서울시 미세먼지 네트워크를 활용한 공기 정화탑의 적정 위치 선정

박범진*, 문상준**, 채상희***

< 요 약 >

최근 중국 과학원 지구환경연구소의 연구결과에 따르면 공기 정화탑은 주변의 미세먼지 농도를 감소시키는 효과를 가지고 있다. 우리나라도 미세먼지 농도를 줄일 수 있는 방안을 마련하는 것이 시급해짐에 따라 서울시도 공기 정화탑 건설을 검토중인 것으로 알려졌다. 하지만 서울시내에 공기 정화탑의 적절한 위치를 선정하는 것은 상당히 어려운 문제이다. 본 연구에서는 허브 그래피컬 라쏘(Hub graphical lasso) 모형을 이용하여 서울시내 지역구별 미세먼지 농도간의 네트워크를 추정하고 네트워크의 연결이 집중되는 중심지를 공기 정화탑의 적정 위치라 제안하였다. 서울시 미세먼지 네트워크를 추정한 결과, PM_{10}과 $PM_{2.5}$ 미세먼지 농도간의 네트워크 연결이 집중되는 중심지는 광진구, 중구 그리고 양천구로 나타났다. 따라서 서울시내의 공기 정화탑이 광진구, 중구, 양천구에 건설된다면 공기 정화탑 주변 지역뿐만 아니라 네트워크가 연결된 광범위한 지역의 미세먼지 농도까지 줄이는 효과가 있을 것이라 판단된다.

주제어 : 그래피컬 모델, 네트워크, 미세먼지, 그래피컬 라쏘, 허브 그래피컬 라쏘

* (02504) 서울특별시 동대문구 서울시립대로 163 (전농동), 서울시립대학교 통계학과, 박사과정
** (02504) 서울특별시 동대문구 서울시립대로 163 (전농동), 서울시립대학교 통계학과, 박사과정
*** (02504) 서울특별시 동대문구 서울시립대로 163 (전농동), 서울시립대학교 통계학과, 석사과정

Ⅰ. 서 론

최근 여러 연구 결과에 의하면 미세먼지는 호흡기 및 심혈관계 질환을 유발하고 사망률과도 연관되어 있는 것으로 보고되고 있다(신동천, 2007). 우리나라도 2003년 '수도권 대기환경 개선에 관한 특별법'이 재정되면서 수도권 지역을 중심으로 미세먼지 농도를 줄이고자 하는 정책들이 수립되었다. 하지만 2017년 경제협력개발기구(OECD)의 '삶의 질' 보고서에 따르면 우리나라의 초미세먼지 평균 노출도는 41개국 중 가장 나쁜 것으로 보고되고 있어 아직까지도 미세먼지로 인해 위협을 받고 있고 미세먼지 절감을 위한 방안 마련이 시급한 상황이다.

세계적으로 악명 높은 수준의 미세먼지가 발생하는 중국은 미세먼지 절감 방안으로 세계 최대 규모의 공기 정화탑을 건설하였다. 중국 서부 시안에 세워진 높이 100m가 넘는 공기정화용 탑이 실제 효과가 있는지에 대해 의구심이 있었지만 중국과학원 지구환경연구소에 따르면 "탑 주변 12곳의 공기 질 측정소에서 효과를 살핀 결과 $10km^2$ 지역에 매일 1000만m^3의 깨끗한 공기가 생산됐으며, 특히 대기 오염이 심각한 날 초미세먼지($PM_{2.5}$) 평균 농도가 15% 줄었다"고 한다(주간조선, 2018).

최근 서울시도 중국의 공기 정화탑을 방문해 서울시에 공기 정화탑을 건설할 수 있는지에 대해 검토할 계획이라고 알려져 있다. 서울시에도 중국과 같은 규모의 공기 정화탑을 건설한다면 "어떤 위치에 공기 정화탑을 건설해야 가장 효과적일까?"라는 문제에 대한 논의가 필요하다. 단순하게 공기 정화탑의 적절한 위치를 생각한다면 연평균 미세먼지의 농도가 높은 곳이나 "미세먼지 나쁨" 이상의 농도가 자주 발생하는 곳으로 정할 수 있다. 그러나 공기 정화탑을 건설한 뒤에 구조물이 위치한 좁은 지역구에만 국소적으로 효과를 보인다면 많은 지역에 공기 정화탑을 건설해야하기 때문에 건설비용 측면에서 비효율적이다. 예를 들어 서울역의 미세먼지 농도가 주변 많은 지역구의 미세먼지 농도와 연관성이 강하다면 서울역에 공기 정화탑이 건설됨으로 인해 인근 지역구인 용산구, 중구뿐만 아니라 멀리는 강남구까지의 미세먼지 농도도 낮아질 수 있다. 반면, 영등포구의 연평균 미세먼지 농도는 높지만 다른 지역구들과 미세먼지 농도의 연관성이 없는 지역구라고 한다면 영등포구에 공기 정화탑을 건설하더라도 인근 지역에서의 공기 질만 개선할 수 있다. 따라서 공기 정화탑 건설 위치는 건설비용과 지리적 위치를 고려하여 구조물이 위치한 지역구뿐만 아니라 주변 지역구에 가능한 좋은 영향을 끼치는, 주변 지역구와의 미세먼지 농도 연관성이 큰 곳으로 선정해야 할 필요성이 있다.

지역구와 지역구간의 미세먼지 농도에 대한 연관성을 파악하기 위한 방법으로 네트워크를 이용하는 방법이 있다. 지역구간 미세먼지 농도의 연관성을 네트워크 그래프를 통해 파악한다면 지역구간의 미세먼지 농도의 연관성을 한눈에 알 수 있고 연관성이 집중되는 곳을 쉽게 찾아 낼 수 있는 장점을 가지고 있다. 하지만 여기서의 문제점은 이러한 연관성 네트워크를 어떻게 찾아내는 것이다. 연관성 네트워크를 찾는 방법은 다양하지만 통계학에서는 네트워크를 찾는 모형으로 그래피컬 모형(graphical model)을 이용한다.

그래피컬 모형은 크게 베이지안 네트워크(Bayesian network; directed graphical model)과 마르코프 랜덤 필드(Markov random field; undirected graphical model)로 나눌 수 있다. 베이지안 네트워크는 방향성이 있는 그래피컬 모형으로 인과관계를 설명할 수 있는 장점을 가지고 있지만 일반적으로 사전정보, 즉 그래픽 구조에 대한 초기 값이나 순서 등을 분석 전에 결정해야 하고 네트워크의 순환(cycle)이 형성되지 않는다는 제약이 있기 때문에 사전 정보가 없이 변수 간의 연관성을 추정 할 때는 적절하지 않을 수 있다. 반면에 마르코프 랜덤 필드는 네트워크의 방향성은 가지고 있지 않지만 네트워크 구조를 추정함에 있어서 사전정보가 적게 필요하고 네트워크 구조의 비순환(acyclic)이라는 제약이 없기 때문에 변수 간 모든 연관성을 추정하기에 더 적절하다. 본 연구에서는 미세먼지 네트워크의 특성상 네트워크 구조의 사전정보가 부족하고 가능한 모든 지역구간의 연관성을 모형화하기 위해서 마르코프 랜덤 필드를 고려하였다.

