The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.7.1445
29(7), 1445–1458

Temperature network analysis of the Korean peninsula linking by DCCA methodology

Seungsik Min[a,1]

[a]Department of Natural Science, Korea Naval Academy

(Received November 14, 2016; Revised December 17, 2016; Accepted December 23, 2016)

Abstract

This paper derives a correlation coefficient using detrended cross-correlation analysis (DCCA) method for 59 regional temperature series for 40 years from 1976 to 2015. The average temperature, maximum temperature, and minimum temperature series for 4 year units are analyzed; consequently, we estimated that a temperature correlation exists between the two regions during the unit period where the correlation coefficient is greater than or equal to 0.9; subsequently, we construct a network linking the two regions. Based on network theory, average path length, clustering coefficient, assortativity, and modularity were derived. As a result, it was found that the temperature network satisfies a small-worldness property and is a network having assortativity and modularity.

Keywords: temperature, detrended cross-correlation analysis (DCCA), temperature network, average path length, clustering coefficient, assortativity, modularity, small-worldness

1. 서론

적도를 지나는 동태평양 해안의 기온이 비정상적으로 상승하여 각종 기상이변을 일으키는 엘니뇨는 한동안 잠잠하다가 2015년부터 다시 강력해지기 시작하여 2016년에는 전세계적으로 유례없는 폭염을 유발했다. 2015–2016년 사이 매우 강한 엘니뇨가 발생할 것이며 이로 인해 올해(2016년)는 역사상 가장 더운 해가 될 것이라는 것은 세계기상기구(WMO)에서 이미 예상한 바 있다. 프레온가스와 각종 온실가스 등으로 인한 오존층 파괴, 대기 환경 변화 등으로 인한 각종 기후 변화는 10–20년 후를 내다보기 힘들 정도이며, 특히 매우 빠른 주기를 가지고 변화하는 게릴라성 폭우, 태풍 등의 정확한 예보는 정교한 알고리즘과 성능 좋은 슈퍼컴퓨터를 동원하고 있는 지금도 불가능에 가깝다.

기상 요소 중 일상생활과 가장 밀접한 관계가 있는 것이 바로 기온과 강수량이다. 강수일은 준주기적이며 그 양의 변동성이 커서 예측이 어렵고 홍수나 가뭄으로 인한 피해가 막심하다. 반면 기온 변화는 주기적이며, 변동성이 적어 예측이 상대적으로 쉽다. 하지만 기온은 일상생활과 24시간 관련되어 있기 때문에 그 특성을 파악하는 것은 사회적으로나 경제적으로 매우 중요하다. 기온은 강수량과 달리 매일 매일 측정이 되는 항목이기 때문에 상대적으로 풍부한 데이터를 확보할 수 있는 이점이 있다.

[1]Department of Natural Science, Korea Naval Academy, 1, Jungwon-ro, Changwon-si, Gyeongsangnam-do 51704, Korea. E-mail: fieldsmin@gmail.com

Table 2.1. 59 surveyed areas for air temperature analysis for Jan. 1. 1976–Dec. 31. 2015

권역	관측지역 수	관측 지역
강원	7	강릉, 대관령, 속초, 원주, 인제, 춘천 홍천
경기	6	서울, 인천, 강화, 수원, 양평, 이천
충북	5	보은, 제천, 청주, 추풍령, 충주
충남	6	대전, 금산, 보령, 부여, 서산, 천안
전북	5	남원, 부안, 임실, 전주, 정읍
전남	7	광주, 고흥, 목포, 여수, 완도, 장흥, 해남
경북	10	대구, 구미, 문경, 영덕, 영주, 영천, 울릉도, 울진, 의성, 포항
경남	10	부산, 울산, 거제, 거창, 남해, 밀양, 산청, 진주, 통영, 합천
제주	3	서귀포, 성산, 제주
총계	59	

South Korea can be divided into 9 districts, each of which shows the number of observations and specific regional names.

기상에 관한 선행연구들은 많은 분야의 연구자들에 의해 수행되었는데, 궁극적인 목적은 계절적인 큰 추세(mega trend)를 정확히 파악하거나, 그 속에 숨어 있는 작은 규칙(micro trend)을 찾는 것이라고 하겠다. Kang과 Ahn (2006)은 분산분석을 통해 기온과 강수량 자료의 추세 함수를 찾고자 하였고, Ko (2007)는 카이제곱 적합도 검정을 통해 일일 최고기온 분포 함수를 추정하였다. 또한 Kim 등 (2013)은 일반화 선형 모형(generalized linear model)을, Lee와 Sohn (2008)은 구조적 시계열 모형(structural time series model)을 이용하여 기온의 분석 및 예측을 시도하였다. 한편, Sohn 등 (2008)은 자기회귀 모형(auto-regression model)을, Kim과 Kim (2013)은 역거리 가중법(inverse distance weighting)을 이용하여 대한민국 기온 분포의 공간적 특성을 조사하였다. 앞의 연구들은 기상 현상의 큰 추세(mega trend)를 연구한 것이라고 볼 수 있다.

물리학계에서는 Podobnik과 Stanley (2008)가 detrended cross-correlation analysis(DCCA) 방법을 창시했는데, 비정상 시계열(non-stationary time series)의 추세를 제거하여 상관관계를 분석하는 방법을 제시하였다. 이후, DCCA에 대한 연구는 Horvatic 등 (2011)과 Podobnik 등 (2011) 등에 의해 많은 연구가 수행되었다.

본 논문에서는 DCCA 방법을 이용하여 기온 시계열 자료의 상관성을 분석하여 대한민국의 기온 네트워크를 구축하였다. 이로부터 평균 경로 길이(average path length), 결집 계수(clustering coefficient), 유사성(assortativity), 모듈성(modularity) 등의 값들을 도출하고 그들의 의미를 도출하였다.

2. 분석 데이터

2016년 9월 1일 현재 기상청에서 제공되는 과거 기상 관측 자료는 총 95개소이다. 본 연구는 한반도의 장기적 기온 분포를 분석하는 것이 목적이다. 따라서 1976년 1월 1일부터 2015년 12월 31일까지 40년간, 14,610일을 조사 대상일로 설정하고, 각 일자 별 평균기온, 최고기온, 최저기온을 조사대상으로 하였다. 지난 40년 간 결측치 없이 기온 데이터가 측정된 지점은 총 59개이다. 이 같은 양은 한반도의 기온 특성을 조사하기에 적절한 것으로 여겨진다(Table 2.1).

