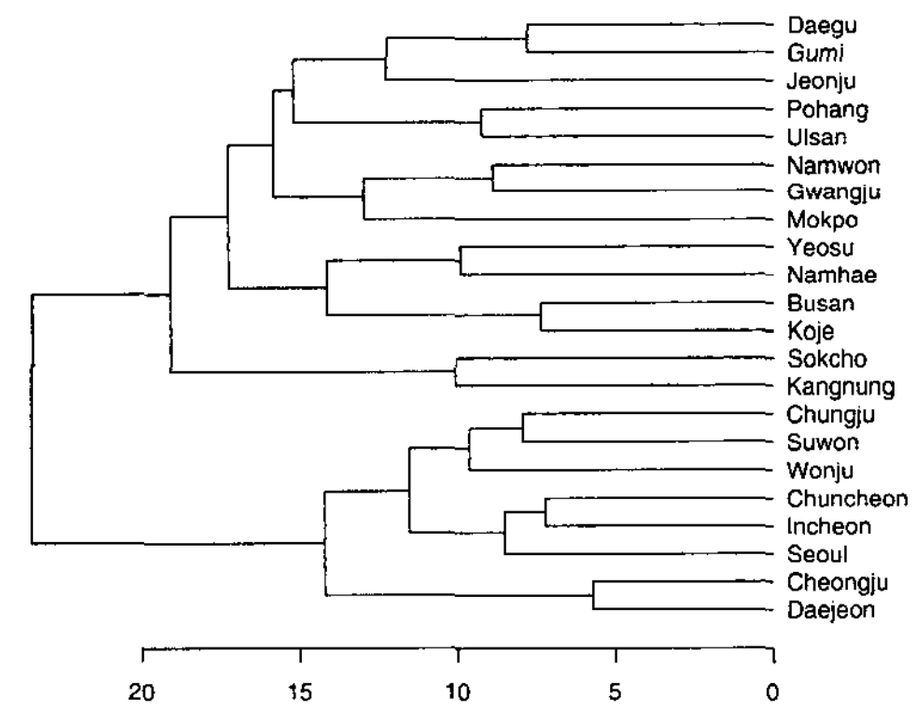
다음으로 그림 3.3 (b)의 결과에 의한 G_1^{*NI}, G_2^{*NI}, G_3^{*NI}의 군집별 특성을 알아보고자 한다. 표 3.2은 각 지역별 상위 5개의 $N\widetilde{I}_n(\omega_j)$에 해당하는 j를 정리한 것이다. 군집별 특징을