미세먼지 농도와 같은 연속형 데이터에서 마르코프 랜덤 필드의 대표적인 네트워크 추정 모형으로 가우시안 그래피컬 모형(Gaussian graphical model)이 사용된다. 하지만 가우시안 그래피컬 모형은 일반적으로 고차원 데이터에서 정밀행렬(precision matrix)이라 불리는 공분산행렬(covariance matrix)의 역행렬이 존재하지 않는 문제점을 가지고 있다. 따라서 많은 연구자들은 ℓ_1 노름(norm) 벌점(penalty)을 통해 정밀 행렬의 희소성(sparsity)을 조절하여 정밀행렬을 추정하는 모형을 연구하였다(Yuan과 Lin, 2007a; Friedman 등, 2007; Rothman 등, 2008; Yuan, 2008). ℓ_1 노름 벌점 가우시안 그래피컬 모형의 다른 장점으로는 조율 모수(tuning parameter)를 통해 네트워크의 추정과 식별을 동시에 할 수 있으며 정밀행렬의 희소성을 조절해 네트워크를 단순화하여 해석을 더 용이하게 할 수 있다.

본 연구의 목적은 서울시의 지역구 간에 존재하는 미세먼지 농도의 연관성을 추정하고 연관성이 집중되는 지역구를 찾아 공기 정화탑 위치를 선정하는 방법을 제안하는 것이다. 본 연구에서는 지역구간 미세먼지 농도의 연관성 네트워크를 추정하는 것과 더불어 다른 지역구와 연관성이 집중되는 허브(hub)를 식별하는 것 또한 중요한 문제이다. 그래피컬 모형을 통해 추정된 네트워크의 허브를 식별하는 것은 많은 연구자들에 의해 연구 되고 있으며(Hero와 Rajaratnam, 2012; Firouzi와 Hero, 2013; KM Tan 등, 2014) 본 연구에서는 KM Tan 등이 제안한 허브 그래피컬 라쏘(Hub graphical lasso)를 이용하였다. 허브 그래피컬 라쏘는 조율 모수를 통해 네트워크의 희소성과 허브의 희소성을 조절 할 수 있고 다른 허브 식별 방법보다 좋은 성능을 보이며(KM Tan 등, 2014) R 프로그램의 hglasso 패키지를 통해 쉽게 이용이 가능한 이점이 있다.

본 연구는 다음과 같이 구성된다. 먼저 2장 연구모형에서는 ℓ_1 벌점화 기반의 가우시안 그래피컬 모형과 허브 그래피컬 라쏘를 소개하고 3장 자료 분석에서 실제 연구모형에서 소개한 허브 그래피컬 라쏘 모형을 통해 서울시의 미세먼지 데이터에 대한 분석결과를 제시한다. 4장에서는 결론 및 향후 연구과제에 대해 설명한다.

Ⅱ. 연구모형

본 장에서는 서론에서 소개한 네트워크 추정 모형인 ℓ_1 벌점화 기반의 가우시안 그래피컬 모형을 설명하고자 한다. 먼저 $\boldsymbol{X}=(X_1,X_2,...,X_p)$는 $n\times p$차원인 데이터 행렬이고 $X_i,\ i=1,2,...,p$는 $n\times 1$차원의 변수 벡터이다. 가우시안 그래피컬 모형에서는 일반적으로 데이터 행렬 $\boldsymbol{X}$는 평균이 $\boldsymbol{\mu}=[E(X_1),E(X_2),\dots,E(X_p)]^T$이고 공분산 행렬이 $\Sigma=E[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T]$인 다변량 정규분포를 따른다고 가정한다.

네트워크 구조를 표현하는 그래프는 $G=(V,E)$로 정의되며 여기서 $V=\{1,2,...,p\}$는 노드(node) 집합이고 E는 연결선(edge) 집합을 나타낸다. 그래피컬 모형에서 두 노드의 순서쌍 $(a,b),\ a,b\in V$가 집합 E에 포함된다는 것은 두 노드가 연결되어 있음을 의미하며 동시에 $X_{-(a,b)}=\{X_k;k\in V\backslash\{a,b\}\}$가 주어졌을 때 X_a와 X_b가 조건부 종속임과 동치이다.

X_a와 X_b의 조건부 독립성(종속성)을 추정하기 위한 방법으로 여러 접근 방법이 존재한다. 가장 일반적으로 사용하는 접근법은 가능도 함수(likelihood) 기반의 접근 방법이다. 이 방법은 정밀 행렬 $\Theta=\Sigma^{-1}$의 (a,b)번째 원소가 0이면 X_a와 X_b가 조건부 독립이라는 성질을 이용해 변수 사이의 조건부 독립성을 추정하는 방법이다.

1. ℓ_1 노름 벌점 가우시안 그래피컬 모형

고차원 데이터의 경우 정밀행렬 Θ가 존재하지 않아 일반적인 가우시안 그래피컬 모형을 사용하기 어렵다. 또한 전진 선택법(forward selection), 후진 소거법(backward selection) 그리고 단계적 선택법(stepwise selection)과 같은 기존의 모형 선택 방법은 부정확하고 계산량이 많아 고차원 데이터의 네트워크 구조를 추정하는 문제에는 적절하지 않다(Yuan and Lin, 2007). 이러한 문제점을 보완하기 위해 고차원 데이터의 분석에는 정밀 행렬의 원소에 ℓ_1 노름 벌점을 고려한 추정 방법을 사용한다. ℓ_1 노름 벌점 가우시안 그래피컬 모형은 다음과 같은 식(1)을 최소화 하는 정밀 행렬을 추정함으로써 그래프에서 노드 간의 종속성을 추정한다.

$$\min_{\Theta}\ \ell(\Theta;X)+\lambda\|\Theta\|_1 \tag{1}$$

여기서 Θ는 정밀 행렬로 음의 정부호가 아닌 행렬(non-negative definite matrix)이고 일반적으로 손실 함수(loss function) $\ell(\Theta;X)$를 $\ell(\Theta;X)=-\log\det(\Theta)+tr(S\Theta)$로 정의한다. 또한 손실 함수에서 S는 표본 공분산 행렬로 $S=(\boldsymbol{X}-\boldsymbol{\mu})^T(\boldsymbol{X}-\boldsymbol{\mu})/n$로 정의되며 $\det(A)$는 행렬 A의 행렬식(determinant), $tr(A)$는 행렬 A의 대각 합(trace)을 나타낸다. 또한 λ는 음이 아닌 조율 모수로 정밀 행렬의 희소성을 조절하게 되는데 값이 커지면 희소(sparse)한 정밀 행렬이 추정되며 작으면 조밀(dense)한 정밀 행렬

이 추정된다. 쉽게 말해서 조율 모수가 커지면 식(1)에서 $\|\Theta\|_1$이 전체 손실 함수에 큰 비중을 차지하게 되는데 이로 인해 전체 손실 함수에서 $\|\Theta\|_1$를 최소화하기 위해서 정밀 행렬에 0인 원소가 많아지게 된다. 반대로 조율모수가 작아지면 정밀 행렬에 0이 아닌 원소가 많아진다.

