应用概率统计 第 35 卷
第 5 期 2019 年 10 月

Chinese Journal of Applied Probability and Statistics
Oct., 2019, Vol. 35, No. 5, pp. 495-507
doi: 10.3969/j.issn.1001-4268.2019.05.005

基于稀疏图模型的 $PM_{2.5}$ 分布的网络结构学习 *

张 海* 郭 骁 任 飒 邓亚景

(西北大学数学学院, 西安, 710127)

摘 要: 本文聚焦于中国 31 省会城市 $PM_{2.5}$ 污染的网络结构分析. 基于稀疏图模型研究 $PM_{2.5}$ 污染网络的中心点和 $PM_{2.5}$ 污染网络的社区结构, 结果表明: $PM_{2.5}$ 污染严重的城市同时也是 $PM_{2.5}$ 污染网络的中心点; $PM_{2.5}$ 污染网络存在明显的区块特征, 同一区块内的城市可认为其污染具有某种共性. 基于研究结果, 对于 $PM_{2.5}$ 污染的治理, 我们建议关注重点区域城市, 开展分地区治理, 并重点关注西部污染.

关键词: 图模型; 网络; 社区; 无标度; $PM_{2.5}$

中图分类号: O212.4

英文引用格式: ZHANG H, GUO X, REN S, et al. Structure learning of $PM_{2.5}$ distribution using sparse graphical models [J]. Chinese J Appl Probab Statist, 2019, 35(5): 495–507. (in Chinese)

§1. 引 言

2013 年, “雾霾” 成为年度关键词, 雾霾笼罩中国各省 (区、市), 对全国人民的生活、健康、出行产生了极大的危害. 雾霾天气是一种大气污染状态, 是对大气中各种悬浮颗粒物含量超标的笼统表述. $PM_{2.5}$ 被认为是造成雾霾天气的元凶之一. 不同于 PM_{10} (可吸入颗粒物), $PM_{2.5}$ 是指大气中直径小于或等于 2.5 微米的颗粒物, 属于细颗粒物. 2013 年之前, 中国在大气污染颗粒物方面的监测对象主要是 PM_{10}. 2013 年 1 月 1 日, 中国环保部开始正式将 $PM_{2.5}$ 列入空气监测指标中, $PM_{2.5}$ 成为了一个重要的监测空气污染程度的指数. $PM_{2.5}$ 来源主要有自然和人为两种. 自然过程如尘土、森林火灾、细菌等均可产生 $PM_{2.5}$. 而其绝大部分来自于人类的生产生活过程. 化石燃料的燃烧、机动车尾气的排放、生物质燃烧等均会产生 $PM_{2.5}$. $PM_{2.5}$ 不仅会影响环境, 造成大气污染, 更重要的是长期吸入 $PM_{2.5}$ 会对人体造成极大的危害, 甚至会增加致死和致病概率. 开展关于 $PM_{2.5}$ 的数据分析有着非常重要的意义, 可为治理 $PM_{2.5}$ 提供依据.

$PM_{2.5}$ 受如气象、温度、湿度、风向等很多因素影响, 且其生成机制复杂. $PM_{2.5}$ 近期受到了广泛关注, 如文献 [1–3]. 其中, 一系列关于 $PM_{2.5}$ 的研究试图从数据分析的角度解释 $PM_{2.5}$ 的规律, 如文献 [4, 5]. 他们以北京等中心城市为研究对象, 对空气质量进行量化

*国家自然科学基金项目 (批准号: 11571011) 资助.

*通讯作者, E-mail: zhanghai@nwu.edu.cn.

本文 2017 年 12 月 13 日收到, 2018 年 10 月 17 日收到修改稿.

