희박 벡터자기상관회귀 모형을 이용한 한국의 미세먼지 분석[†]

이원석¹, 백창룡²

¹²성균관대학교 통계학과

접수 2014년 6월 16일, 수정 2014년 7월 3일, 게재확정 2014년 7월 8일

요약

본 논문은 최근 많은 관심을 받는 미세먼지 (PM10)의 일별 평균농도에 대해서 전국 16개 시도에서 2008년부터 2011년까지 관측한 다변량 시계열 자료에 대한 연구이다. 다변량 시계열 모형을 이용해서 시간 및 공간에 대한 상관관계를 동시에 고려, 일변량 혹은 특정 지역에 국한해서 분석한 기존의 연구와 차별성을 두었다. 또한 Davis 등 (2013)이 제안한 부분 스펙트럼 일관성 (partial spectral coherence)을 통해 다른 지역간의 상호 의존성을 파악하고 이를 토대로 변수 선택을 통해 희박벡터자기회귀모형 (sVAR; sparse vector autoregressive model)을 적합하는 방법론을 적용하여 고차원 자료 분석의 단점 및 한계를 보완하였으며 예측력 비교를 통해서 sVAR 모형 적합의 타당성을 검증하였다.

주요용어: 미세먼지, 부분 스펙트럼 일관성, 장기간 종속 시계열, 희박벡터자기회귀모형, 희박성.

1. 서론

최근 들어 대기오염 예보제가 시행되며 미세먼지에 대한 관심이 높아졌다. 미세먼지 (PM10; particulate matter less than 10 μm)란 대기 중에 직경이 10 μm 이하의 크기로 존재하는 매우 작은 입자를 뜻한다. 미세먼지는 다양한 원인에 의해서 생성된다. 크게 자연적인 미세먼지와 인간의 활동에 의해서 생기는 미세먼지로 구분할 수 있다. 자연적인 원인으로는 산불, 화산재, 먼지 등을 예로 들 수 있고, 인간의 활동에 의한 원인으로는 화석연료의 사용, 자동차 배기가스, 화력 발전소 및 산업 공장에서 발생하는 공해 등을 꼽을 수 있다. 특히 직경이 2.5 μm 이하인 초미세 먼지 (PM2.5)의 경우 그 크기가 너무 작아 우리 신체에서 자연적으로 걸러내지 못하고 폐와 장기 깊숙히 침투하여 폐암, 천식, 호흡기 질환 및 순환계 질환을 일으키며 심지어 불임의 원인이 되기도 한다. 이러한 심각성을 깨닫고 우리 나라도 전국에 걸쳐 대기 오염 물질 측정소를 설치하여 운영하고 있으며 미세먼지의 경우 연평균 50 $\mu m/m^3$, 일평균 100 $\mu m/m^3$ 이하로 규제하고 있다. 또한 초미세 먼지의 경우 시간당 60 $\mu m/m^3$ 이상이 두 시간 이상 관측될 경우 초미세 먼지 예보를 발령하는 등 보다 엄격한 관리를 시행 확대 중에 있다.

본 연구에서는 인간의 삶의 질에 밀접한 영향을 미치는 미세먼지를 우리 나라 16개 대도시에서 관측된 일별자료를 토대로 분석하고자 한다. 기존의 미세먼지에 대한 연구는 대부분 우리 나라의 전체 평균 혹은 특정 지역에 국한된 일변량 시계열 분석이었다 (Kim, 2006; Pan과 Chen, 2008; Lee, 2010). Baek (2013)에서는 2011년 시간별 자료 분석을 통해서 PM10의 주변분포 (marginal distribution)는

[†] 이 논문은 2014년도 정부 (미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2014R1A1A1006025).

¹ (110-745) 서울특별시 종로구 성륜관로 25-2, 성균관대학교 통계학과, 대학원생.

² 교신저자: (110-745) 서울특별시 종로구 성륜관로 25-2, 성균관대학교 통계학과, 조교수.