하지만 그래피컬 라쏘 모형은 함축적으로 각 연결선이 다른 연결선들과 같이 동일하게 등장함을 가정한다. 따라서 대부분의 노드가 비슷한 개수의 연결선을 가질 수 있고 이러한 가정은 실제 네트워크의 허브를 추정하는 과정에 적절하지 않다(KM Tan 등, 2014). 이러한 이유로 KM Tan 등은 허브 그래피컬 라쏘 모형을 제안했다.

2. 허브 그래피컬 라쏘(Hub graphical lasso)

허브 그래피컬 라쏘는 식(1)의 ℓ_1 노름 벌점을 대신해 식(2)와 같이 허브노드를 모델링 할 수 있는 벌점 함수를 고려한다.

$$P(\Theta) = \min_{\Theta} \left\{ \lambda_1 \|Z - diag(Z)\|_1 + \lambda_2 \|W - diag(W)\|_1 + \lambda_3 \sum_{j=1}^{p} \|(W - diag(W))_j\|_q \right\} \quad (2)$$
$$subject\ to\ \Theta = W + W^T + Z$$

여기서 Z는 희소 대칭 행렬(sparse symmetric matrix)로 Z의 원소는 허브노드가 아닌 노드 간의 연결성을 표현하며 W는 몇 개의 특정 열은 0이 아닌 원소로 이루어져 있고 나머지 열은 모든 원소가 완전히 0으로 이루어진 행렬로 0이 아닌 원소로 이루어져 있는 열은 허브노드에 해당한다. 또한 λ_1, λ_2, 그리고 λ_3는 음이 아닌 조율 모수로 λ_1은 Z 즉, 허브노드가 아닌 요소의 희소성을 조절하고 λ_2와 λ_3는 W의 희소성을 조절하게 되는데 λ_2는 허브노드의 희소성을 조절하고 λ_3는 선택되는 허브노드의 개수를 조절한다.

허브 그래피컬 라쏘는 식(2)의 벌점 함수를 이용하여 식(1)을 확장해 다음과 같은 식을 최소화하는 문제를 풀어 정밀 행렬 Θ를 추정한다.

$$\min_{\Theta} \left\{ \ell(\Theta, X) + \lambda_1 \|Z - diag(Z)\|_1 + \lambda_2 \|W - diag(W)\|_1 + \lambda_3 \sum_{j=1}^{p} \|(W - diag(W))_j\|_q \right\} \quad (3)$$
$$subject\ to\ \Theta = W + W^T + Z$$

여기서 벌점 함수의 q는 모든 양의 정수가 가능한데 특히, 허브 그래피컬 라쏘 모형은 추정되는 정밀 행렬에 조밀한 허브를 포함시키기 위해 $q=2$로 고정하였다. 제약조건을 만족시키면서 식(3)의 손실 함수와 벌점 함수를 함께 최소화 시키는 정밀 행렬을 찾는 것은 어려운 문제이다. 따라서 alternating direction method of

multipliers(ADMM) 알고리즘을 이용해 식(3)의 해를 찾는다. ADMM을 이용해 식(3)의 해를 찾는 자세한 알고리즘은 KM Tan 등(2014)를 참고할 수 있다.

3. 조율 모수 선택

그래피컬 모형을 통해 네트워크를 추정할 때의 어려운 문제 중 하나는 조율 모수를 선택하는 것이다. 그래피컬 라쏘 문제에서는 많은 연구자들이 다음 식(4)와 같은 Bayesian information criterion(BIC)를 최소화 하는 조율 모수 λ를 선택할 것을 제안한다.

$$BIC(\lambda) = -n\log\det(\widehat{\Theta_\lambda}) + ntr(S\widehat{\Theta_\lambda}) + \log(n)|\widehat{\Theta_\lambda}| \qquad (4)$$

여기서 $\widehat{\Theta_\lambda}$는 조율 모수 λ에 대해 추정되는 정밀 행렬을 나타내고 $|\widehat{\Theta_\lambda}|$는 $\widehat{\Theta_\lambda}$의 카디널리티(Cardinality)로 $\widehat{\Theta_\lambda}$의 0이 아닌 유일한 값의 개수를 나타낸다. 허브 그래피컬 라쏘에서는 Θ가 허브노드를 포함하지 않는 Z와 허브노드를 포함하는 W로 분해되기 때문에 기존 BIC에 이를 반영할 필요가 있다. 따라서 식(5)와 같이 Z와 W가 반영된 BIC를 통해 조율 모수를 선택한다.

$$BIC(\lambda_1,\lambda_2,\lambda_3) = -n\log\det(\hat{\Theta}) + ntr(S\hat{\Theta}) + \log(n)|\hat{Z}| + \log(n)(\nu + c[|\widehat{W}| - \nu]) \qquad (5)$$

여기서 $\hat{Z}$와 $\widehat{W}$는 조율 모수 λ_1,λ_2 그리고 λ_3에 의해 추정되는 행렬이고 $\hat{\Theta} = \hat{Z} + \widehat{W} + \widehat{W^T}$이다. 또한 $|\hat{Z}|$와 $|\widehat{W}|$는 $\hat{Z}$와 $\widehat{W}$의 카디널리티, $\nu = \sum_{j=1}^{p} 1_{\{\|\widehat{W}_j\|_0 > 0\}}$ 그리고 c는 0과 1사이의 상수이다. 만약 상수 c가 0에 가까울수록 BIC는 $\widehat{W}$에 더 많은 허브노드가 있는 것을 선호하고 1에 가까울수록 적은 허브노드가 있는 것을 선호하게 된다.

Ⅲ. 자료 분석

1. 데이터 설명

본 연구에서는 서울시 지역구별 미세먼지 농도간의 네트워크를 찾기 위해서 에어코리아에서 제공하는 2017년 대기환경 최종확정자료[1]를 사용하였다. 2017년 최종확정자료에는 2017년 01월 01일~09월 30일까지 전국의 356개의 측정소에서 측정한 SO_2, CO, O_3, NO_2, PM_{10} 그리고 $PM_{2.5}$의 실시간 농도가 들어있다. 그 중 서울시에 위치한 관측소의 PM_{10} 농도와 $PM_{2.5}$ 농도를 추출하였고 최종적으로 PM_{10} 농도가 측정되고 있는 39개 관측소와 $PM_{2.5}$ 농도가 측정되고 있는 25개 관측소 측정치를 사용하였다. 각 지역구별 미세먼지 농도의 대략적인 특성을 파악하기 위해 PM_{10}과 $PM_{2.5}$ 미세먼지 농도의 연 평균, 최솟값, 최댓값 그리고 일평균 농도가 환경공단기준의 '나쁨' 이상인 일수를 산출하였다. 각 지역구별 PM_{10}과 $PM_{2.5}$ 미세먼지 농도의 개요는 <표 1>과 같다.