分析, 主要研究了风向、风速、降水、冬季供暖、APEC、大阅兵等对 $PM_{2.5}$ 的影响, 比较了几大城市各季节 $PM_{2.5}$ 的主要特征和动态趋势. 其研究成果为人们理解 $PM_{2.5}$, 认识 $PM_{2.5}$ 与各因素的关系提供参考. 他们以每个城市的污染为研究对象, 开展了城市局部污染分析. 事实上, 各城市的 $PM_{2.5}$ 污染并不是相互独立的, 由于地理结构、经济结构等因素影响, 同一地理结构和经济发展区域内城市之间的 $PM_{2.5}$ 污染必然存在一定的相互影响, 比如北京地区的 $PM_{2.5}$ 污染可能会受到河北等相邻地区 $PM_{2.5}$ 污染的影响. 因此, $PM_{2.5}$ 数据分析需要更大范围开展, 需要研究全国范围内 $PM_{2.5}$ 污染的网络结构, 从而为有效治理 $PM_{2.5}$ 提供帮助.

开展 $PM_{2.5}$ 污染的治理需要研究几个关键问题. 首先, 我们需要确认 $PM_{2.5}$ 污染源地区. 一般地, 长期污染严重的区域有很大可能是污染源区域. 在污染源确认后, 可开展重点治理. 其次, 需要了解 $PM_{2.5}$ 污染在全国范围内有没有区域特征, 中国国土面积大, 存在工业分布、人口分布、以及地形地理结构的差异. 因此, $PM_{2.5}$ 污染必然会存在成因差异, 其分布具有结构特征. 为了更有效治理 $PM_{2.5}$, 需按区、分块、分结构制定治理政策. 本文旨在从数据分析的角度, 基于近年来发展的稀疏图模型方法, 研究中国省会城市 $PM_{2.5}$ 污染的网络结构, 为理解 $PM_{2.5}$ 污染和 $PM_{2.5}$ 的治理提供统计参考和帮助. 我们收集了从 2014 年 1 月 1 日到 2017 年 12 月 31 日中国主要城市的日平均 $PM_{2.5}$ 浓度数据. 由于香港、澳门、台湾的 $PM_{2.5}$ 数据并未公开, 我们收集了包括除香港、澳门、台湾之外的 22 个省会城市、4 个直辖市、5 个自治区首府, 共 31 个城市的 $PM_{2.5}$ 浓度数据. 数据来源于中国空气质量在线监测分析平台*.

将 31 个省会城市的 $PM_{2.5}$ 污染分别看做 $X_1, X_2, \ldots, X_{31}$ 这 31 个随机变量, 并假设其联合分布为多维 Gaussian 分布, 即 $\boldsymbol{X} = (X_1, X_2, \ldots, X_{31}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. 将 2014 至 2017 年的 $PM_{2.5}$ 污染数据视为从分布 $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ 中抽取的样本. 我们目的是基于 31 个城市 2014 至 2017 年的 $PM_{2.5}$ 数据, 学习出各城市 $PM_{2.5}$ 污染网络结构, 包括网络的中心点, 及各城市之间 $PM_{2.5}$ 污染的社区结构. 图模型 [6–8] 是从数据中估计网络结构的有效工具. 它是一类用图 (Graph) 来可视化表示随机变量联合概率分布的模型, 被广泛应用于建立各种相互作用的单元之间形成的网络结构, 如基因调控网络、蛋白质网络、社交网络等, 从而为研究和理解变量之间的相互关系提供了工具和参考 [9]. 本文中, 我们将交替使用图和网络来描述变量之间的连接关系. 图模型的主要构成要素为节点 (Nodes) 和边 (Edges), 其中节点与随机变量对应, 边与变量之间的条件依赖 (Conditional Dependence) 关系对应. 具体地, 用二元组 $G = (V, E)$ 表示一个图, 其中, $V = 1, 2, \ldots, p$ 分别代表节点变量 $X_1, X_2, \ldots, X_p$, $E \subseteq V \times V$ 代表边的集合. 马尔科夫性质 [6–8] 是图模型将图与分布联系起来的桥梁. 对于 Gaussian 图模型, 令 $\boldsymbol{\Theta} = (\boldsymbol{\Sigma})^{-1} = (\theta_{ij})_{i,j=1,2,\ldots,31}$, 则由多元 Gaussian 分布的性质, $\theta_{ij} = 0$ 等价于 X_i 与 X_j 条件独立, 同时等价于节点 i 和节点 j 无边相连 [6–8]. 因此, 构建 31 个省会城市的 $PM_{2.5}$ 污染所形成的网络结构等价于估计协方差矩阵的逆 $\boldsymbol{\Theta}$. 当数据不服从

*https://www.aqistudy.cn/

Gaussian 分布时, Gaussian 图模型的损失函数可以解释为待估参数 $\boldsymbol{\Theta}$ 与样本协方差矩阵的逆的 Bregman 散度, 网络结构仍然对应于 $\boldsymbol{\Theta}$. 对于 $\boldsymbol{\Theta}$ 的估计, 在如今高维数据爆炸的背景下, 已有大量关于高维图模型的研究成果, 如文献 [10–17] 等. 上述图模型理论及方法为各类应用问题提供了有力的工具.