E-mail: crbaek@skku.edu

두꺼운 꼬리를 가지고 있으며 매우 높은 시차에서도 매우 강한 종속성을 가짐을 밝혔다. 이러한 강한 종속성은 장기간 종속 시계열 (long range dependence; LRD)라고 불리우는 시계열의 대표적인 특징이다 (Beran, 1994; Doukhan 등, 2003; Palma, 2007). 하지만, 이러한 강한 종속성은 우리나라 전체의 미세먼지 수준이 여름철에 다량의 강수에 의해서 낮게 나타나는 특성, 즉 평균변화 때문에 매우 긴 종속관계를 가지는 LRD 시계열처럼 보이는 가짜효과임을 밝혔다. 이번 연구는 우리 나라의 규제의 기준이 되는 일별 자료를 사용하여 보다 거시적인 분석에 초점을 두었으며 각 지역별 미세먼지의 차이, 특징, 상호 상관관계 (cross-correlation)에 좀 더 중점을 둔 연구이다.

통계적인 방법론에 있어서도 본 연구는 Davis 등 (2013)이 제안한 방법에 따라 부분 스펙트럼 일관성 (partial spectral coherence)을 통하여 변수간의 상관관계의 정도를 파악하였다. 그리고 차원이 16인 다변량 시계열 자료의 분석에 있어서 효율적인 변수 선택을 위하여 희박 벡터자기회귀모형 (sVAR; sparse vector autoregressive model)을 사용하였다. 그리하여 다변량 시계열을 다룰 때 흔히 쓰이는 벡터자기회귀모형 (Kim 등, 2012)의 단점인 차원의 복잡성을 피하고 모형의 정확성을 높였다.

본 논문의 구성은 다음과 같다. 먼저 우리는 본 논문의 핵심적 방법론인 희박 벡터자기회귀모형에 대해서 2절에서 살펴본다. 그리고 2008년부터 2011년까지 전국 16 개 시도에서 관측된 일별 평균 자료에 대한 사전분석을 3절에서 보고하여 우리나라 미세먼지의 전반적인 특징에 대해서 먼저 살펴본다. 희박 벡터자기회귀모형 분석에 대한 결과 보고 및 예측값 비교를 통해서 희박 벡터자기회귀모형에 대한 우수성을 4절에서 다루며 5절에서 결과에 대한 토의를 한다.

2. 희박 벡터자기상관회귀 모형

차원이 K 인 다변량 시계열 자료 $Y_1, \dots, Y_n$ 이 있다고 가정하자. 이에 대해서 차수 p 를 갖는 VAR(p) 모형은

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \dots + A_p Y_{t-p} + Z_t, \quad t = 1, \dots, T$$

으로 주어진다. 여기서 Z_t 는 평균이 0이고 분산-공분산 행렬이 Σ 인 K 차원의 *i.i.d.* 정규분포를 따르는 확률변수열이다. 또한 $\{Y_t\}$ 가 인과과정 (causal process)임을 가정하여 Z_t 가 $\{Y_s, s < t\}$ 와 독립임을 가정한다. $A_1, \dots, A_p$ 는 크기가 $K \times K$ 인 실수 행렬들로 AR계수이다. 따라서 매개변수인 $A_1, \dots, A_p$ 를 추정하기 위해서는 $K^2 p$ 개의 모수에 대한 추정이 필요하다. 하지만 차원 (K)이 커질수록 추정해야 하는 매개변수의 개수는 기하급수적으로 증가하여 비단 모형이 복잡해질 뿐만 아니라 추정에 있어서도 매우 불안하게 되며, 이는 곧 시계열 자료의 궁극적인 목적인 예측력을 현저히 저하시킨다. 따라서 절약성의 원리에 따라서 간결한 모형을 선호하게 된다. 하지만 어떠한 변수를 제거할 지, 정확하게는 어디서 0 계수 (zero coefficient)를 가지게 하여 불필요한 정보를 없앨지에 대해서는 자명하지 않다.