PM_{10} 미세먼지의 연평균 농도가 가장 높은 지역구는 공항대로, 한강대로, 영등포로 순으로 대로변의 미세먼지 농도가 대체로 높게 나타났으며 가장 낮은 지역구는 강북구, 용산구, 종로구 순으로 나타났다. 또한 연중 미세먼지 농도의 중앙값이 가장 높은 지역구는 공항대로, 영등포로, 한강대로 순으로 나타나 연평균 미세먼지 농도와 같이 대로변의 PM_{10} 미세먼지 농도의 중앙값이 높게 나타났다. PM_{10} 미세먼지 농도의 중앙값이 가장 낮은 지역구는 강북구, 용산구, 중구 순으로 나타났다. 연중 PM_{10} 미세먼지 농도가 가장 높았던 지역구는 송파구로 06월 06일 $243\mu g/m^3$으로 나타났으며 '나쁨' 이상인 일수가 가장 많았던 지역구는 공항대로, 한강대로, 영등포로로 연평균 미세먼지 농도와 중앙값과 같이 대로변의 미세먼지 농도가 '나쁨' 이상을 자주 넘는 것으로 나타났다.

$PM_{2.5}$ 미세먼지의 연평균 농도가 가장 높은 지역구는 마포구, 양천구, 관악구 순으로 나타났으며 낮은 지역구는 강북구, 동대문구 순으로 나타났다. 또한 $PM_{2.5}$ 미세먼지 농도의 중앙값이 가장 높은 지역구로는 마포구, 양천구로 나타났으며 가장 낮은 지역구로는 강북구로 나타나 강북구 PM_{10}과 $PM_{2.5}$ 농도의 연평균과 중앙값으로 봤을 때 서울시 지역구 중 대기환경이 가장 좋은 지역구인 것으로 나타났다. 연중 $PM_{2.5}$ 미세먼지 농도가 가장 높았던 지역구는 강남구로 03월 21일 $107\mu g/m^3$로 나타났으며 '나쁨' 이상인 일수가 가장 많았던 지역구는 양천구, 마포구, 성북구 순으로 나타났다.

1) AirKorea 홈페이지 www.airKorea.co.kr

<표 1> 서울시 지역구별 PM_{10}과 $PM_{2.5}$ 미세먼지 농도의 데이터 개요

지역구	PM_{10}					$PM_{2.5}$				
	평균	중앙값	최솟값	최댓값	나쁨 횟수	평균	중앙값	최솟값	최댓값	나쁨 횟수
강남구	47.4	45	6	200	20	25	22.5	4	**107**	49
강남대로	59.0	55	10	230	42					
강동구	49.8	47	6	233	25	24	21	4	82	43
강변북로	55.2	52	6	208	40					
강북구	**37.0**	**33**	**5**	160	9	**22**	**19**	**3**	87	36
강서구	50.0	46	7	223	20	25	22	4	71	49
공항대로	**67.8**	**64**	13	217	**70**					
관악구	45.4	42	9	174	13	28	25	4	80	65
광진구	42	38	6	165	15	25	21	**3**	95	58
구로구	46.7	44	7	210	20	25	22	5	78	46
금천구	42.5	39	**5**	187	13	27	23	**3**	80	58
노원구	42.1	38	8	165	16	25	23	**3**	83	51
도봉구	45.1	41	6	209	19	24	22	4	72	47
도산대로	52.7	49	12	196	23					
동대문구	44.2	40	6	173	19	23	20	4	89	37
동작구	42.0	38	7	159	14	25	23	4	80	55
동작대로	58.0	54	12	218	41					
마포구	43.4	40	**5**	164	17	**30**	**26**	6	91	71
서대문구	48.5	45	6	219	25	26	24	4	67	54
서초구	48.1	45	7	233	23	24	22	4	84	44
성동구	48.0	44	6	237	23	25	23	4	80	49
성북구	48.9	43	8	233	21	27	24	4	98	66
송파구	47.0	44	6	**243**	22	24	22	**3**	74	42
신촌로	52.0	49	8	199	22					
양천구	43.0	39	7	147	15	29	**26**	5	84	**73**
영등포구	50.6	46	8	219	25	24	21	4	75	40
영등포로	60.0	57	10	216	50					
용산구	40.5	37	6	133	11	25	22	5	97	50
은평구	45.9	41	6	232	18	24	22	**3**	79	46
정릉로	56	52	10	229	40					
종로	46.6	43	11	148	18					
종로구	40.9	38	**5**	159	15	25	22	**3**	87	55
중구	41.6	37	6	227	16	25	22	4	82	53
중랑구	47.6	44	6	227	21	26	23	4	95	58
천호대로	49.2	46	13	141	21					
청계천로	48.2	45	9	151	15					
한강대로	60.6	56	11	240	51					
홍릉로	49.7	47	11	160	20					
화랑로	54.2	52	11	209	33					

2. 데이터 전처리

허브 그래피컬 라쏘를 통한 서울시 지역구간의 미세먼지 농도 네트워크를 추정하기에 앞서 서울시 전역의 미세먼지 농도에 영향을 주는 특정 요인이 지역구간의 미세먼지 농도 네트워크에도 영향을 줄 것이라 판단하여 그 요인의 효과를 제거하기로 한다. 서울시 미세먼지 농도에 광범위하게 영향을 미친다고 판단한 요인으로는 시간과 계절 그리고 기상 정보인 기온, 강수량, 습도 그리고 풍향으로 이러한 요인들은 관련연구에서 미세먼지 농도에 영향을 준다고 밝혀졌다(박애경 등, 2011; 이종현 등, 2017; 박충선, 2017). 고려한 요인 중 기상 정보는 기상자료개방포털[2]에서 제공하는 2017년 종관기상관측 자료를 이용하였고 미세먼지 농도에 영향을 미치는 요인 효과를 제거하기 위해서 회귀모형을 사용하였다. 회귀모형은 일반적으로 교란변수(confounding variable)의 효과를 제어하는데 사용할 수 있는데 반응변수(response variable)와 설명변수의 회귀모형에서 교란변수를 설명변수(explanatory variable)로 함께 고려하는 공변량분석(ANCOVA)이나 교란변수로 회귀모형을 적합하여 구한 잔차를 다시 데이터로 사용하는 방법이 있다(SA ESREY, 1990; RP Freckleton, 2002). 본 연구에서는 각 구의 미세먼지 농도를 반응변수로 하고 시간, 계절 그리고 기상 정보를 설명변수로 하는 모형을 적합 후 잔차를 사용하는 것으로 요인들의 효과를 제거하였다. 또한 미세먼지의 농도는 구별로 갖는 임의효과(random effect)가 있을 것으로 생각하여 회귀모형 중 임의효과가 반영된 혼합효과모형(mixed effects model)을 사용하였다. 미세먼지 농도에 영향을 미치는 요인 효과를 제거하기 위한 회귀 모형은 식(6)과 같다.

$$Y = \beta_{00}t + \beta_{01}t^2 + \sum_{i=1}^{12}\beta_{1i}I_i + \sum_{j=1}^{4}\beta_{2j}X_j + A_k + \varepsilon \tag{6}$$