기온 데이터는 계절적 요인이 가장 극명하게 반영되는 시계열이므로, 계절 추세를 제거하기 위해 detrended cross-correlation analysis(DCCA) 분석 기법을 사용하였다. 본 연구에서는 DCCA 분석으로 추세를 제거한 후 시계열 간의 상관 계수가 0.9 이상인 경우 각 지역들끼리 상관성이 있는 것으로 해석하였다. 이 때 임계값을 반드시 0.9로 설정할 필요는 없다. 다만 임계값을 0.95 이상으로 설정할 경우

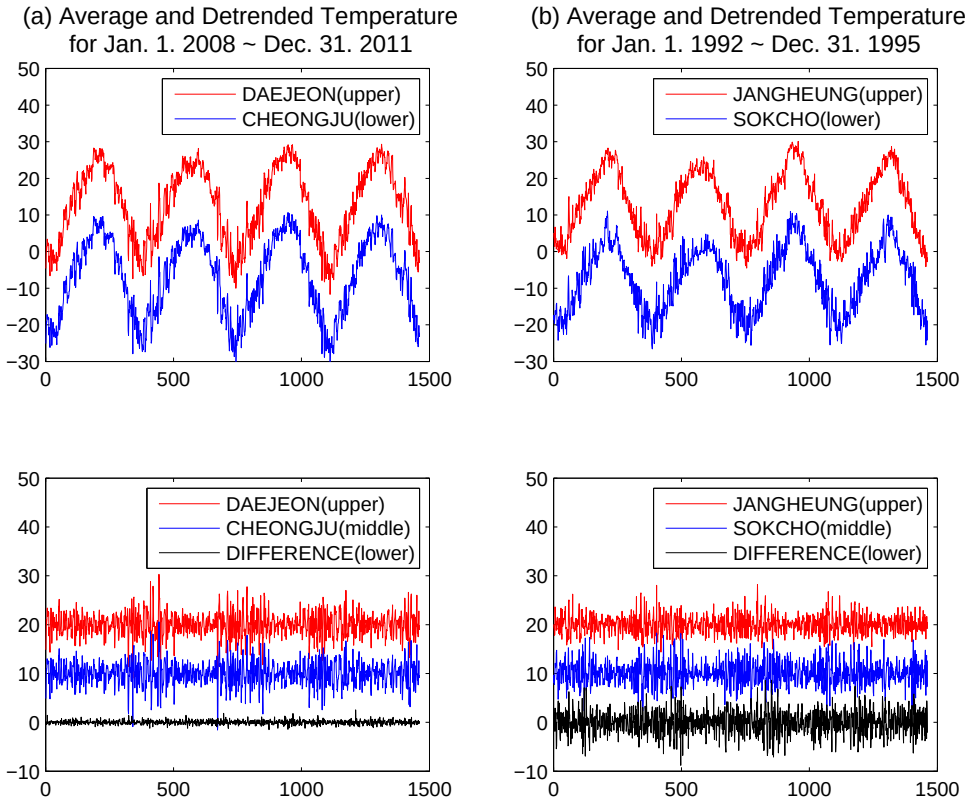


Figure 2.1. Example of time series pairs to analyze 4-year units. For visualization, the lower time series is shifted downward by 20. (a) maximal correlated pairs: Daejeon and Cheongju, (b) minimal correlated pairs: Jangheung and Sokcho. (*data source: Korea Meteorological Administration Homepage, http://www.kma.go.kr)

기온 네트워크는 연결선이 거의 없는 구조가 되고 0.85 이하로 설정할 경우 연결선 수가 기하급수적으로 늘어나 지역적 요인 이외의 요소가 개입될 여지가 있어 가능한 연결선 개수 1,711개 중 48개가 형성되는 0.9를 임계값으로 설정하였다. 가능하다면 임계값을 변화시켜 가며 네트워크 특성의 동역학적 변화를 보는 것도 의미 있는 연구가 될 것이다. 우리는 각 4년(1,461일) 단위 데이터를 가지고 평균기온, 최고기온, 최저기온 모두에서 상관성이 있을 경우 두 지역을 연결하였다. 모든 지역 간의 상관성을 조사하여 연결한 후 이들 10개 기간의 동역학적 네트워크 특성 변화를 분석하였다.

Figure 2.1 (a)는 상관관계가 가장 높은 지역인 대전과 충북 청주, (b)는 가장 낮은 지역인 전남 장흥과 강원 속초 기온 시계열의 예시이다. 전체적인 모양은 각각의 쌍에서 서로 비슷하게 나타나지만 추세를 제거한 시계열 쌍은 (a)의 경우 차이가 적지만, (b)는 차이가 크게 나는 것을 확인할 수 있다(가장 하단의 DIFFERENCE(lower)로 표시된 부분). 기온 특성이 지리적인 요인에 큰 영향을 받는 것으로 예상할 수 있듯이, 대전과 청주는 매우 인접한 도시들이고, 장흥과 강원도는 남서쪽과 북동쪽 끝에 위치하여 지리적으로 가장 먼 지역들이다. 피어슨 상관계수는 모든 지역들에서 10개 기간 모두 0.90 이상으로 매우 강한 관계를 가지나, DCCA 방법에 의한 추세 제거 후에는 가장 강한 상관계수가 2008–2011년 기간 동안 대전-청주 쌍에서 0.9858, 가장 약한 상관계수가 1992–1995년 기간 동안 장흥-속초 쌍에서 0.3804로 극명하게 차이가 났다.

3. 배경 이론

3.1. Detrended cross-correlation analysis (Podobnik과 Stanley, 2008; Min과 Lim, 2016)

두 시계열 자료 간의 상관성을 파악하는 가장 일반적인 방법 중의 하나가 피어슨 상관계수(Pearson correlation coefficient)를 알아보는 것이다. 하지만 피어슨 상관계수는 정상시계열(stationary time series)에만 적용할 수 있으며, 추세를 제거할 수 없다는 단점이 있다. 하지만 detrended cross-correlation analysis(DCCA) 방법은 시계열의 정상성에 관계없이 추세를 제거하는 것이 가능하다.