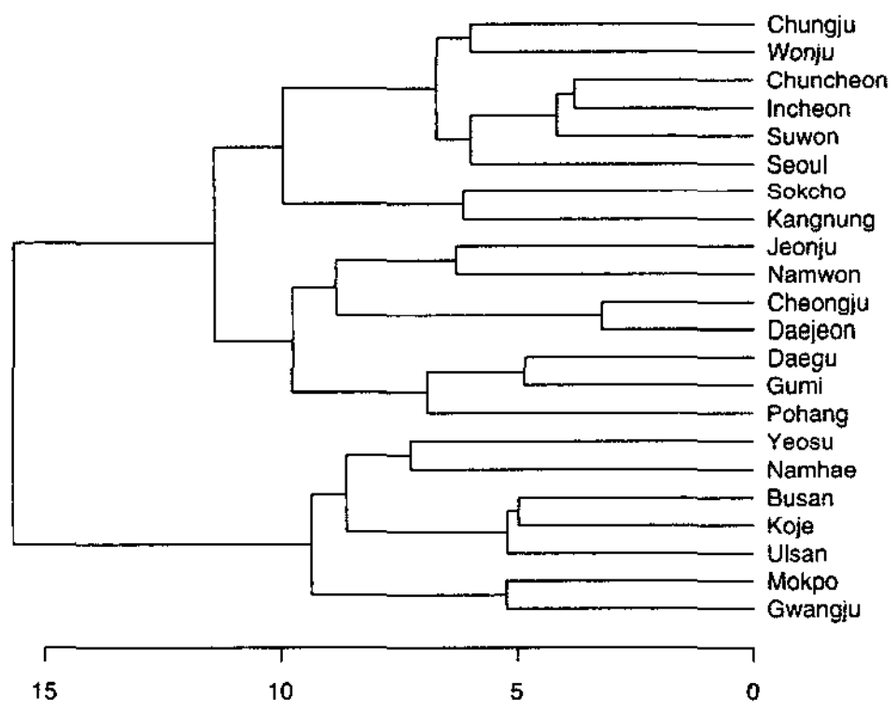
(a) autocorrelation



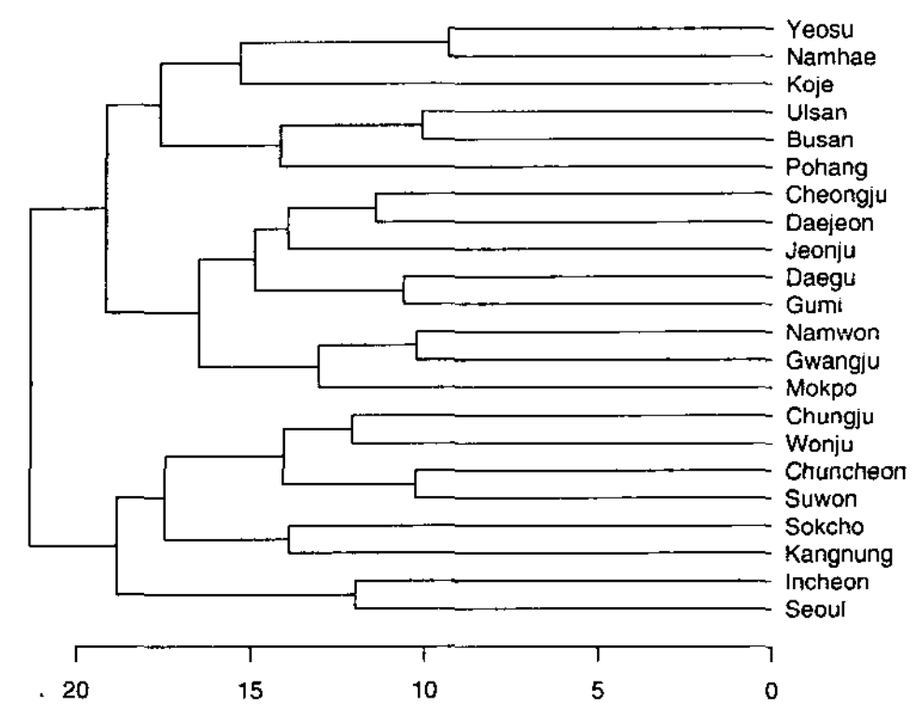
(b) partial autocorrelation



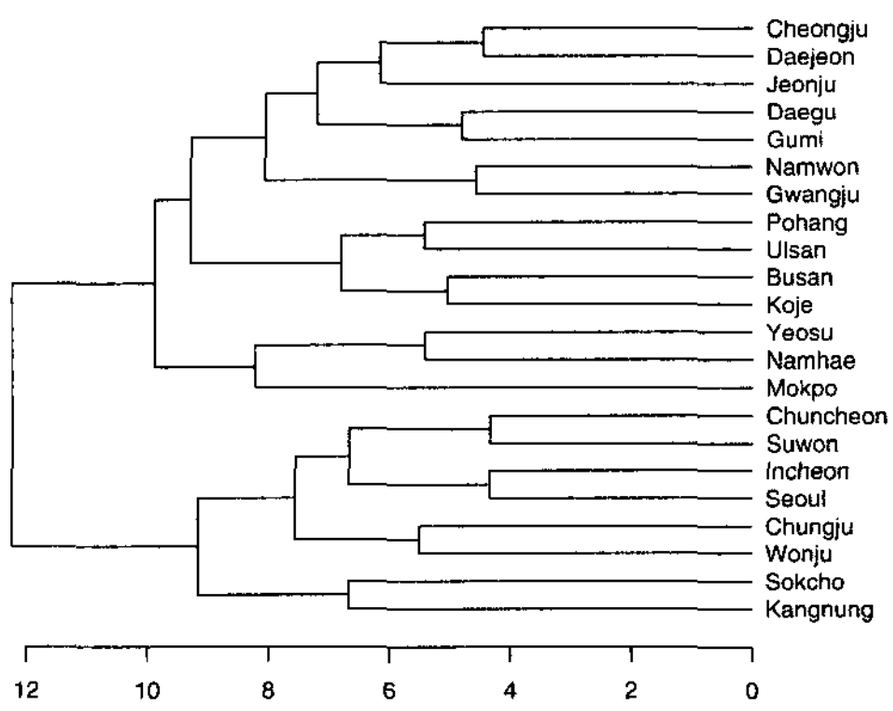
(c) NI



(d) $N\tilde{I}$



(e) $\log(NI)$



(f) $\log(N\tilde{I})$

그림 3.4: 계절차분된 자료의 군집분석에 의한 덴드로그램

표 3.2: 지역별 상위 5개 $N\widetilde{I}_n(\omega_j)$의 j

군집1						군집2						군집3					
지역	1	2	3	4	5	지역	1	2	3	4	5	지역	1	2	3	4	5
광주	31	32	90	33	11	포항	51	72	31	50	46	강릉	72	92	49	71	31
목포	31	32	14	12	13	구미	71	107	72	48	87	속초	49	48	111	92	72
남해	31	30	51	50	32	대구	107	106	87	31	51	서울	48	68	47	92	28
여수	31	51	107	14	30	대전	68	48	28	31	29	수원	48	92	68	91	72
거제	31	50	30	51	52	청주	68	48	28	31	51	인천	48	92	28	68	72
부산	31	50	30	49	10	남원	91	31	90	10	110	춘천	68	48	92	72	28
울산	31	51	32	46	50	전주	68	31	10	30	28	원주	92	72	91	71	111
												충주	48	68	72	31	92

알아보자. 군집 1은 평활된 주기도를 표준화한 $N\widetilde{I}_n(\omega_j)$에서 주로 j가 30–32와 50, 51이 지배적임을 알수 있다. 따라서, 12차 차분후의 월별 강수량이 $240/31 = 7.74$의 주기와 $240/51 = 4.7$ 주기에 의해 설명되는 시계열 구조를 가진다고 할 수 있다. 