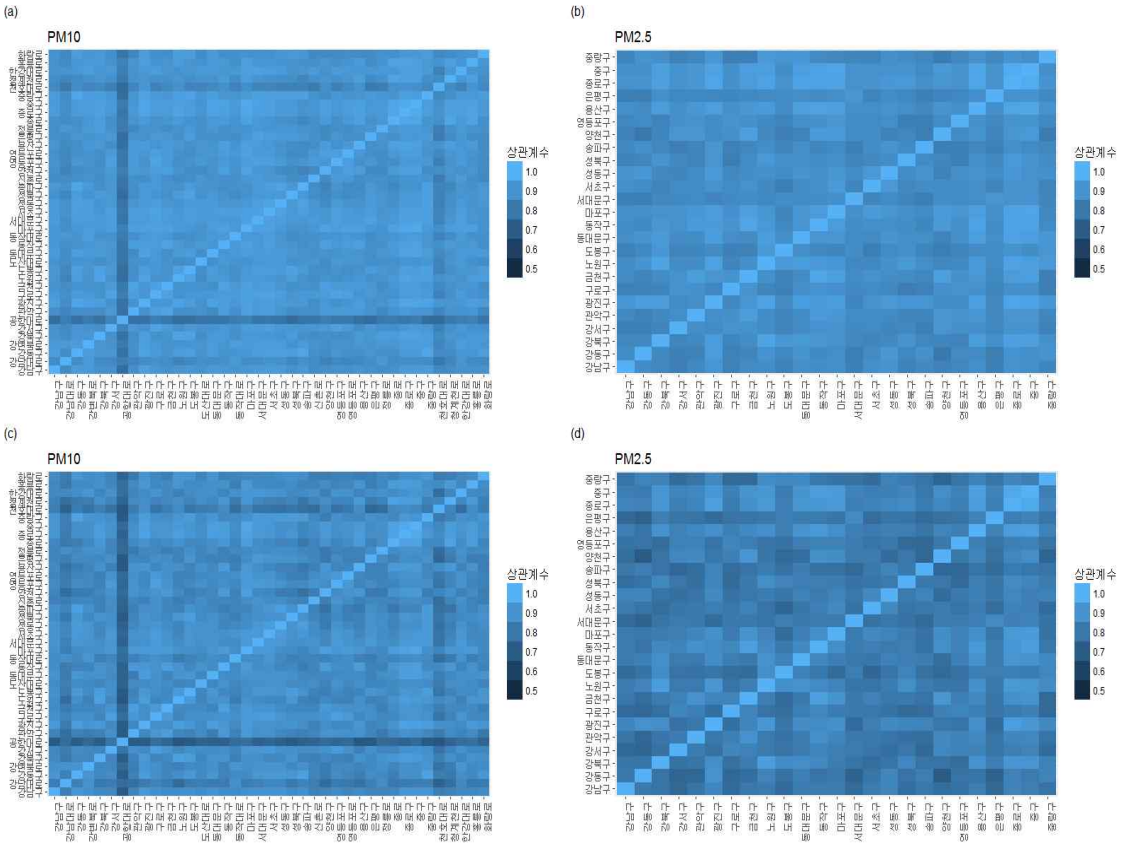
- Y는 실시간 측정한 PM_{10} 또는 $PM_{2.5}$ 미세먼지 농도
- t는 미세먼지 농도의 측정 시간
- I_i는 관측값이 i번째 달($i=1,...,12$)에 관측된 경우 1인 지시변수
- X_j $(j=1,\dots,4)$는 실시간 기상변수(각각 기온, 강수량, 습도, 풍향)
- A_k $(k=1,...,K)$는 지역구별 임의 효과(random effect)를 나타내는 더미변수로 PM_{10} 농도는 $K=39$이고 $PM_{2.5}$는 $K=25$

다음 <그림 1>은 요인 효과가 제거되기 전 미세먼지 농도의 지역구별 상관관계와 요인 효과가 제거된 후 상관관계를 나타낸다. <그림 1>에서 (a)와 (b)는 요인 효과가 제거되기 전 PM_{10}과 $PM_{2.5}$ 미세먼지 농도의 지역구별 상관관계를 나타낸 히트맵(heatmap)으로 PM_{10} 농도의 경우 대부분의 상관계수 값이 0.9 근처로 높은

2) 기상자료개방포털 data.kma.go.kr

상관관계를 가지면서 비슷한 값임을 알 수 있고 $PM_{2.5}$의 경우에도 대부분의 상관계수 값이 0.7 이상의 높은 값을 가지면서 비슷한 값임을 알 수 있다. (c)와 (d)는 요인 효과가 제거된 후 PM_{10}과 $PM_{2.5}$ 미세먼지 농도의 지역구별 상관관계를 나타낸 히트맵으로 여전히 상관관계가 높은 것을 알 수 있지만 미세먼지 농도의 지역구별 상관관계에 영향을 주는 고려한 요인의 효과가 제거되어 격자무늬 색의 차이가 뚜렷해짐을 알 수 있다.

<그림 1> 서울시 지역구별 PM_{10}과 $PM_{2.5}$ 미세먼지 농도의 상관관계



3. 미세먼지 농도 네트워크

본 절에서는 앞서 혼합효과모형을 통해 전처리한 데이터를 사용하여 서울시 지역구간의 미세먼지 네트워크를 추정한다. 네트워크를 추정하는 모형으로는 앞서 설명한 허브 그래피컬 라쏘 모형을 사용하였으며 허브 그래피컬 라쏘 모형은 R 프로그램의 hglasso 패키지를 이용하였다.

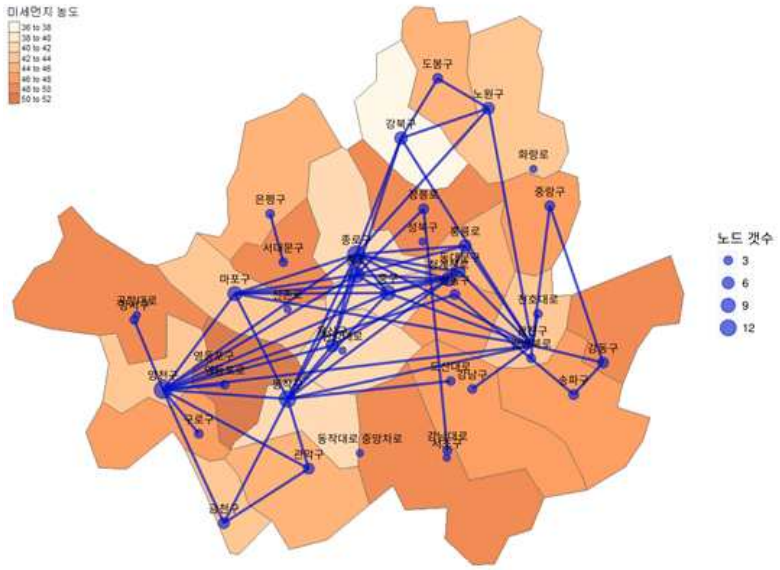
분석의 목적은 실제 네트워크 구조를 추정하기보다는 네트워크의 연결선이 집중되는 허브를 찾는 것이다. 따라서 Z의 희소성을 조절하는 조율모수 λ_1을 PM_{10} 미세먼지 데이터에서는 17로 고정하였고 $PM_{2.5}$ 미세먼지 데이터에서는 21.5로 고정하였다. 또한 허브의 개수를 조절하는 조율 모수 λ_3를 각각 85와 100으로 고정하였고 허브노드의 희소성을 조절하는 조율 모수 λ_2는 각각 0.007부터 7까지, 0.5부터 20까지 변화시키면서 미세먼지의 네트워크를 추정하였다. 각 조율 모수 λ_2의 범위에서 식(5)의 BIC가 가장 작은 최적의 네트워크는 <그림 2>, <그림 3>와 같다. 또한 λ_2를 변화시키면서 네트워크가 추정되는 과정은 <그림 4>, <그림 5>와 같다.