일반적인 시계열 자료는 평균과 분산이 시간의 흐름에 따라 변동하는 경우가 많다(non-stationarity). 이런 경우 일종의 광역 추세(global trend)인 평균과 분산을 분석하는 것은 바람직하지 않으며, 전체 시계열을 구간으로 나누어 지역 추세(local trend)를 분석하여 종합하는 것이 바람직하다. 전체 길이가 N인 어떤 시계열 $X_1(t)$와 $X_2(t)$가 있을 때, 이를 길이 n인 구간으로 나누었다고 하자. 그러면 k번째 구역에서의 변동(fluctuation)은 다음과 같이 나타난다.

$$f^2_{\mathrm{DCCA}}(X_1, X_2; n, k) = \overline{(X_1 - \tilde{X}_{1,k})(X_2 - \tilde{X}_{2,k})}.$$

이 때 $\tilde{X}_{i,k}$는 k번째 구역에서의 지역 추세인데 일반적으로 선형함수(linear fitted line), 혹은 이차함수(quadratic fitted curve)를 많이 사용한다. 본 논문에서는 선형함수를 사용하였다. 그러면 전체 변동의 기댓값은 다음과 같이 도출된다.

$$F^2_{\mathrm{DCCA}}(X_1, X_2; n) = \overline{f^2_{\mathrm{DCCA}}(X_1, X_2; n)}.$$

특히 $X_1 = X_2$인 경우 우리는 F^2_{DFA}(F-square of detrended fluctuation analysis)로 표현한다. 만약 시계열이 정상적(stationary)이라면, 지역 추세를 광역 추세로 바꿀 수 있으며, 선형함수나 이차함수 대신 상수함수를 사용한다면 $F^2_{\mathrm{DCCA}}(X_1, X_2; N)$는 $\mathrm{Cov}(X_1, X_2)$와 정확하게 일치한다. 결과적으로 추세를 제거하기 위해 분할한 구역의 길이에 대해 다음과 같이 $F^2_{\mathrm{DFA}}(X_1)$의 경향성을 살펴볼 수 있다.

$$F^2_{\mathrm{DFA}}(X_1) \sim n^{2\alpha}, \quad n\text{은 각 구역의 길이.}$$

이때 시계열 자료가 무작위적(randomized)이면 $\alpha = 0.5$가 되고, 지속성을 띤다면(persistent) $\alpha > 0.5$, 반지속성을 띤다면(anti-persistent) $\alpha < 0.5$가 된다. F^2_{DCCA} 역시 F^2_{DFA}와 비슷한 성향을 나타내는 것으로 알려져 있다. 그러면 두 시계열 간 상관계수가 다음과 같이 도출되며 -1에서 1사이의 범위 값을 갖는다.

$$\rho_{\mathrm{DCCA}}\,(X_i, X_j) = \frac{F^2_{\mathrm{DCCA}}(X_i, X_j)}{\sqrt{F^2_{\mathrm{DFA}}(X_i)}\sqrt{F^2_{\mathrm{DFA}}(X_j)}} \sim n^{2\beta_{ij} - \alpha_i - \alpha_j}.$$

본 연구에서는 40년간의 기온 데이터를 4년 단위로 구분하였고 4년 치 데이터 (N = 1,461)를 시계열의 길이로 취급하여 분석을 실시하였다. 또한 각 구역의 길이 n은 30부터 10씩 늘려가며 90까지 시행하였다. 예를 들어, 구역의 길이가 30일인 경우, 1,461일 간의 데이터를 49 등분하여 매 구역마다 선형함수를 적합시키고(fitting), 추세를 제거한 후 $F^2_{\mathrm{DCCA}}(X_1, X_2; 30)$ 값을 구하였다. 같은 방법으로 구역의 길이가 90일인 $F^2_{\mathrm{DCCA}}(X_1, X_2; 90)$ 값까지 구하였다. 결과적으로 두 시계열에 대해서 구한 ρ_{DCCA}는 각 구역의 길이 n에 대해 멱함수 형태를 띠게 되는데, 이번 연구에서는 두 지역의 고온, 저온, 평균기온 시계열에서 각각 구한 $\rho_{\mathrm{DCCA}}(X_1, X_2; 30) \sim \rho_{\mathrm{DCCA}}(X_1, X_2; 90)$ 모두에서 임계값을 넘어설 경우, 두 시계열 X_1과 X_2는 4년의 기간 동안 연계성이 있다고 파악하였다. 나아가 40년 동안 10개의 단위 기간 중

9개 이상의 기간에서 연계성이 파악된 경우 두 지역 간 네트워크 연결선을 구축하였다. 온도 분포에 대해 DCCA 분석을 실시한 이유는 각 지역 별 온도 시계열의 피어슨 상관계수가 강력한 계절적 추세 때문에 대부분 0.95이상, 심지어 0.99를 초과하는 경우도 많은 데에 반해, 추세를 제거한 시계열은 계절 이외의 요인(특히 지역적 요인)에 의한 상관성을 볼 수 있다는 이점이 있기 때문이다.

3.2. Network theory

최근 들어 복잡계(complex system)에 대한 연구가 물리를 비롯한 자연현상 및 사회, 경제 현상 전반에 대해서도 활발하게 수행되고 있다. 복잡계란 계를 구성하고 있는 하부 구조의 상호작용이 하부 구조 개별 작용의 합보다 큰 시너지 효과를 발휘하는 계를 말한다 (Kang, 2010). 이 같은 복잡계를 기술할 수 있는 훌륭한 모형의 하나가 바로 네트워크 이론인데, 복잡계에 대한 관심과 더불어 네트워크 이론 역시 생물학, 화학, 경제학, 사회과학 등을 대상으로 최근 십여 년 동안 폭발적인 연구가 수행되고 있다 (Wang 등, 1989; Plerou 등, 2000; Albert과 Barabasi, 2002; Abe과 Suzuki, 2003; Greene와 Higman, 2003; Yook 등, 2004; Estrada와 Hatano, 2008).

네트워크는 노드와 링크로 구성된다. 노드는 개별 입자일 수도 있고, 조직 구성원, 특정 지역일 수도 있다. 한편 링크는 무엇을 노드들 간의 상호작용으로 해석하느냐에 따라 매우 다양하게 정의될 수 있다. 본 논문에서는 각 지역 기상 관측소가 노드에 해당하며, 평균기온과, 최고기온, 최저기온 모두에서 상관계수가 0.9 이상이 나올 경우 두 노드들 간에 링크가 형성된다고 판단한다. 노드와 링크를 구성하면 다양한 분석이 가능하며, 우리는 네트워크 이론의 대표적인 속성 값들을 계산할 것이다.