효율적인 변수선택을 위해서 Davis 등 (2013)은 부분 스펙트럼 일관성에 기반한 변수 선택법을 제안하였으며 다음과 같이 요약된다. 관측벡터 Y_t 의 i 번째 원소를 $Y_{t,i}$, j 번째 원소를 $Y_{t,j}$ 라 쓰고 i, j 번째 원소를 제거한 나머지 벡터를 $Y_{t,-ij}$ 라고 하자. 먼저 $Y_{t,i}$ 와 $Y_{t,j}$ 의 조건부 상관관계를 구하기 위해서 $Y_{t,-ij}$ 의 선형 효과를 제거해 주어야 한다. $Y_{t,i}$ 에서 $Y_{t,-ij}$ 의 효과를 제거한 최적의 선형 결합은 다음과 같이 구할 수 있다.

$$(D_{k,i}^{opt} \in \mathbb{R}^{K-2}, k \in \mathbb{Z}) = \underset{(D_{k,i}, k \in \mathbb{Z})}{\operatorname{argmin}} \mathbb{E} \left(Y_{t,i} - \sum_{k=-\infty}^{\infty} D_{k,i} Y_{t-k,-ij} \right)^2.$$

이를 통해 얻어진 잔차는

$$\epsilon_{t,i} := Y_{t,i} - \sum_{k=-\infty}^{\infty} D_{k,i}^{opt} Y_{t-k,-ij}$$

으로 주어지며 비슷한 방법을 통해서 $\epsilon_{t,j}$ 를 구할 수 있다. 따라서 $Y_{t,-ij}$ 의 효과가 제거된 $Y_{t,i}$ 와 $Y_{t,j}$ 가 조건부 무상관이 (conditionally uncorrelated) 되어 있다면 $\epsilon_{t,i}$ 와 $\epsilon_{t,j}$ 도 모든 시차에서 무상관성을 지닌다. 즉 모든 정수 k 에 대하여 $\gamma_{ij}^{\epsilon}(k) = \text{Cov}(\epsilon_{t+k,i}, \epsilon_{t,j}) = 0$ 이다. 이는 진동수 영역 (frequency domain)으로 바꾸어서 생각하면 두 잔차의 교차 스펙트럼 밀도함수

$$f_{ij}^{\epsilon}(\omega) := \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_{ij}^{\epsilon}(k) e^{-ik\omega}, \quad \omega \in (-\pi, \pi], \quad (2.1)$$

가 모든 주파수 ω 에서 0의 값을 가짐을 의미한다.

교차 스펙트럼 밀도함수 $f_{ij}^{\epsilon}(\omega)$ 은 $Y_{t,-ij}$ 가 주어졌을 때, $Y_{t,i}$ 와 $Y_{t,j}$ 와의 조건부 상관을 의미하며 이를 스케일 조정한 것이 바로 부분 스펙트럼 일관성 (PSC; partial spectral coherence) (Brillinger, 1981)으로 다음과 같이 주어진다.

$$\text{PSC}_{ij}(\omega) := \frac{f_{ij}^{\epsilon}(\omega)}{\sqrt{f_{ii}^{\epsilon}(\omega) f_{jj}^{\epsilon}(\omega)}}, \quad \omega \in (-\pi, \pi]. \quad (2.2)$$

좀 더 손쉬운 계산을 위해서 Brillinger (1981)는 $f_{ij}^{\epsilon}(\omega)$ 를 Y_t 의 스펙트럼 밀도 (spectral density) $f^Y(\omega)$ 을 이용하여 다음과 같이

$$f_{ij}^{\epsilon}(\omega) = f_{ii}^Y(\omega) - f_{i,-ij}^Y(\omega) f_{-ij,-ij}^Y(\omega)^{-1} f_{-ij,j}^Y(\omega) \quad (2.3)$$

계산할 수 있음을 보였다. 하지만 (2.3)식에 포함된 $f_{-ij,-ij}^Y(\omega)^{-1}$ 로 인해 $(K-2) \times (K-2)$ 차원의 역행렬 계산이 필요하여 $\binom{K}{2}$ 번의 역행렬 계산을 하여야하는 불편함이 여전히 존재한다. Dahlhaus (2000)은 다음의 보다 효율적인 방식을 제안하였다.

$$\text{PSC}_{ij}(\omega) = -\frac{g_{ij}^Y(\omega)}{\sqrt{g_{ii}^Y(\omega) g_{jj}^Y(\omega)}}, \quad \omega \in (-\pi, \pi].$$

여기서 $g^Y(\omega)$ 은 $f^Y(\omega)$ 의 역함수이며, $g_{ii}^Y(\omega)$, $g_{jj}^Y(\omega)$, $g_{ij}^Y(\omega)$ 은 각각 $g^Y(\omega)$ 의 i 번째, j 번째의 대각원소 그리고 (i, j) 번째 원소를 의미한다. 이로써 우리는 크기 $K \times K$ 의 역행렬을 한 번만 계산하여 PSC를 얻을 수 있다.