<그림 2>는 지역구별 연 평균 PM_{10} 미세먼지 농도와 추정된 미세먼지 네트워크를 나타낸다. 조율 모수 λ_2의 값이 클 때, 즉 지역구 간 미세먼지 농도의 강한 연관성을 가진 네트워크는 대부분이 인접한 지역구와 연결된 네트워크로 추정됐다. 따라서 PM_{10} 미세먼지 농도는 인접한 지역구 간에 연관성이 큰 것을 알 수 있다. 또한 조율 모수 λ_2 값이 점점 작아짐에 따라 미세먼지 농도가 다른 지역구들과 강한 연관성을 가진 허브노드가 순차적으로 등장하게 되는데 종로구, 중구 그리고 양천구가 제일 먼저 허브노드로 식별이 되며 다음으로 광진구와 종로 그리고 동작구와 청계천로 순으로 허브노드가 나타나게 된다. <그림 4>는 BIC 기준으로 BIC가 가장 작은 최적의 PM_{10} 미세먼지 네트워크가 추정된 그림이다. <그림 4>를 통해 식별할 수 있는 허브노드로는 광진구, 동작구, 종로구, 종로, 중구, 양천구, 마포구 그리고 청계천로로 이러한 지역구들은 연평균 PM_{10} 미세먼지 농도는 다른 지역구에 비해 크게 높지는 않지만 인근 지역구뿐만 아니라 비교적 멀리 떨어진 지역구와도 연관성이 높은 지역으로 나타났다.

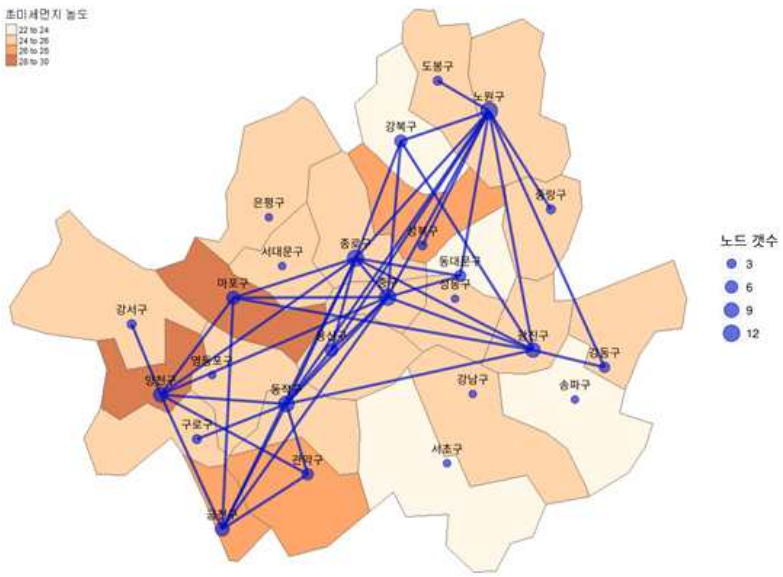
<그림 3>은 지역구별 연평균 $PM_{2.5}$ 미세먼지 농도와 추정된 미세먼지 네트워크를 나타낸다. 조율 모수 λ_2의 값이 큰 네트워크를 살펴보면, PM_{10} 미세먼지 네트워크와 다르게 인접한 지역구뿐만 아니라 지리적으로 더 멀리 떨어진 지역구까지도 강한 연관성을 보여준다. 이는 $PM_{2.5}$ 미세먼지가 PM_{10} 미세먼지보다 입자의 크기가 작아 더 먼 거리까지 이동하기 때문인 것으로 생각된다. 또한 조율 모수 λ_2 값이 점점 작아지면 가장 먼저 금천구, 종로구 그리고 중구가 허브노드로 나타나며 그 다음 광진구, 노원구, 동작구 그리고 양천구가 나타나게 된다. <그림 5>는 BIC 기준으로 찾은 최적의 $PM_{2.5}$ 미세먼지 네트워크가 추정된 그림으로 광진구, 금천구, 노원구, 동작구, 종로구, 중구 그리고 양천구가 허브노드로 나타났고 금천구와 양천

구는 연평균 $PM_{2.5}$ 미세먼지 농도가 높으면서 다른 지역구와의 연관성이 집중되는 곳으로 나타났다.

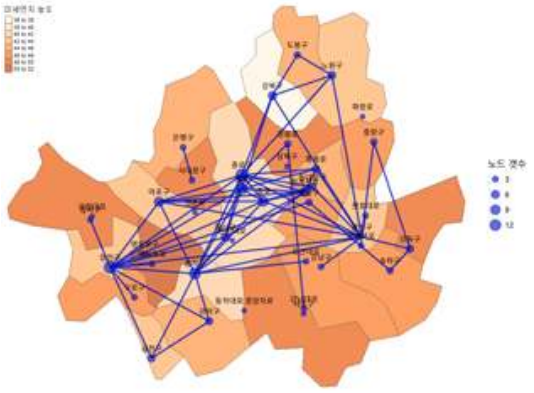
<그림 2> BIC 기준의 최적의 PM_{10} 미세먼지 네트워크 그림



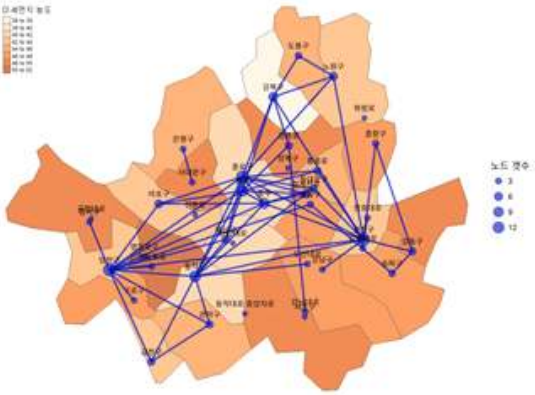
<그림 3> BIC 기준의 최적의 $PM_{2.5}$ 미세먼지 네트워크 그림



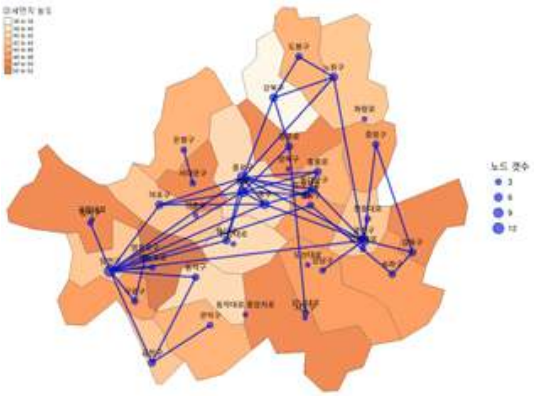
<그림 4> 조율 모수 λ_2에 따른 PM_{10} 미세먼지 네트워크 그림