3.2.1. 평균 경로 길이(average path length) i번째 노드와 j번째 노드 사이의 거리를 L_{ij}고 할 때 평균 경로 길이(average path length)는 각 노드들 사이 거리 L_{ij}의 평균으로 정의된다.

$$L = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{i-1} L_{ij}.$$

이 때, 노드들 사이의 거리는 i번째 노드에서 j번째 노드로 가기 위해 거쳐 가는 최소 노드 수에 1을 더한 값이다. 만약 두 노드를 잇는 경로가 존재하지 않는다면 $L_{ij} = \infty$가 되어 L이 발산하게 된다. 이런 경우 L_{ij}가 유한한 값들에 대해서만 평균을 하게 된다. 이와 같은 과정으로 구해진 평균 경로 거리는 네트워크가 좁은 세상(small-worldness) 성질이 있는지 여부를 판가름하는 지표로 이용된다. 전체 노드의 수 N인 네트워크에 대해 만약 L이 $\log N$에 비례하거나 그보다 작다면 그 네트워크는 좁은 세상 네트워크(small world network)라고 부른다.

3.2.2. 결집 계수(clustering coefficient) i번째 노드와 직접 연결된 노드 수가 k_i개라고 하면, k_i개의 링크들 중 임의의 2개는 삼중선(triplet)을 이룬다. 이 중 삼각형을 이루는 비율을 i번째 노드의 결집 계수가 된다. 즉,

$$C_i = \frac{\text{number of triangle}}{\text{number of triplet}} = \frac{2}{k_i(k_i-1)} \times \text{number of triangle},$$
$$\text{for the neighbourhood of node } i.$$

이들 C_i의 평균을 결집 계수(clustering coefficient)라고 부른다.

$$C = \frac{1}{N} \sum_{i}^{N} C_i.$$

결집 계수는 삼중선(triplet)에 대한 삼각형(triangle)의 비중을 나타내는 것이다. 결집 계수 역시 주어진 네트워크의 좁은세상 성질을 알아보거나, 네트워크가 얼마나 견고하게 결합되어 있는지 알아보는 도구로 이용된다.

3.2.3. 유사성(assortativity) (Newmann, 2002) 네트워크가 많은 연결선을 가진 노드들끼리, 또는 적은 연결선을 가진 노드들끼리 연결되려는 성향이 강하면 유사성 네트워크(assortative network)라고 말하며, 다음의 유사성 계수를 이용하여 값을 도출한다.

$$r = \frac{1}{\sigma_q^2} \sum_{j,k} jk \left(e_{jk} - q_j q_k \right),$$

이때 $q_j = (j+1)p_{j+1} / \sum_{i=1}^{N} i p_i$는 임의의 연결선에 대해 양 끝 노드가 여분의 연결선이 j개일 확률을 나타낸다. 여기서 p_i는 임의의 노드에 대해 연결선 수가 i개일 확률이다. 또한 e_{jk}는 양 끝 노드가 j개, k개일 결합 확률(joint probability)을 나타낸다. 한편, σ_q^2는 q_j 분포의 표준편차와, q_k 분포의 표준편차의 곱인데, 둘의 분포는 동일하므로 q_j 분포의 분산이라고 표현해도 무방하다. 결국 유사성 계수는 여분 연결선 분포의 상관계수이기 때문에 -1에서 1사이의 값을 갖는다. 어떤 네트워크가 유사성 네트워크라면 r의 값은 양수로 나타나는데, 수학이나 과학 저널의 공저자 네트워크나 영화배우, 회사의 임원 네트워크 등은 유사성 계수의 값이 0.1–0.3 정도의 양수값을 가지는 반면 인터넷, 월드 와이드 웹, 단백질 상호작용, 신경망, 먹이사슬 등은 -0.1∼0.3 정도의 음수값을 가지는 것으로 알려져 있다. 한편, 단백질끼리의 상호작용 네트워크가 -0.2 정도의 음수값을 가지는 반면 단백질 내의 아미노산 상호작용 네트워크는 0.1–0.5 정도의 양수값을 가져, 같은 물질이더라도 관측 범위(resolution)에 따라 다른 유사성을 나타내는 것이 알려진 바 있다 (Min과 Kim, 2014).

3.2.4. 모듈성(modularity) (Newmann, 2004, 2006) 어떤 네트워크의 군집 구조(community structure)를 알아보는 방법 중의 하나로 모듈성을 계산할 수 있다. N개의 노드로 구성된 네트워크에서 임의의 노드 i와 j에 대해 연결선의 수를 k_i, k_j라 하면 모듈성은 다음과 같이 표현된다.

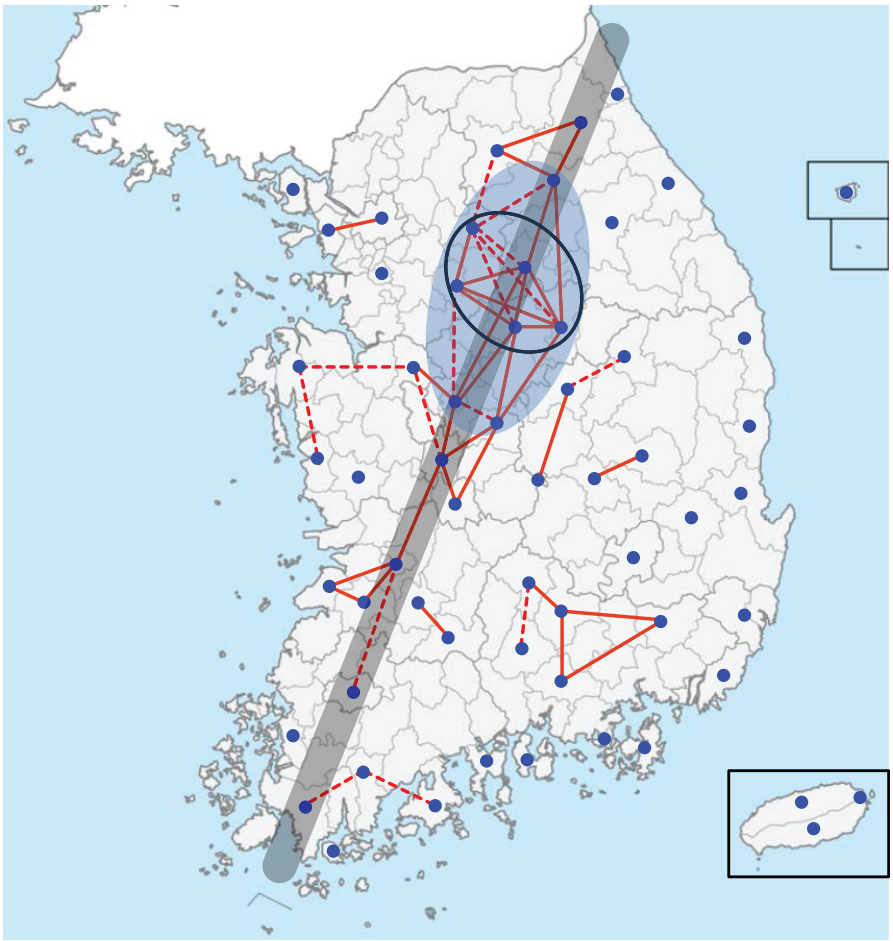
$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta \left(c_i, c_j \right).$$