따라서 $Y_{t,i}$ 와 $Y_{t,j}$ ($i \neq j$)가 조건부 비상관성을 가짐은 $g_{ij}^Y(\omega) = 0$ 이 모든 주파수(ω)에서 0을 가짐과 동치인 명제가 된다. 이는 곧 PSC값을 계산함으로써 $Y_{t,i}$ 와 $Y_{t,j}$ 와의 상관관계를 구할 수 있음을 의미한다. 이를 바탕으로 Davis 등 (2013)은 sVAR 모형 적합을 위해 다음의 방법을 제안하였다.

(1단계) 앞서 설명했듯이, PSC값이 0을 가지는 것은 두 변수간에 조건부 무상관성을 의미하므로 이를 이용하여 조건부 무상관성을 가지는 쌍에 대해서 VAR 모형의 AR 계수를 0으로 둔다. 즉 $Y_{t,i}$ 와 $Y_{t,j}$ 의 PSC값이 0이라면

$$A_k(i, j) = A_k(j, i) = 0, \quad i \neq j, \quad k = 1, \dots, p. \quad (2.4)$$

하지만 실측 관측값에 대해서는 $\hat{\text{PSC}}_{ij}(\omega)$ 가 정확하게 0의 값을 주지 않을 수도 있기에 Dahlhaus (2000)에서 제안한 다음의 방법을 통하여 VAR 모형의 0이 아닌 (i, j) 번째 계수 (nonzero coefficient)를 선택한다. 요약 통계량으로

$$\hat{S}_{ij} := \sup_{\omega} |\hat{\text{PSC}}_{ij}(\omega)|^2, \quad \omega = \frac{2\pi k}{T}, \quad (k = 1, \dots, T) \quad (2.5)$$

를 정의하여 총 $\binom{K}{2}$ 개의 $\hat{S}_{ij}$ 를 구한다. 위 통계량이 큰 값을 가진다는 것은 두 변수가 서로 조건부 상관성을 지님을 의미한다. 따라서 큰 값에서 부터 작은 값으로 정렬한 다음에 순차적으로 AR계수를 0이 아닌 계수로 설정한 후, BIC를 통해서 VAR 모형의 차수 (p)와 0이 아닌 계수의 개수 (M)을 선택하게 된다. BIC는 다음과 같이 계산된다.

$$\text{BIC}(p, M) = -2 \log L(\hat{A}_1, \dots, \hat{A}_p) + \log T \times (K + 2M)p,$$

여기서 $L(\hat{A}_1, \dots, \hat{A}_p)$ 은 $\text{VAR}(p)$ 모형의 최대우도함수이다. p 와 M 의 범위를 미리 정해준 후, BIC값을 계산하며 이를 통해 $(K + 2\hat{M})\hat{p}$ 개의 0이 아닌 AR 계수들을 얻는다.

(2단계) 하지만 1단계를 거친 AR 계수들 중에서도 여전히 0의 값을 가질 수 있는 계수들이 존재한다. 왜냐하면 PSC는 오직 $(i, j), i \neq j$ 인 자료의 쌍들만 계산되기 때문에 AR 계수의 대각 원소에 대해서는 누락되며, 일반적인 VAR 모형에서 나타나듯이 (i, j) 번째 원소와 (j, i) 번째 원소들이 같은 값을 가지지 않아도 되기 때문이다. 따라서 (1단계)에서 구한 $(K + 2\hat{M})\hat{p}$ 개의 0이 아닌 AR 계수들에 대해서 t -통계량의 절대값을 구하고 이를 통하여 다시 한번 전진 변수 선택 (forward selection)을 실시한다. AR 계수들의 t -통계량은 다음을 통해 구하며

$$t_{i,j,k} := \frac{\hat{A}_k(i, j)}{\text{S.E.}(\hat{A}_k(i, j))}, \quad (2.6)$$