군집 3은 평활된 주기도를 표준화한 $N\widetilde{I}_n(\omega_j)$에서 j가 48, 49와 68–72, 91, 92일 때 큰 값을 보인다. 즉, $240/48 = 5.00$, $240/70 = 3.43$, $240/91 = 2.64$ 주기가 군집 3을 특징짓는다고 할 수 있다. 군집 2는 군집 1과 3에 비해 다소 동질성이 떨어지는 것으로 보아 군집 1과 3에 속하지 않은 나머지 지역들의 경향을 보여주고 있다.

4. 결론

본 연구에서는 시계열 자료를 군집화하기 위한 여러 시도들을 수행하였다. 기본적으로 시계열 자료분석시 모형설정과 예측을 수행해야 할 계열의 수가 상당히 많은 경우에, 주어진 계열 각각의 모형을 적합하고 예측하는 것보다는 유사한 경향을 보이는 계열들을 몇 개의 군집으로 그룹화한 후, 각 군집에 포함된 계열들을 군집분석의 결과를 토대로 재표현하는 것이 상당히 효율적일 것이다. 이에 본 연구에서는 시계열 자료를 군집화하기 위하여 사용되는 거리를 시간 영역에서와 주파수 영역에서로 나누어 간략히 소개하였다. 최근 주파수 영역에서의 거리들 중 김희영과 박만식 (2007)이 제시한 평활된 주기도에 의한 거리가 기존의 주기도에 의한 거리보다 정상시계열들을 AR모형과 MA모형으로 그리고 정상 시계열과 비정상 시계열로 보다 잘 군집화함을 보였다. 그리고 그 연구의 연장선에서 실제 우리나라의 22개 지역의 1987년 1월부터 2007년 12월까지 월별 강수량 자료를 이용하여, 시간영역에서 정의된 거리, 주기도에 의한 거리 그리고 평활된 주기도에 의한 거리를 사용하여 군집화하였다. 그 결과 군집의 개수를 3으로 했을때, 평활된 주기도에 의한 거리가 자료의 군집화 결과와 해석에 있어서 보다 만족스럽다는 것을 알 수 있었다. 만약 관측지점의 수를 보다 많이 확보한다면, 좀더 세분화된 군집결과를 얻을 것으로 기대된다.