이때 네트워크에 연결된 연결선의 총 개수는 $m = \sum_{i=1}^{N} k_i / 2$이므로, $k_i k_j / 2m$은 각 노드에 대해 연결선 수의 기댓값이 된다. 또한 A_{ij}는 인접행렬(adjacency matrix)을 나타내며 일반적으로 0 또는 1을 성분으로 가진다. 한편 $\delta(c_i, c_j)$는 노드 i가 속한 그룹 c_i와 j가 속한 그룹 c_j가 같을 경우에는 1, 다를 경우에는 0의 값을 갖는 크로네커 델타 함수(Kronecker delta function)이다. 일반적으로 0.3 이상의 값을 가지면 그룹 간의 분리가 잘된 경우로 판단한다.

4. 분석 결과 및 논의

4.1. DCCA 방법을 이용한 기온 네트워크 구축

본 연구에서는 지난 40년 동안 결측치 없는 대한민국 59개 지역의 기온 특성을 분석하였다. 매 4년 단위로 총 10개의 기간으로 구분하여 DCCA 분석을 실시하였는데, 개별 기간마다 총 1,711개(= $59 \times 58/2$) 지역 쌍들의 평균, 최고, 최저 기온 시계열의 상관계수(ρ_{DCCA})를 계산하여 세 값들이 모두 0.9를 넘어서면 그 지역 쌍들은 연계성이 있는 것으로 파악하였다.



* source of blank map : http://cafe.naver.com/geoarchive/993

Figure 4.1. Map of the Korean air temperature network: red solid line(—) means all correlation coefficients of 10 periods are greater than or equal to 0.9, and red dotted line(- - -) means 9 correlation coefficients are greater than or equal to 0.9.

Figure 4.1은 조사 지역을 파란색 점(•)으로 표시하고, 총 10개의 기간 중 9개 이상의 기간에서 연계성이 파악되는 경우 두 노드 간 빨간색 연결선(—, 또는 - - -)을 구성한 것이다. 이 때 실선은 10개의 기간 모두에서 연계성이 파악되는 경우이고, 점선은 9개 기간에서 연계성이 파악되는 경우이다. 총 59개의 지역 중 36개 지역에서 48개의 연결선을 가지는 것을 알 수 있다. 기온 네트워크는 강원도 북부지역에서부터 전라도 남쪽지역으로 대각선 모양(음영 처리된 선분)으로 큰 축을 이루는 것을 확인할 수 있다. 이 같은 현상의 원인은 산맥을 주축으로 하는 지형적인 요인과 더불어 편서풍에 의한 영향 등으로 추정해 볼 수 있으며, 정확한 원인은 후속 연구를 통해 조사할 필요가 있을 것으로 사료된다. 한편 대부분의 연결선은 충청도를 중심으로 형성되는데(음영 처리된 타원 내부), 특히 {양평, 이천, 원주, 충주, 제천}의 다섯 지역(검은 실선으로 표시된 원 내부)은 모든 노드들이 서로 연결된 완전 그래프(complete graph) 형태를 띠고 있어 매우 강한 기온 네트워크를 형성하는 것을 알 수 있다.

Table 4.1. The number of connections (degree) for air temperature network

연결선 수	해당 지역 수	해당 지역(연결된 지역)
6	5	**양평**(원주, 이천 제천, 춘천, 충주, 홍천), **원주**(양평, 이천, 제천, 청주, 충주, 홍천), **제천**(보은, 양평, 원주, 이천, 충주, 홍천), **청주**(대전, 보은, 원주, 이천, 천안, 충주), **충주**(보은, 양평, 원주, 이천, 제천, 청주)
5	4	**대전**(금산, 보은, 전주, 천안, 청주), **보은**(금산, 대전, 제천, 청주, 충주), **이천**(양평, 원주, 제천, 청주, 충주), **홍천**(양평, 원주, 인제, 제천, 춘천)
4	1	**전주**(광주, 대전, 부안, 정읍)
3	3	**천안**(대전, 서산, 청주), **춘천**(양평, 인제, 홍천), **합천**(거창, 밀양, 진주)
2	10	**거창**(산청, 합천), **금산**(대전, 보은), **문경**(영주, 추풍령), **밀양**(진주, 합천), **부안**(전주, 정읍), **서산**(보령, 천안), **인제**(춘천, 홍천), **장흥**(고흥, 해남), **정읍**(부안, 전주), **진주**(밀양, 합천)
1	13	**고흥**(장흥), **광주**(전주), **구미**(의성), **남원**(임실), **보령**(서산), **산청**(거창), **서울**(인천), **영주**(문경), **의성**(구미), **인천**(서울), **임실**(남원), **추풍령**(문경), **해남**(장흥)
연결선(link) 총수: 48개		**연결된 지역(node) 총수: 36개**

Areas in parentheses refer to areas connected to the area outside of parentheses. Since the administrative district is often divided into hexagonal shapes, it is rare to exceed the maximum number of connecting lines are greater than 6.

Table 4.1은 Figure 4.1에서 나타낸 각 지역들을 연결선 수에 따라 분류하여 나타낸 표이다. 행정 구역은 육각형 형태로 방사되는 것이 일반적이므로 인접 지역과 모두 연결되더라도 연결선 수는 최대 6개를 넘기기 힘들다. 따라서 연결선 수가 6개인 것은 주변 지역과의 상관성이 매우 높은 지역이라는 것을 추정할 수 있다. 표에서 연결선을 6개, 또는 5개 가지고 있는 대부분의 지역이 충청도임을 알 수 있다. 한편, 전라도와 경상도의 지역들은 상대적으로 약한 기온 네트워크를 형성하고 있다. 충청, 경기, 강원을 제외한 지역에서 가장 많은 수의 연결선을 가지고 있는 지역은 전북 전주로 4개의 연결선을 가지고 있다. 그 다음이 경남 합천으로 3개의 연결선을 가지고 있다. 전국적인 기준으로는 미약하지만, 지역적으로는 전주와 합천이 전라도와 경상도 지역 기온의 허브 역할을 하고 있는 것을 확인할 수 있다.