여기서 $\hat{A}_k(i, j)$ 의 표준편차는 다음의 극한 이론을 사용하여 구한다

$$\sqrt{T}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}\left(0, \tilde{R}[\tilde{R}'(\tilde{\Gamma}_Y(0) \otimes \tilde{\Sigma}_Z^{-1})\tilde{R}]^{-1}\tilde{R}'\right).$$

여기서 $\alpha := \text{vec}(A_1, \dots, A_p)$ 이며 $\hat{\alpha}$, $\tilde{\Gamma}_Y(0)$, $\tilde{\Sigma}_Z$ 는 각각 α , $\text{Cov}((Y_t', \dots, Y_{t-p+1}')')$ 와 Σ_Z 의 최대우도추정량이다. $\tilde{R}$ 은 제약행렬로 Remark 2.1에서 추가 설명하였다.

요약하자면 t -통계량 (2.6)의 절대값이 큰 AR계수에서 부터 작은 계수까지 정렬을 한 뒤에 이를 바탕으로 변수를 하나씩 추가한 뒤 다시한번 $\text{BIC} = -2 \log L + \log Tm$ 를 계산하고 BIC를 가장 작게 만드는 모형을 가장 좋은 모형으로 선택한다. 최종적으로 선택한 모형은 $\text{sVAR}(p^*, m^*)$ 이며 여기에서 p^* 는 AR의 차수이고 m^* 은 영이 아닌 AR 계수들의 개수이다.

Remark 2.1 (2단계)에서 사용한 제약행렬의 예제는 다음과 같다. 간단하게 2차원의 $\text{VAR}(1)$ 모형을 생각하고 $(1, 2)$, $(2, 2)$ 번째 원소가 0 이라고 가정하자. 그러면 모형은 다음과 같다

$$\begin{pmatrix} Y_{t,1} \\ Y_{t,2} \end{pmatrix} = \begin{pmatrix} A_{1,1} & 0 \\ A_{2,1} & 0 \end{pmatrix} \begin{pmatrix} Y_{t-1,1} \\ Y_{t-1,2} \end{pmatrix} + \begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \end{pmatrix}.$$

이에 대해서 vec 연산을 통하면 R 행렬은 다음과 같이 주어진다

$$\alpha := \begin{pmatrix} A_{1,1} \\ A_{2,1} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} A_{1,1} \\ A_{2,1} \\ A_{1,2} \\ A_{2,2} \end{pmatrix} =: R\gamma.$$

3. 미세먼지 사전분석

본 절에서는 대기 오염도를 측정하는 중요한 지표중에 하나인 미세먼지 (PM10)를 2008년 부터 2011년까지 관측한 자료의 사전분석을 보고한다. 사용한 자료는 국립환경과학원 확정자료로 전국 16개 시도 지역 (서울, 부산, 대구, 인천, 광주, 대전, 울산, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주)의 일평균을 분석하였다. 총 1461개의 자료를 분석하였으며 43개의 결측치에 대해서는 선형보간 (linear interpolation)을 적용하였다.

Figure 3.1 Time plot of daily average of PM10 in Korea from 2008-2011

Figure 3.2 SACF of daily average of PM10 in Korea from 2008-2011

Table 3.1 Summary statistics of PM10 in Korea from 2008-2011

	Q25	Q50	Q75	Mean	Std		Q25	Q50	Q75	Mean	Std
서울	32	46	62	52	33	강원	33	48	66	53	31
부산	33	43	59	49	28	충북	39	54	74	61	33
광주	28	40	56	46	32	충남	29	41	57	47	29
대전	27	39	54	45	27	전북	35	46	62	52	27
인천	38	52	68	58	32	전남	29	38	51	44	27
울산	32	44	62	50	29	경북	34	44	58	50	26
대구	33	44	61	51	30	경남	30	41	56	47	27
경기	39	54	70	59	33	제주	26	36	51	44	37

Figure 3.1은 서울, 부산, 광주 및 대전지역에 대한 시계열 (time plot)를 나타낸다. 간결한 요약에 위하여 4개 대도시를 선택하였으며 Table 4.2에 나와 있듯이 다른 지역 역시 비슷한 패턴을 보이고 있다. 먼저 대부분의 미세먼지 농도는 우리 나라의 허용기준치인 $100\mu\text{m}/\text{m}^3$ 이하에 머무르고 있으나 몇몇 관측시점에서는 최고 $500\mu\text{m}/\text{m}^3$ 로 매우 큰 요철 (spike)이 관측되고 있다. 이러한 점은 봄철 황사의 영향으로 생긴 이상치 (outlier)라고 판단된다.