基于稀疏图模型方法并结合 $PM_{2.5}$ 污染治理的关键问题, 我们主要做了如下工作:

1) 基于无标度图模型发现 $PM_{2.5}$ 污染的中心点 (Hub). 结果表明, $PM_{2.5}$ 污染严重的城市同时也是 $PM_{2.5}$ 污染网络的中心点, 从统计意义来讲, 是与其余城市 $PM_{2.5}$ 污染相互影响最多的城市.

2) 基于分块对角图模型分析 $PM_{2.5}$ 污染的分区结构. 结果表明, $PM_{2.5}$ 污染存在明显的区块特征, 同一区块内部的城市 $PM_{2.5}$ 污染的相互影响强于不同区块的城市 $PM_{2.5}$ 污染的相互影响. 另外, 所得城市区块与地理位置有很高的一致性.

§2. $PM_{2.5}$ 城市污染趋势分析

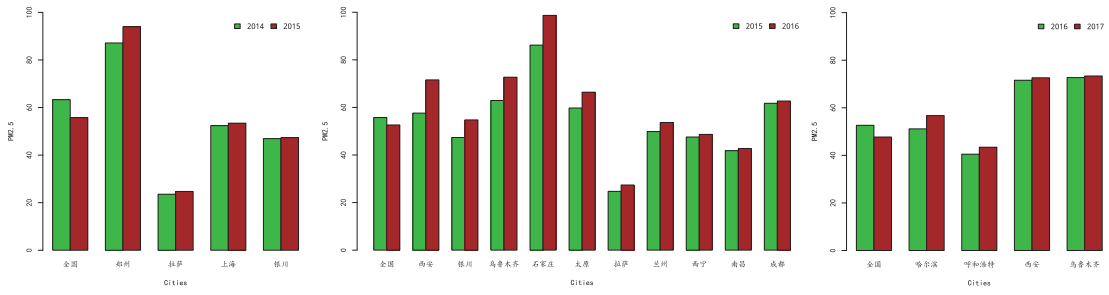
近年来, 国家各级部门均加大力度治理 $PM_{2.5}$ 污染, 出台多项政策措施, 推行节能减排目标任务, 本节分析近 4 年全国省会城市 $PM_{2.5}$ 污染趋势, 通过比较近 4 年 $PM_{2.5}$ 平均浓度说明污染变化趋势.

我们统计了全国 2014 年至 2017 年 4 年的年平均 $PM_{2.5}$ 浓度, 具体结果见表 1. 结果显示, 在 2014 至 2017 年 4 年中, 全国 $PM_{2.5}$ 平均浓度逐年下降, 其中 2015 年相比 2014 年下降 11.94%, 2016 年相比 2015 年下降 5.57%, 2017 年相比 2016 年下降 9.35%, 说明这 4 年尤其是 2015 年 $PM_{2.5}$ 污染治理有一定效果. 特别是东部省份, $PM_{2.5}$ 污染浓度均有大幅下降趋势. 但是, 不同于全国 $PM_{2.5}$ 平均浓度下降的趋势, 相比于 2014 年, 有 4 个城市 2015 年的 $PM_{2.5}$ 平均浓度上升, 如图 1(a) 所示, 分别是郑州、拉萨、上海、银川, 其 $PM_{2.5}$ 平均浓度分别增加了 7.83%、4.97%、1.97%、0.94%. 而相比于 2015 年, 有 10 个城市 2016 年的 $PM_{2.5}$ 平均浓度较 2015 年有所增加, 如图 1(b) 所示, 其中西安上升了 24.25%, 银川上升了 15.63%, 乌鲁木齐上升了 15.56%, 石家庄上升了 14.49%, 太原上升了 11.11%, 拉萨上升了 10.80%, 兰州上升了 7.68%, 西宁上升了 2.31%, 南昌上升了 2.20%, 成都上升了 1.52%. 基于上述结果, 我们发现 2015 年相比于 2014 年, $PM_{2.5}$ 平均浓度增加的城市较少而且增加幅度小, 大部分城市都同全国的下降趋势一致. 但是 2016 年相比于 2015 年, 在全国 $PM_{2.5}$ 浓度下降的大趋势下, 有多达 10 个省会城市 $PM_{2.5}$ 浓度上升, 特别地, 中国西北地区五省省会西安、银川、乌鲁木齐、兰州、西宁 $PM_{2.5}$ 浓度全部上升, 而西安上升幅度最大. 而 2017 年相比于 2016 年, 仅有 4 个省会城市的 $PM_{2.5}$ 浓度上升, 如图 1(c) 所示, 哈尔滨、呼和浩特、西安、乌鲁木齐分别上升了 10.93%、7.26%、1.45%、0.92%.