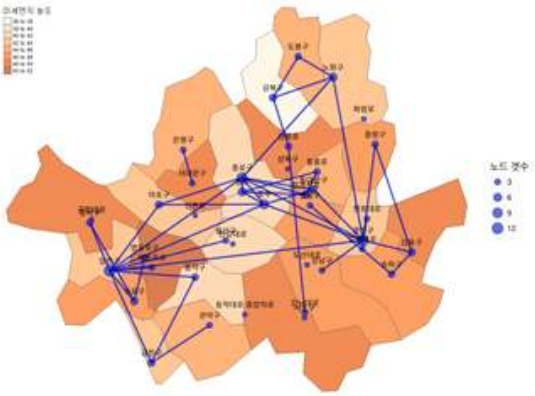
$\lambda_2 = 0.007$



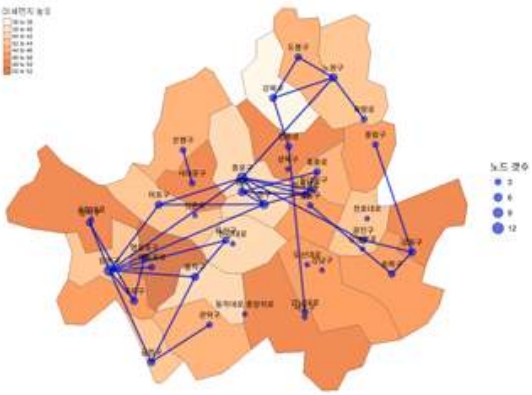
$\lambda_2 = 0.82$



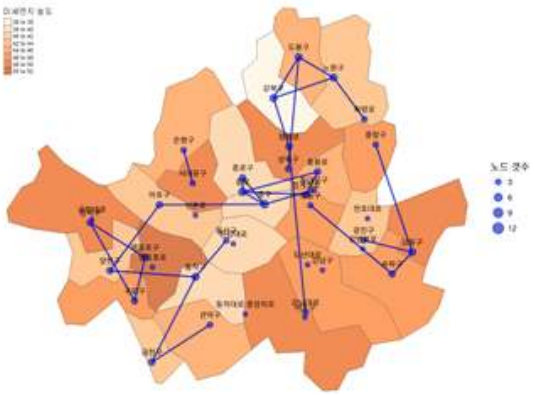
$\lambda_2 = 2.70$



$\lambda_2 = 4.35$

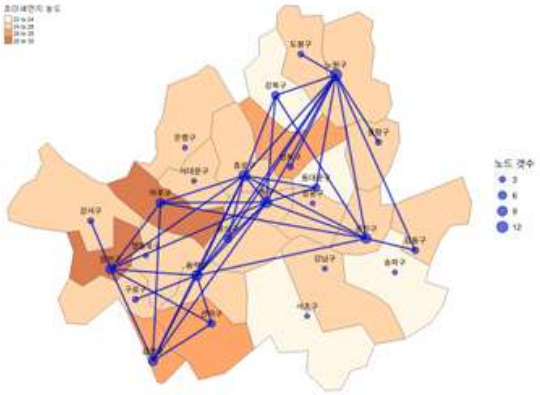


$\lambda_2 = 5.52$

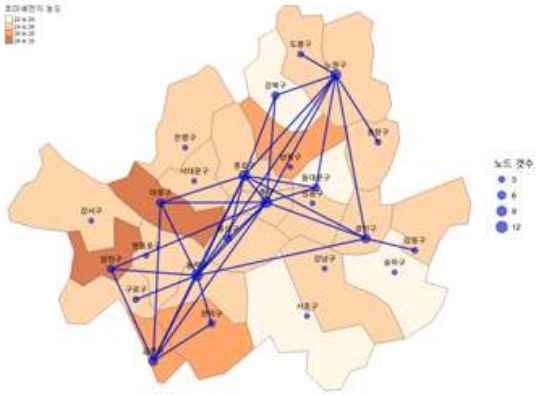


$\lambda_2 = 7.00$

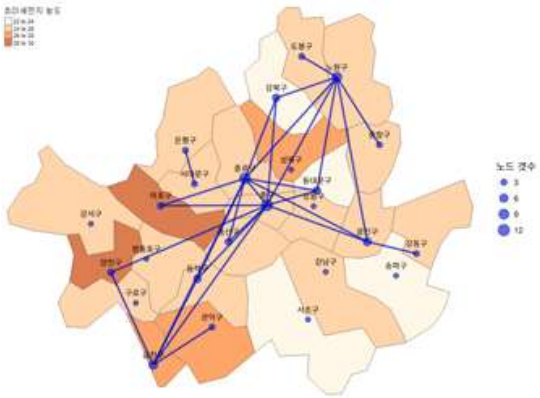
<그림 5> 조율 모수 λ_2에 따른 $PM_{2.5}$ 미세먼지 네트워크 그림