4.2. 네트워크 분석

Figure 4.2–Figure 4.5는 4년 단위로 구분된 10개 기간 동안 대한민국 59개 지역의 온도 네트워크에서 도출된 대표적인 속성 값들을 표시한 그림이다. Figure 4.2는 평균 경로 길이(average path length)를 나타낸 것이다. 이번 분석에서는 전체 노드의 수가 59개인 소규모 네트워크이므로 노드 수를 변화시켜 가며 그에 따른 평균 경로 길이의 값 변화 추이를 살펴보기는 힘들다. 다만 1983–2011년까지의 기간 중 무작위로 섞은 기온 자료(shuffled temperature data, 파란색 점선)에 비해 고려 대상 기온 자료(temperature data, 빨간색 실선)의 평균 경로 길이가 비교적 작게 나타나는 것으로 보아 좁은 세상 성질을 어느 정도 만족하는 것으로 여겨진다.

Figure 4.3은 결집 계수(clustering coefficient)를 나타낸 것으로 네트워크가 얼마나 견고한지를 알아보는 지표로 사용된다. 결집 계수는 무작위 자료에 비해 고려 대상 자료의 값이 유의미하게 큰 값을 가진다. 따라서 본 네트워크는 무작위 네트워크(random network)와는 확연히 다르게 결집도가 매우 높은 네트워크인 것을 알 수 있다.

앞서 평균 경로 길이 L이 $\log N$에 비례할 경우 좁은 세상 네트워크라고 지칭하였는데, 이번 분석과 같이 전체 노드의 수가 적거나 노드 수를 변형시키는 것이 어려울 경우 평균 경로 길이와 결집 계수를 결

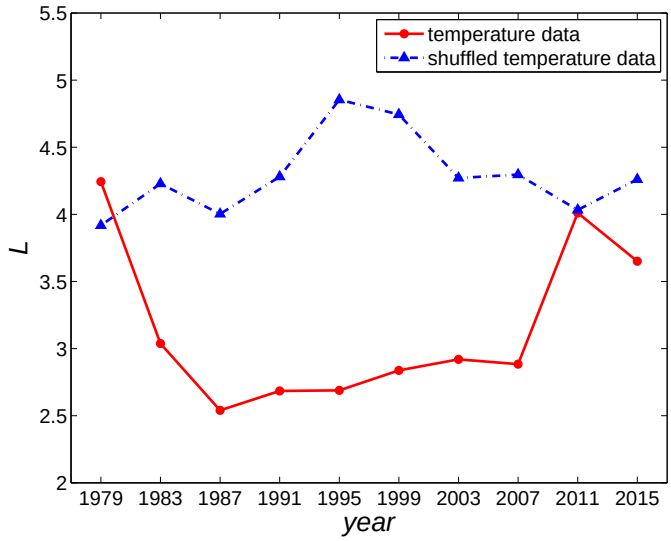


Figure 4.2. Characteristic path lengths of the temperature network for 10 periods of the South Korean 59 regions.

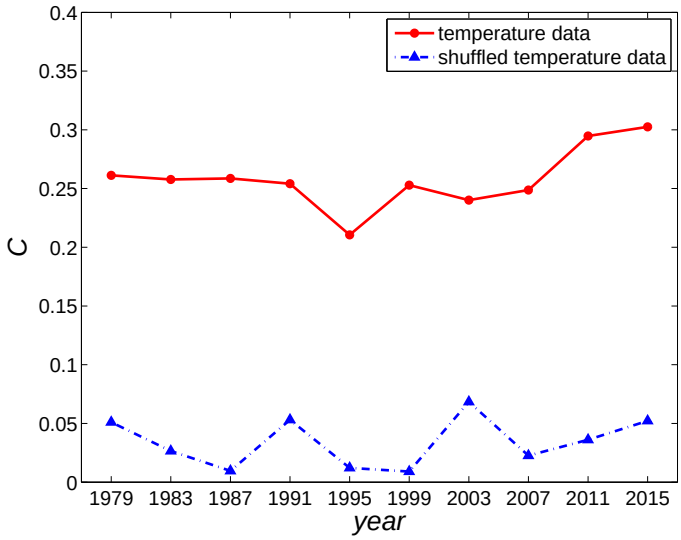


Figure 4.3. Clustering coefficients of the temperature network for 10 periods of the South Korean 59 regions.

합하여 보다 간단하게 좁은 세상 성질을 확인할 수 있다. 주어진 온도 네트워크의 평균 경로 길이와 결집 계수를 각각 L, C라 두고 무작위로 섞은 기온 데이터의 평균 경로 길이와 결집 계수를 각각 L_{ran}, C_{ran}이라 두면 좁은 세상 성질은 다음과 같이 계산된다.

$$\text{SW} = \frac{C/C_{\text{ran}}}{L/L_{\text{ran}}}.$$

따라서 주어진 네트워크는 L/L_{ran}이 1–2 정도의 범위를 갖고, C/C_{ran}이 5 이상의 범위를 가지므로

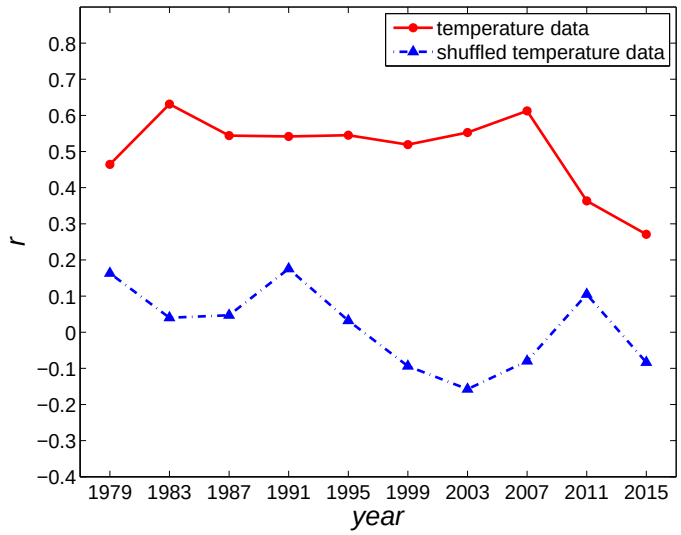


Figure 4.4. Assortativities of the temperature network for 10 periods of the South Korean 59 regions.