Figure 3.2는 위 4개 도시 (서울, 부산, 광주, 대전)에 대한 시계열 자료의 표본자기상관도 (sample autocorrelation plot; SACF)를 나타낸다. 차수 (lag)가 60에 이르면 매우 천천히 감소하는 표본자기상관계수를 확인해 볼 수 있으며 약간의 불규칙한 계절성도 눈에 띈다. 이렇게 천천히 감소하는 SACF는 장기간종속시계열 (long range dependence; LRD) 모형으로 효과적으로 분석할 수 있다. 하지만, Baek (2013)에서 밝혔듯이 장기간종속시계열은 유한 표본의 경우 단순한 평균변화모형 (changes-in-mean model)과 매우 쉽게 혼동되므로 주의가 필요하며 국소로컬 (LW; local Whittle) 방법론에 기반하여 두 모형을 구분 짓는 통계방법론을 통해 매시간 미세먼지 자료에 적용할 경우 평균변화에 기반한 가짜 효과임을 밝혔다. 따라서 이번 분석에서도 이러한 매우 강한 상관관계를 설명하기 위해서 Baek (2013)에서 사용한 가설검정법을 적용하였으며, 그 결과 강한 상관관계는 장기간종속성보다

는 평균변화에 따른 변화에 기인한 것으로 보인다. 예를 들어 Figure 3.3은 강원지역의 일평균 미세먼지의 시도표와 LW 방법에 기반하여 평균변화모형을 추정한 그림이다. 같은 그림의 아래 패널에서 살펴 보듯이 SACF 및 SPACF 모두 빠르게 감소하고 있다. 또한 추정된 평균변화모형의 경우 여름철에 낮은 미세먼도를 보임으로서, 여름에 한국에 집중된 강수와의해서 대기질이 개선되었음을 의미한다. 따라서 본 연구에서는 각 지역에 대해서 평균변화모형을 LW 방법에 근거하여 추세를 제거한 뒤 자료 분석을 실시하였다.

Figure 3.3 Estimated changes-in-mean by LW estimation; Residual series are no longer strongly correlated

Figure 3.4 Sample cross-correlation plot between Seoul and other areas; All are positively correlated

또한, 이번 연구는 한국의 미세먼지에 대한 일변량 시계열 자료 분석을 확장하여 전국 16개 시도에 대한 종합적인 모형을 세우는데 있다. 예를 들어 Figure 3.4는 서울 지역과 인천, 부산, 광주 그리고 제주

간의 교차상관도를 나타낸다. 지역의 거리에 상관없이 모두 유의한 양의 상관관계를 가지고 있음을 살펴볼 수 있다. 따라서 한국의 미세먼지에 대해서 더 잘 이해하기 위해서는 이러한 교차상관을 고려한 벡터자기회귀모형 (VAR)이 적절할 것으로 보인다.

4. 회박 벡터자기상관회귀 모형을 이용한 미세먼지 실증 분석

하지만, 데이터의 차원이 $K = 16$ 이므로 일반적인 VAR 모형을 사용하기에는 추정해야 하는 변수가 너무 많다. 예를 들어 VAR(1) 모형을 적합할 경우 $16 \times 16 = 256$ 개의 모수를 실제 데이터의 개수인 1461개로 추정하여야 하는 어려움이 존재한다. 따라서 추정해야 하는 변수의 개수를 효과적으로 제거해주는 sVAR 모형이 매우 적합하다. 2절에서 소개한 방법에 따라 sVAR 모형을 적합한 결과는 다음과 같다. 먼저 $\binom{16}{2} = 120$ 개의 순서쌍에 대한 PSC 값의 최대값인 $\hat{S}_{ij}$ 은 Figure 4.1과 같다. 원의 크기가 클수록 강한 상관관계를 의미한다. 예를 들어 (2,7)-(부산, 울산), (2,15)-(부산, 경남), (3,7)-(대구, 울산), (8, 15)-(경기, 경남), (13,15)-(전남, 경남)이 매우 강한 상관관계를 보임을 알 수 있다. 일차적으로는 지역적인 거리의 가까움이 가장 큰 상관관계를 줌을 알 수 있다. 하지만 중간 크기의 원을 가지는 (2,5)-(부산, 광주), (4,16)-(인천, 제주)의 경우를 통해서 거리의 가까움 만이 종속관계를 다 설명해주지는 못함을 알 수 있다.