以上分析说明全国 $PM_{2.5}$ 平均浓度逐年下降, $PM_{2.5}$ 污染治理初见成效. 虽然西北部地区在 2017 年污染得到了缓解, 但由于其在 2016 污染严重, 故仍然需要进一步关注.

表 1　各城市四年 $PM_{2.5}$ 平均浓度 ($\mu g/m^3$) 排序

	2014 年		2015 年		2016 年		2017 年	
排序	城市	$PM_{2.5}$	城市	$PM_{2.5}$	城市	$PM_{2.5}$	城市	$PM_{2.5}$
1	石家庄	122.99	郑州	93.97	石家庄	98.69	石家庄	81.46
2	济南	89.94	济南	89.16	郑州	78.55	乌鲁木齐	73.37
3	郑州	87.14	石家庄	86.20	济南	75.40	西安	72.62
4	天津	86.53	北京	79.91	乌鲁木齐	72.70	郑州	70.90
5	北京	84.56	沈阳	71.50	西安	71.58	济南	64.34
6	武汉	81.78	哈尔滨	69.54	北京	69.30	太原	64.27
7	合肥	79.96	天津	68.92	天津	68.54	天津	60.40
8	西安	75.78	武汉	68.84	太原	66.40	哈尔滨	56.73
9	长沙	73.90	合肥	65.13	成都	62.68	北京	56.65
10	南京	73.82	长春	64.94	合肥	57.35	合肥	55.28
11	哈尔滨	72.81	乌鲁木齐	62.91	武汉	56.94	成都	52.59
12	成都	72.72	成都	61.74	银川	54.76	长沙	51.82
13	沈阳	72.00	长沙	60.20	长沙	54.32	武汉	51.81
14	太原	68.06	太原	59.76	重庆	53.87	沈阳	49.17
15	长春	65.47	西安	57.61	沈阳	53.79	兰州	47.91
16	乌鲁木齐	64.08	南京	56.20	兰州	53.68	长春	45.59
17	重庆	62.88	杭州	55.25	哈尔滨	51.14	银川	44.34
18	西宁	61.46	重庆	53.94	西宁	48.66	重庆	43.71
19	杭州	61.42	上海	53.42	杭州	48.62	呼和浩特	43.43
20	兰州	58.40	兰州	49.85	南京	47.78	杭州	42.91
21	上海	52.39	西宁	47.56	长春	46.23	南京	40.75
22	南昌	51.21	银川	47.36	上海	44.59	南昌	40.19
23	南宁	48.32	呼和浩特	42.37	南昌	42.74	上海	38.64
24	广州	47.79	南昌	41.82	呼和浩特	40.49	西宁	38.10
25	银川	46.92	南宁	40.62	贵阳	36.70	南宁	34.80
26	贵阳	45.56	广州	38.49	南宁	36.57	广州	34.53
27	呼和浩特	43.94	贵阳	37.74	广州	35.96	贵阳	31.58
28	昆明	32.51	昆明	28.79	昆明	27.79	昆明	27.02
29	福州	32.03	福州	28.17	拉萨	27.39	福州	25.90
30	拉萨	23.55	拉萨	24.72	福州	27.30	海口	19.49
31	海口	22.39	海口	21.36	海口	21.17	拉萨	18.86