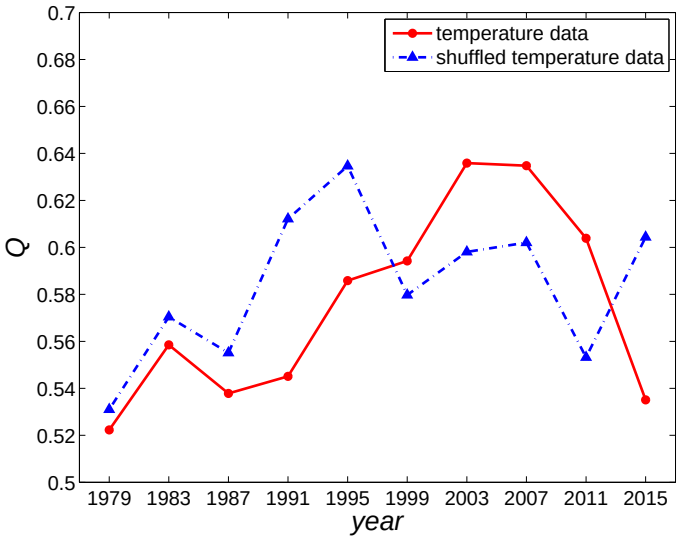


Figure 4.5. Modularities of the temperature network for 10 periods of the South Korean 59 regions.

SW는 5보다 큰 값을 가지게 되어 좁은 세상 성질을 만족한다고 판단할 수 있다.

Figure 4.4는 유사성(assortativity) 계수를 나타낸 것이다. 일반적인 네트워크는 -0.3~0.3 범위의 값을 가지는데, 본 연구에서 유사성은 0.5 수준에서 매우 높은 값을 가진다. 유사성 계수만 보더라도 연결선의 수가 5–6개 수준인 노드들(rich nodes) 간의 연결성이 강하게 나타나는 것을 손쉽게 추정할 수 있다.

마지막으로 Figure 4.5는 네트워크의 커뮤니티 구조를 나타내는 모듈성(modularity)을 구하여 표시한 것이다. 어떤 네트워크를 몇 개의 커뮤니티(그룹)로 구분할 때 그룹 내에서의 연결선은 평균 연결선 개

Table 4.2. Network properties of air temperature network of South Korea

기간	k	L	L_{rand}	C	C_{rand}	SW	r	r_{rand}	Q	Q_{rand}	NC	NC_{rand}
–1979	2.54	4.24	3.92	0.26	0.05	4.72	0.46	0.16	0.52	0.53	7	7
–1983	2.31	3.04	4.23	0.26	0.03	13.51	0.63	0.04	0.56	0.57	7	7
–1987	2.68	2.54	4.00	0.26	0.01	42.43	0.54	0.05	0.54	0.56	7	7
–1991	2.24	2.68	4.28	0.25	0.05	7.63	0.54	0.18	0.55	0.61	9	7
–1995	1.93	2.69	4.85	0.21	0.01	31.40	0.55	0.03	0.59	0.63	10	5
–1999	2.41	2.84	4.74	0.25	0.01	46.77	0.52	−0.09	0.59	0.58	7	7
–2003	2.34	2.92	4.27	0.24	0.07	5.14	0.55	−0.16	0.64	0.60	9	7
–2007	2.14	2.88	4.3	0.25	0.02	16.40	0.61	−0.08	0.63	0.60	9	9
–2011	2.75	4.01	4.03	0.29	0.04	8.21	0.36	0.10	0.60	0.55	8	6
–2015	2.68	3.65	4.26	0.30	0.05	6.76	0.27	−0.08	0.54	0.60	6	7
전체 기간	3.97	3.29	3.06	0.42	0.10	4.10	0.41	−0.01	0.54	0.45	6	7

k: average degree (average number of links); L: average path length; C: clustering coefficient; SW: small-worldness; r: assortativity; Q: modularity; NC: number of clusters grouped by Newmann modularity method (groups of a person are not counted); $\{\cdot\}_{\text{rand}}$: property of randomly shuffled data.

수에 비해 많고, 그룹 간의 노드들의 연결선은 평균 연결선 개수에 비해 적다면 모듈성은 1에 가까운 큰 값을 나타내고, 그 반대의 경우 0에 가까운 낮은 값을 나타낸다. 앞에서 구한 평균 경로 길이, 결집 계수, 유사성과 다르게 모듈성에 있어서 기온 자료와 무작위 자료는 큰 차이를 보이지 않는다. 다만 그 값이 0.5 이상의 높은 값을 가지므로 기온 네트워크는 커뮤니티 내에서 긴밀한 연계성을 띠고 있으며, 커뮤니티 간에는 거의 연계성을 가지지 않는 것을 추정할 수 있다.

Table 4.2는 Figure 4.2–Figure 4.5에서 나타낸 평균 경로 길이(L), 결집 계수(C), 유사성(r), 모듈성(Q)을 비롯하여, 평균 연결선 수(k), 좁은 세상 성질(SW), 커뮤니티 수(NC) 등을 나타내었다. 4년 단위로 기간을 분리하여 분석한 속성 값 외에 40년 전체 기간을 한꺼번에 분석한 속성 값을 별도로 표현하였다. 평균 경로 길이(L)와 모듈성(Q), 커뮤니티 수(NC)는 무작위 자료(shuffled data)와 조사 대상 기온 자료의 차별성이 크지 않아 모듈성 분석으로 기온 네트워크를 규명하는 것은 힘든 것으로 판단된다. 하지만 결집 계수(C)는 무작위 자료에 비해 조사 대상 기온 자료의 값이 매우 높게 나타나서 결과적으로 좁은 세상 성질(SW)이 높은 것을 확인할 수 있다. 또한 유사성(r)이 매우 높게 나타는 것을 알 수 있다. 따라서 기온 네트워크는 무작위 네트워크와도 차이를 보이지만, 대체적으로 -0.3∼0.3 범위의 유사성 계수를 갖는 자연, 생물, 사회망 등과도 차별성을 보이는 것을 다시 한 번 확인할 수 있다.

5. 결론

지금까지 2개의 단계를 통해 최근 40년간 대한민국 59개 지역의 기온 특성을 분석하였다. 1단계는 detrended cross-correlation analysis(DCCA)를 이용한 네트워크를 구축하는 단계이고, 2단계는 구축된 네트워크를 분석하여 의미 있는 값들을 도출해내는 단계이다.

시계열의 평균과 분산이 일정하지 않거나(non-stationary), 계절 요인 등 강한 추세를 가지는 경우 피어슨 상관계수를 이용하여 연계성을 파악하기 어려운데, DCCA는 주어진 시계열을 여러 개의 구간으로 분할하여 개별 구간 내에서 추세를 제거한 시계열의 상관계수를 구하는 방법이다. 1,711개의 지역 쌍 각각에 대해 총 40년의 기간을 4년 단위로 세분하여 각 기간에서 평균기온, 최고기온, 최저기온 모두 상관계수가 0.9를 넘어서면 두 지역은 상관성이 있는 것으로 판단하였다. 특히, 10개의 기간 중 9개 이상의 기간에서 상관성이 있는 것으로 판단되는 링크는 Figure 4.1에서 별도 표기하였다.