참고문헌

고정웅, 백희정, 권원태 (2005). 한반도 우기의 강수 특성과 지역 구분, *Asia-Pacific Journal of Atmospheric Sciences*, **41**, 101-114.

김성렬, 양진석 (1995). 한국의 온대 저기압성 강수지역 구분, <한국지역지리학회지>, **1**, 45-60.

김희영, 박만식 (2007). Clustering time-series based on frequency domain, <한국통계학회 추계학술발표회 논문집>, **73**.

문영수 (1990). 클러스터분석에 의한 한국의 강수지역 구분, *Asia-Pacific Journal of Atmospheric Sciences*, **26**, 203-215.

이동규, 박정균 (1999). 군집 분석을 이용한 남한의 여름철 강수 지역 구분, *Asia-Pacific Journal of Atmospheric Sciences*, **35**, 511-518.

이승호 (1993). 계량적 분석에 의한 학국의 강수지역구분, <지역과 환경>, **11**, 1–15.

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time series, *Supplement to the Journal of the Royal Statistical Society*, **8**, 27-41.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, Springer-Verlag, New York.

Caiado, J., Crato, N. and Peña, D. (2006). A periodogram-based metric for time series classification, *Computational Statistics & Data Analysis*, **50**, 2668–2684.

Corduas, M. and Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric, *Computational Statistics and Data Analysis*, **52**, 1860–1872.

Fu, T. C., Chung, F. L., Ng, V. and Luk, R. (2001). Pattern discovery from stock time series using self-organizing maps, *KDD 2001 Workshop on Temporal Data Mining*, August 26-29, San Francisco, 27-37.

Galeano, P. and Peña, D. (2000). Multivariate analysis in vector time series, *Resenhas*, **4**, 383–403.

Goldstein, D. R., Ghosh, D. and Conlon, E. M. (2002). Statistical issues in the clustering of gene expression data, *Statistica Sinica*, **12**, 219–240.

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

Kalpakis, K., Gada, D. and Puttagunta, V. (2001). Distance measures for effective clustering of ARIMA time series, In *Proceedings of the 2001 IEEE international conference on data mining*, 273-280.

Liao, T. W. (2005). Clustering of time series data-a survey, *Pattern Recognition*, **38**, 1857-1874.

Maharaj, E. A. (2000). Clustering of time series, *Journal of Classification*, **17**, 297–314.

Pattarin, F., Paterlini, S. and Minerva, T. (2004). Clustering financial time series: An application to mutual funds style analysis, *Computational Statistics & Data Analysis*, **47**, 353-372.

Piccolo, D. (1990). A distance measure for classifying ARIMA models, *Journal of Time Series Analysis*, **11**, 153-164.

Shumway, R. H. (2003). Time-frequency clustering and discriminant analysis, *Statistics & Probability Letters*, **63**, 307-314.

[2008년 3월 접수, 2008년 4월 채택]

Classification of Precipitation Data Based on Smoothed Periodogram

Man Sik Park[1)] Hee-Young Kim[2)]

ABSTRACT

It is well known that spectral density function determines auto-covariance function of stationary time-series data and smoothed periodogram is a consistent estimator of spectral density function. Recently, Kim and Park (2007) showed that smoothed-periodogram based distances performs very well for the classification. In this paper, we introduce classification methods with smoothed periodogram and apply the approaches to the monthly precipitation measurements obtained from January, 1987 through December, 2007 at 22 locations in South Korea.

Keywords: Periodogram, smoothing, spectral density, clustering, precipitation.

1) Research professor, Dept. of Preventive Medicine, Korea University, Seoul 136-701, Korea.
E-mail: bayesia@korea.ac.kr
2) Corresponding author. Research professor, Dept. of Preventive Medicine, Korea University, Seoul 136-701,
Korea.E-mail: starkim@korea.ac.kr