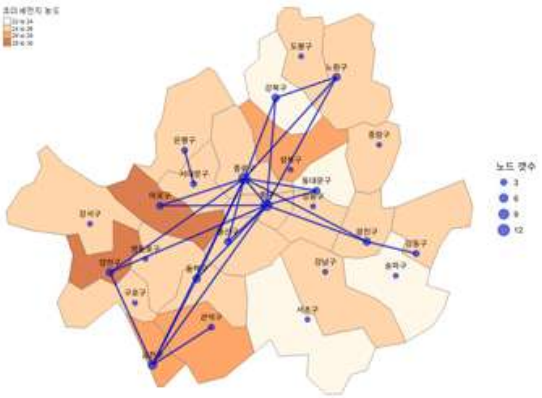
$\lambda_2 = 0.02$



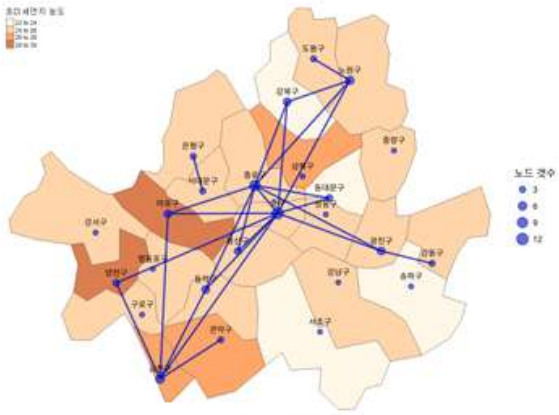
$\lambda_2 = 1.85$



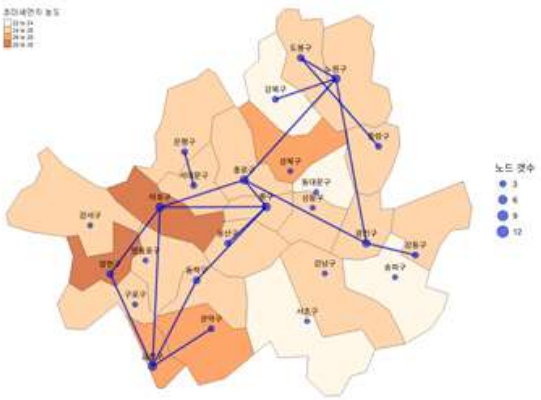
$\lambda_2 = 4.79$



$\lambda_2 = 6.08$



$\lambda_2 = 9.80$



$\lambda_2 = 20.0$

Ⅳ. 결 론

본 연구는 최근 서울시의 공기 정화탑 건설 검토 계획에 맞춰 공기 정화탑의 적절한 위치를 찾는 것을 목적으로 하였다. 공기 정화탑의 적절한 위치를 선별하는 문제는 지리적 요인 또는 풍향 등의 기상 요인을 고려하는 방법 등 다양한 접근 방법이 있을 수 있다. 하지만 특정 지역구가 다른 지역구들의 미세먼지 농도에 영향을 주거나 받는다면 그 지역구의 미세먼지를 정화시킴으로써 다른 지역구의 미세먼지 농도도 줄어드는 효과를 볼 수 있다고 판단하였다. 따라서 본 연구에서는 허브 그래피컬 라쏘 모형을 이용해 지역구간의 미세먼지 농도의 네트워크를 추정하고 네트워크 연결선들이 집중되는 지역구를 공기 정화탑의 위치로 제안하였다.

허브 그래피컬 라쏘 모형을 통해 PM_{10} 미세먼지 네트워크를 추정한 결과, PM_{10} 미세먼지 농도는 대부분 인근 지역구와 강한 연관성을 가지는 것으로 나타났고 연관성이 집중되는 지역구로는 광진구, 동작구, 종로구, 종로, 중구, 양천구, 마포구 그리고 청계천로로 나타났다. 이 지역구들은 연평균 PM_{10} 미세먼지 농도로만 보면 높은 지역은 아니지만 다른 지역구와 미세먼지 농도의 연관성이 집중되는 곳으로 나타났고 특히, 광진구, 종로구 그리고 양천구는 가장 먼저 미세먼지 농도의 연관성이 집중되는 곳일 뿐만 아니라 인접한 지역구의 연평균 PM_{10} 미세먼지가 비교적 높은 곳으로 공기 정화탑을 건설시 가장 먼저 우선순위를 두어야한다고 판단된다.

$PM_{2.5}$ 미세먼지 농도의 경우, PM_{10} 미세먼지 농도보다 비교적 먼 거리의 지역구와 강한 연관성을 가지는 것으로 나타났다. 이는 $PM_{2.5}$ 미세먼지 입자가 PM_{10} 미세먼지 입자보다 크기가 작아 데이터가 측정되는 시간 동안 더 먼 거리로 이동하기 때문인 것으로 보인다. 또한 $PM_{2.5}$ 미세먼지 농도의 연관성이 집중되는 지역구로는 광진구, 금천구, 노원구, 동작구, 종로구, 중구 그리고 양천구로 나타났으며 금천구와 양천구는 연평균 $PM_{2.5}$ 미세먼지 농도가 높은 곳일 뿐만 아니라 연관성이 집중되는 곳으로 $PM_{2.5}$ 미세먼지 농도를 고려한다면 금천구와 양천구에 가장 먼저 공기 정화탑의 건설을 고려해야 할 것을 보인다.

추정된 미세먼지 네트워크를 종합적으로 살펴보면 광진구, 종로, 종로구, 중구 그리고 양천구가 PM_{10}과 $PM_{2.5}$ 미세먼지 네트워크가 공통적으로 집중되는 곳으로 나타났다. 종로와 종로구 중구는 인근 지역구임을 고려한다면 광진구, 중구 그리고 양천구에 공기 정화탑이 건설될 시 서울시 대부분의 지역이 공기 정화탑의 영향권에 포함될 뿐만 아니라 네트워크가 연관된 많은 지역구들의 공기 질을 개선할 수 있을 것으로 보인다.

본 연구에서 살펴본 허브 그래피컬 라쏘 모형은 네트워크의 방향성이 없다는 한계점을 지니고 있다. 추후 연구해 볼 만한 주제는 앞서 언급한 베이지안 네트워크를 이용한 방향성이 있는 미세먼지 네트워크이다. 베이지안 네트워크를 이용해 미세먼지 네트워크를 추정하려면 네트워크 노드의 순서나 구조 등의 사전 정보를 필요로 한다. 따라서 본 연구에서 추정한 미세먼지 네트워크 구조나 다른 연구를 통

해 미세먼지 발생지를 알아낼 수 있다면 이를 베이지안 네트워크의 사전정보로 이용하여 방향성이 있는 서울시 미세먼지의 네트워크를 추정할 수 있다. 예를 들어 가우시안 DAG 모형(Gaussian directed acyclic graphical model) 등을 이용할 수 있으며 추정된 방향성 네트워크를 통해 미세먼지의 방향이나 지역구간의 연관성 등 더 많은 미세먼지 정보를 알 수 있을 것이라 판단된다.

참 고 문 헌

박애경·허종배·김 호(2011년), “서울시 미세먼지 농도에 영향을 미치는 요인 분석”, 『Particle and Aerosol Reasearch』, 7(2): 59-68, (사)한국입자에어로졸학회.

박충선(2017), “2015년 서울시 미세먼지 농도의 변화와 기상 조건과의 관련성”, 『한국사진지리학회지』, 27(2):47-64.

신동천(2007), "미세먼지의 건강영향", 『대한의사협회지』, 50: 175-182, 대한의사협회.

이종현·김영민·김용구(2017), “공간패널모형을 이용한 국내 초미세먼지 농도에 대한 분석”, 『한국데이터정보과학회지』, 28(3): 473-481.

주간조선. (2018). "깨끗해진 중국 ‘괴물’ 청정기 덕분?", 2503호. 2018.04.16.

Firouzi, H., and Hero, A. O. (2013, September). "Local hub screening in sparse correlation graphs", In Wavelets and Sparsity XV (Vol. 8858, p. 88581H), International Society for Optics and Photonics.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). "Sparse inverse covariance estimation with the graphical lasso", Biostatistics, 9(3): 432-441.

Hero, A., and Rajaratnam, B. (2012). "Hub discovery in partial correlation graphs", IEEE Transactions on Information Theory, 58(9): 6064-6078.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). "Sparse permutation invariant covariance estimation", Electronic Journal of Statistics, 2: 494-515.

Tan, K. M., London, P., Mohan, K., Lee, S. I., Fazel, M., and Witten, D. (2014). "Learning graphical models with hubs", The Journal of Machine Learning Research, 15(1): 3297-3331.

Yuan, M. (2008). "Efficient computation of $\ell 1$ regularized estimates in Gaussian graphical models", Journal of Computational and Graphical Statistics, 17(4): 809-826.

Yuan, M. and Lin, Y. (2007). "Model selection and estimation in the gaussian graphical model", Biometrika, 94: 19-35.