Figure 4.1 PSC between variables

Figure 4.2 BIC selection and estimated coefficients

2절에서 소개된 sVAR 모델을 적합한 결과 최종 선택된 모형은 sVAR(1, 152)이다. 이는 VAR(1) 모형을 적합하였을 때 추정해야 하는 변수의 개수 256개의 변수에 비해서 60%의 계수만 사용한 모형이다. Figure 4.2의 왼쪽 패널은 차수 $p = 1, \dots, 4$ 에 대해서 사용된 변수의 개수에 따른 BIC그림을 나타내며 오른쪽은 최종 선택된 모형 sVAR(1, 152)에 대해서 추정된 계수를 나타낸다. 먼저 sVAR에서의 추정된 계수와 Figure 4.1이 유사함을 살펴 볼 수 있다. 또한 대부분의 대각원소들이 양의 상관관계를 가짐을 통해서 PM10의 농도가 같은 지역의 하루 전 농도와 가장 밀접하게 연관되어 있음을 알 수 있다. 예를 들어 첫번째 행인 서울의 경우

$$Y_{t,1} = .71Y_{t-1,1} - .06Y_{t-1,2} - .21Y_{t-1,4} + .19Y_{t-1,5} - .05Y_{t-1,6} + .05Y_{t-1,7} - .07Y_{t-1,14} + z_{t,1}$$

으로 주어지 같은 지역의 하루 전 농도와 가장 큰 상관관계를 보이고 있으며 VAR 모형의 성질에 따라 과거의 차수에 대해서는 그 상관관계가 기하급수적으로 감소함을 알 수 있다. Davis 등 (2013)의 (2 단계)에서 지적하였듯이 AR 계수가 대칭적이지는 않는다. 5, 6, 7, 8월 (각각 광주, 대전, 울산, 경기)의 경우 다른 지역과 양의 상관관계를 가짐에 비해서, 4, 12, 14월 (인천, 전북, 경북)의 경우 음의 상관관계를 가진다. LW 방법에 기반하여 평균 변화 추세를 제거한 뒤에 얻어진 잔차에 대한 sVAR 모형이기 때문에 해석이 용이하지는 않지만 지리학적 설명에 대한 추후 연구가 필요할 것으로 보인다.

마지막으로 시계열 분석에서 중요한 예측력을 다음의 시험자료에 대해서 h -시차 예측을 통해서 비교하였다. 먼저 마지막 $T_{test} = 24$ 개의 시험자료 (test data)에 대해서 t -시점 까지의 자료를 이용해서 h -시차 이후의 값을 예측한 $\hat{Y}_{t+h,k}$ 을 계산한 뒤에 제곱근평균제곱오차 (RMSE; root mean squared error)를 다음과 같이 계산한다.

$$RMSE(h) = \left\{ \frac{1}{K(T_{test} - h + 1)} \sum_{k=1}^K \sum_{t=T}^{T+T_{test}-h} \left(\hat{Y}_{t+h,k} - Y_{t+h,k} \right)^2 \right\}^{1/2}.$$

Table 4.1은 sVAR(1, 152) 및 VAR(1, 256)을 각각 이용하여 h -시차 예측을 할 경우에 얻어진 제곱근 평균제곱오차를 비교한 표이다. 이 표에서 고려한 모든 시차 $h = 1, 2, 3, 4$ 에 대해서 sVAR(1, 152) 모형이 더 좋은 예측력을 보여주어 sVAR 모형이 간결할 뿐만 아니라 예측력도 향상시킴을 알 수 있다.