(a) 2015 年对比 2014 年　　**(b) 2016 年对比 2015 年**　　**(c) 2017 年对比 2016 年**

图 1　相邻年份 $PM_{2.5}$ 污染对比条形图. (a)、(b)、(c) 分别显示了 2015 年比 2014 年、2016 年比 2015 年、2017 年比 2016 年 $PM_{2.5}$ 污染上升的省会城市和全国.

§3. $PM_{2.5}$ 网络中心点分析

本节基于无标度图模型 [18] 构建 31 个省会城市的 $PM_{2.5}$ 污染所形成的无标度网络, 从而发现 $PM_{2.5}$ 污染网络的中心点 (Hub). 我们首先介绍无标度图模型, 然后将无标度图模型应用于 $PM_{2.5}$ 数据并分析结果.

3.1 无标度图模型

为了发现 $PM_{2.5}$ 网络中心节点, 我们利用无标度图模型 [18] 来建立 $PM_{2.5}$ 污染的网络, 顾名思义, 该模型所得网络为无标度网络 [19]. 无标度网络的典型特征是网络中大部分节点度很小, 存在度非常大的少数节点, 后者被称为中心点. 对于 $PM_{2.5}$ 数据, 网络的中心点从统计角度可以认为是与其余城市的 $PM_{2.5}$ 相互影响最多的城市, 网络的中心点所对应的城市不一定就是 $PM_{2.5}$ 的污染源城市, 但是其所在区域往往就是污染严重地区. 无标度图模型的数学框架如下: 假设每一个城市的 $PM_{2.5}$ 值对应一个随机变量 X_i, 不失一般性, 我们假设城市个数为 p (本文 p 即为 31). 假设 $\boldsymbol{X} = (X_1, X_2, \ldots, X_p) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. 令 $\boldsymbol{\Theta} = (\boldsymbol{\Sigma})^{-1} = (\theta_{ij})_{i,j=1,2,\ldots,p}$. 我们目标是估计 $\boldsymbol{\Theta}$ 使其对应网络为无标度的. 对于无标度网络 $G = (V, E)$, 节点的度服从幂律分布, 即

$$P(d) \propto d^{-\alpha},$$

其中, $\alpha > 0$ 为尺度参数, d 为节点的度. 反映到网络中, 节点 i $(i = 1, 2, \ldots, p)$ 的度 d_i 为 $\sum\limits_{j:j\neq i} I_{\{\theta_{ij}\neq 0\}}$, 即 $\boldsymbol{\theta}_{-i} = (\theta_{i1}, \ldots, \theta_{i,i-1}, \theta_{i,i+1}, \ldots, \theta_{ip})'$ 的 l_0 拟范数. 考虑到 l_0 的不连续性及其带来的组合优化问题, 用 $\|\boldsymbol{\theta}_{-i}\|_q^q$ $(0 < q < 1)$ 代替 d_i, 其中 $\|\boldsymbol{\theta}_{-i}\|_q^q = |\theta_{i1}|^q + \cdots + |\theta_{i,i-1}|^q + |\theta_{i,i+1}|^q + \cdots + |\theta_{ip}|^q$. 假设每个节点的度是独立的, 则

$$\ln P(G = (V, E)) \propto \prod_{i=1}^{p} (\|\boldsymbol{\theta}_{-i}\|_q^q)^{-\alpha}.$$

基于 $\boldsymbol{X}$ 的 n 次观测 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, Gaussian 图模型的对数似然函数为

$$P(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \,|\, G) = \ln|\boldsymbol{\Theta}| - \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Sigma}),$$

其中, $|\cdot|$ 为矩阵的行列式, $\mathrm{tr}(\cdot)$ 为矩阵的迹, $\boldsymbol{S}$ 为样本协方差矩阵. 结合网络节点度的分布信息以及边的稀疏先验, 我们提出以下正则化模型:

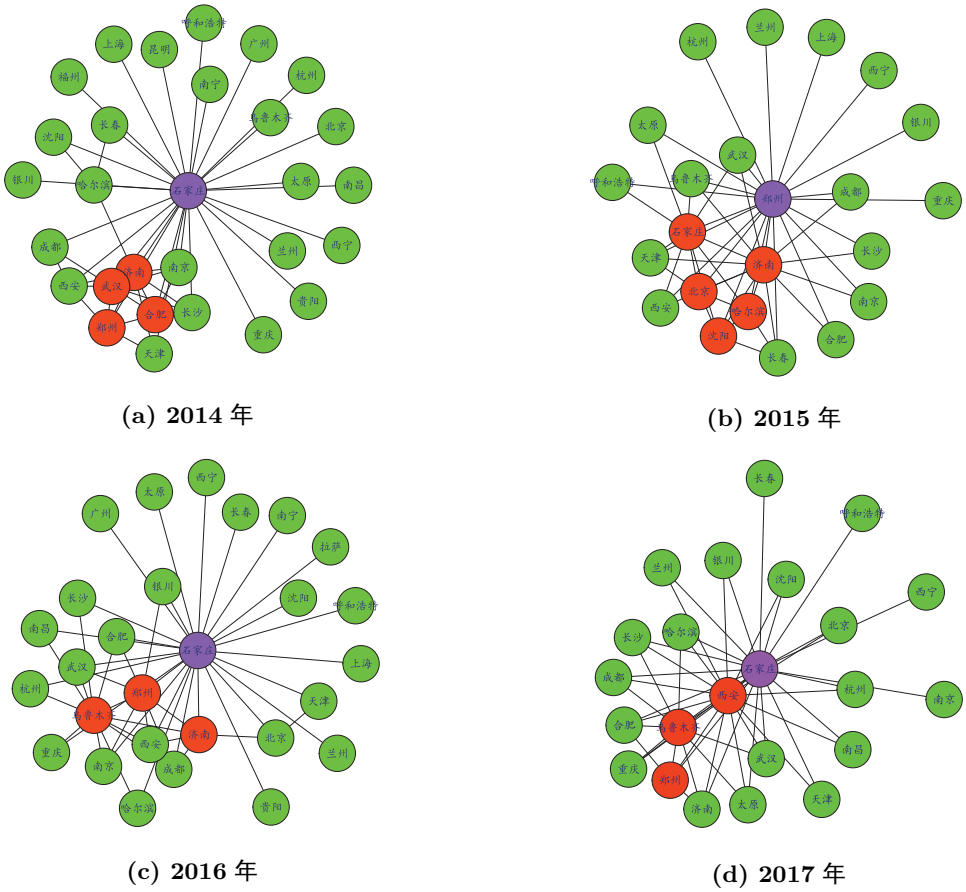
$$\max_{\boldsymbol{\Theta}} \ \ln|\boldsymbol{\Theta}| - \mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{S}) - \lambda \sum_{i=1}^{p} \ln(\|\boldsymbol{\theta}_{-i}\|_q^q + \epsilon_i), \tag{1}$$

其中, λ 为调控参数, $\|\boldsymbol{\theta}_{-i}\|_q^q = |\theta_{i1}|^q + \cdots + |\theta_{i,i-1}|^q + |\theta_{i,i+1}|^q + \cdots + |\theta_{ip}|^q$, $\epsilon_i > 0$ $(i = 1, ..., p)$ 保证 Ln 的指数部分大于 0. 正则化项为 Ln 型和 L_q 型惩罚函数的复合, 相当于引入网络节点度的分布信息以及边的稀疏先验. 模型 (1) 对应于一个非凸优化问题, 文献 [18] 中提出了重赋权迭代算法来求解 (1).