구축된 기온 네트워크는 크게 3가지 특성을 나타내고 있다. 우선 네트워크는 전라도 남부로부터 강원도 북부까지 한반도를 정확히 이등분하는 대각선 형태를 띠며 구축된다. 또한 충청권을 중심으로 강원 서부와 경기 동부에서 네트워크 허브가 형성된다. 특히, 양평, 이천, 원주, 충주, 제천의 5개 지역은 완전 그래프를 이루며 매우 강한 네트워크를 형성한다. 마지막으로 전라도와 경상도는 상대적으로 약한 네트워크를 형성하는데, 그 중 전라북도 전주와 경상남도 합천이 가장 많은 연결선을 가진 허브가 된다.

네트워크 이론을 바탕으로 기온 네트워크를 분석한 결과 2가지 사실을 알 수 있었다. 우선 네트워크의 결집 계수(clustering coefficient)가 매우 높게 나타나 좁은 세상 성질(small-worldness)을 만족하는 것을 알 수 있었다. 그리고 유사성(assortativity)과 모듈성(modularity)이 매우 높은 네트워크임을 알 수 있었다.

본 연구에서 연계성이 파악된 지역들은 산맥이나 분지, 해안 등 지형적 요인, 편서풍 등의 기후적 요인 등이 복합적으로 작용하여 나타난 결과일 것이다. 통계적인 추론을 통해 찾아낸 이들 연계성의 원인을 향후 구체적으로 연구한다면 기상 예보나, 대기 흐름의 예측 능력을 향상시키는 데에 도움이 될 것으로 기대된다.

References

Abe, S. and Suzuki, N. (2003). Law for the distance between successive earthquakes, *Journal of Geophysical Research*, **108**, 2113–2116.

Albert, R. and Barabasi, A. (2002). Statistical mechanics of complex networks, *Reviews of Modern Physics*, **74**, 47–97.

Estrada, E. and Hatano, N. (2008). Communicability in complex networks, *Physical Review E*, **77**, 036111.

Greene, L. and Higman, V. (2003). Uncovering Network Systems within Protein Structures, *Journal of Molecular Biology*, **334**, 781–791.

Horvatic, D., Stanley, H. E., and Podobnik, B. (2011). Detrended cross-correlation analysis for non-stationary time series with periodic trends, *EPL (Europhysics Letters)*, **94**, 18007.

Kang, B. (2010). Science of the complex network: information science of 21st century, *Asan Foundation Research Report*, 124, Jipmoondang Press, Seoul.

Kang, K. and Ahn H. (2006). Functional data analysis of temperature and precipitation data, *The Korean Journal of Applied Statistics*, **19**, 431–445.

Kim, H., Do, H., and Kim, Y. (2013). A modeling of daily temperature in Seoul using GLM weather generator, *The Korean Journal of Applied Statistics*, **26**, 413–420.

Kim, N. and Kim, G. (2013). A study on changes of the spatio-temporal distribution of temperature in Korea peninsular during the past 40 years, *Journal of the Korean Association of Geographic Information Studies*, **16**, 29–38.

Ko, W. (2007). Estimation for the change of daily maxima temperature, *The Korean Journal of Applied Statistics*, **20**, 1–9.

Lee, J. and Sohn, K. (2008). Trends in the climate change of surface temperature using structural time series model, *Atmosphere*, **18**, 199–206.

Min, S. and Kim, K. (2014). Topological properties of networks in structural classification of proteins, *Journal of the Korean Physical Society*, **65**, 1164–1169.

Min, S. and Lim, G. (2016). Analysis of Ocean Time Series by DCCA(detrended cross-correlation analysis) Methodology. *In Proceedings of the 2016 Conference for the Korea Institute of Military Science and Technology.*

Newmann, M. E. (2002). Assortative mixing in networks, *Physical Review Letters*, **89**, 208701.

Newmann, M. E. (2004). Analysis of weighted networks, *Physical Review E*, **70**, 056131.

Newmann, M. E. (2006). Modularity and Community Structure in Networks. *In Proceedings of the National Academy of Sciences of the United States of America*, **103**, 8577.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L., and Stanley, H. (2000). Econophysics: financial time

series from a statistical physics point of view, *Physica A: Statistical Mechanics and its Applications*, **279**, 443–456.

Podobnik, B. and Stanley, H. (2008). Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series, *Physical Review Letters*, **100**, 084102.

Podobnik, B., Jiang, Z., Jhou, W., and Stanley, H. (2011). Statistical tests for power-law cross-correlated processes, *Physical Review E*, **84**, 066118.

Sohn, K., Lee, E., and Lee, J. (2008). Detection and forecast of climate change dignal over the Korean peninsula, *The Korean Journal of Applied Statistics*, **21**, 705–716.

Wang, C., Chan, C., and Ho, K. (1989). Empirical tight-binding force model for molecular-dynamics simulation of Si, *Physical Review B*, **39**, 8586.

Yook, S., Oltvai, Z., and Barabasi, A. (2004). Functional and topological characterization of protein interaction networks, *Proteomics*, **4**, 928–942.

DCCA 방법으로 연결된 한반도의 기온 네트워크 분석

민승식[a,1]

[a]해군사관학교 이학과

(2016년 11월 14일 접수, 2016년 12월 17일 수정, 2016년 12월 23일 채택)

요 약

본 논문에서는 1976년부터 2015년까지 40년 간, 59개 지역 기온 시계열을 대상으로 degrended cross-correlation analysis(DCCA) 방법을 이용한 상관 계수를 도출하였다. 4년 단위의 평균기온, 최고기온, 최저기온 시계열을 분석하여 상관계수 값이 0.9 이상이면 단위 기간 동안 두 지역의 온도 상관성이 존재하는 것으로 판단하고, 두 지역 간의 연결선을 만드는 방식으로 네트워크를 구축하였다. 이후 네트워크 이론을 바탕으로 평균 경로 길이, 결집 계수, 유사성, 모듈성 등의 값들을 도출하였다. 그 결과, 기온 네트워크는 좁은 세상 성질을 만족하고, 유사성과 모듈성이 높은 네트워크임을 알 수 있었다.

주요용어: 기온, detrended cross-correlation analysis (DCCA), 기온 네트워크, 평균 경로 길이, 결집 계수, 유사성, 모듈성, 좁은 세상

[1](51704) 경상남도 창원시 진해구 중원로 1, 해군사관학교 이학과. E-mail: fieldsmin@gmail.com