Table 4.1 The h -step ahead RMSE comparison

	$h = 1$	$h = 2$	$h = 3$	$h = 4$
sVAR(1, 152)	249.75	448.93	532.44	566.81
VAR(1)	256.02	458.24	540.26	567.70

5. 논의

이 논문에서는 한국의 미세먼지의 일별 평균농도에 대해서 2008년부터 2011년까지 전국 16개 시도에서 관측된 다변량 시계열 자료에 대한 분석을 실시하였다. 그 결과 일별 평균 농도에 대해서도 장기간 종속성 (LRD)을 보였다. 하지만 이는 여름철에 집중된 강수에 의해 미세먼지 농도가 감소하는 평균 변화 추세 때문에 나타나는 가짜효과임을 통계적 가설 검증을 통해 밝혔다. 이러한 추세를 제거한다 할지라도 상관관계는 남아있었으며 교차상관 (cross correlation)의 경우 많은 부분이 양의 상관관계를 보였다. 고차원의 다변량 시계열을 보다 효과적으로 적합하기 위해서 Davis 등 (2013)이 제안한 PSC에 기반한 sVAR 모형을 적합시켰다. 그 결과 VAR 모형보다 더 간결하면서도 예측력은 좋은 결과를 주었다. 본 연구는 일변량 혹은 특정 지역에 국한된 미세먼지에 대한 분석이 아니라 시간적 공간적 종속관계를 동시에 고려한 다변량 모형을 사용하여 분석하였다는데 큰 의의를 두며 최적의 모형에 대한 지리학, 기상학, 환경학적 의미에 대해서는 추가 연구가 필요할 것으로 보인다.

References

- Baek, C. (2013). Time series modelling of air quality in Korea: Long range dependence or changes in mean? *The Korean Journal of Applied Statistics*, **26**, 987–998.
- Beran, J. (1994). *Statistics for long-memory processes*, Vol. 61 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, New York.
- Brillinger, D. R. (1981). *Time series: Data analysis and theory*, Society for Industrial and Applied Mathematics, Philadelphia, PA.
- Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, **51**, 157–172.
- Davis, R. A., Zang, P. and Zheng, T. (2013). Sparse vector autoregressive modeling. Preprint.
- Doukhan, P., Oppenheim, G. and Taqqu, M. S. (2003). *Theory and applications of long-range dependence*, Birkhäuser Boston Inc., Boston, MA.
- Kim, J., Jin D., Lee J., Kim S. and Son Y. (2012). Prediction of the interest spread using VAR model. *Journal of the Korean Data & Information Science Society*, **23**, 1093–1102.
- Kim, Y. P. (2006). Air pollution in Seoul caused by aerosols. *Journal of Korean Society for Atmospheric Environment*, **22**, 535–553.
- Lee, H. (2010). Analysis of time series models for PM10 concentrations at the Suwon city in Korea. *Journal of the Korean Data & Information Science Society*, **21**, 1117–1124.
- Palma, W. (2007). *Long-memory time series*, John Wiley & Sons, Inc., Hoboken, New Jersey, USA.
- Pan, J.-N. and Chen, S.-T. (2008). Monitoring long-memory air quality data using arfima model. *Environmetrics*, **19**, 209–219.

The sparse vector autoregressive model for PM10 in Korea[†]

Wonseok Lee¹ · Changryong Baek²

¹²Department of Statistics, Sungkyunkwan University

Received 16 June 2014, revised 3 July 2014, accepted 8 July 2014

Abstract

This paper considers multivariate time series modelling of PM10 data in Korea collected from 2008 to 2011. We consider both temporal and spatial dependencies of PM10 by applying the sparse vector autoregressive (sVAR) modelling proposed by Davis *et al.* (2013). It utilizes the partial spectral coherence to measure cross correlation between different regions, in turn provides the sparsity in the model while balancing the parsimony of model and the goodness of fit. It is also shown that sVAR performs better than usual vector autoregressive model (VAR) in forecasting.

Keywords: Long range dependence, partial spectral coherence, PM10, sparse vector autoregressive model, sparsity.

[†] This research was supported by the Basic Science Research Program from the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1006025).

¹ Graduate student, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea.

² Corresponding author: Assistant professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea. E-mail: crbaek@skku.edu.