3.2 结果分析

以年为数据单元, 对每一年数据, 我们分别用模型 (1) 建立无标度图模型, 试图发现 $PM_{2.5}$ 污染的网络中影响力最大的城市. 需要注意的是, 模型 (1) 中 q 越靠近 0, 则 $\|\boldsymbol{\theta}_{-i}\|_q^q$ 越靠近 $\boldsymbol{\theta}_{-i}$ 的 l_0 拟范数, 从而越靠近节点 i 的度. 文献 [20] 研究了 $L_{1/2}$ 在 L_q $(0<q<1)$ 正则化中的代表性, 即 $L_{1/2}$ 比 L_q $(0.5<q<1)$ 更稀疏, 而 $L_{1/2}$ 与 L_q $(0<q<0.5)$ 压缩表示能力相当, 因此我们选取 $q=0.5$. 另外, 为了网络的可视化效果以及为了便于比较, 使用对应 120 条边的Graphical Lasso[13] 作为初值, 用两步重赋权迭代算法求解模型, 并最终选取 50 条边的网络.

所得 4 个网络如图 2 所示, 其中我们采用不同颜色来显示节点度的不同, 紫色和红色代表网络中度最大的 3 至 5 个节点 (Hubs), 特别地, 紫色表示网络中度最大的节点, 红色表示除度最大节点之外的度较大的若干节点. 其余节点我们用绿色表示.



(a) 2014 年 **(b) 2015 年**

(c) 2016 年 **(d) 2017 年**

图 2 $PM_{2.5}$ 污染的无标度网络. (a)、(b)、(c)、(d) 分别对应 2014、2015、2016、2017 年的无标度网络. 4 个网络中边数均为 50, 紫色和红色被用来表示网络中的中心点, 即节点度远大于平均的节点, 其余节点用绿色表示.

从图 2 我们可以看到, 用无标度图模型建立的网络均有明显的中心点, 说明 $PM_{2.5}$ 污染确实存在影响大的城市. 石家庄、郑州、石家庄、石家庄分别为 2014、2015、2016、2017 年全年 $PM_{2.5}$ 污染网络的度最大的点. 此外, 2014 年, 郑州、济南、合肥、武汉为边较多的城市; 2015 年, 济南、石家庄、北京、沈阳、哈尔滨为边较多的城市; 2016 年, 郑州、济南、乌鲁木齐为边较多的城市; 2017 年, 乌鲁木齐、西安、郑州为边较多的城市. 从图模型的原理来讲, 这些城市为与其他城市 $PM_{2.5}$ 污染相互影响较多的几个城市. 为了进一步验证此结果, 我们分析了 2014 至 2017 年全年各城市的 $PM_{2.5}$ 平均浓度排名 (由高到低), 见表 1. 从表 1 中, 我们看出, 图模型寻找的中心点与 $PM_{2.5}$ 污染较大的城市高度一致. 4 个网络的最中心点石家庄、郑州、石家庄、石家庄, 分别为 2014 至 2017 年 $PM_{2.5}$ 平均浓度最高的城市. 另外, 2014 年, 模型找出的中心点郑州、济南、合肥、武汉为 $PM_{2.5}$ 平均浓度排名前 7 的城市. 2015 年, 边较多的济南、石家庄、北京、沈阳、哈尔滨为 $PM_{2.5}$ 平均浓度前 6 的城市. 2016 年, 边较多的郑州、济南、乌鲁木齐为 $PM_{2.5}$ 浓度第 2、第 3、第 4 的城市. 2017 年, 边较多的乌鲁木齐、西安、郑州为 $PM_{2.5}$ 浓度第 2、第 3、第 4 的城市.

为说明算法稳健性, 我们分别基于 2014 年至 2017 年 $PM_{2.5}$ 污染的 4 个网络, 依次删除网络中度最大的节点后重新建立无标度网络. 为节省空间, 我们略去结果. 结果表明, 原网络中度第二大的节点成为新网络中度最大节点, 且这些网络的度最大的节点与 $PM_{2.5}$ 平均浓度由大到小的城市有一定的一致性. 从而说明无标度图模型可应用于污染源地区辨识, 且算法具有一定的稳健性.

以上分析说明, 与其余城市 $PM_{2.5}$ 互相影响最多的城市 (Hub) 往往为 $PM_{2.5}$ 浓度高的城市, 因此可以通过比较各城市 $PM_{2.5}$ 浓度来衡量各城市 $PM_{2.5}$ 污染的影响力. 由本节结果我们建议要重点治理 $PM_{2.5}$ 污染严重的包括石家庄、郑州、济南等城市所在区域.

§4. $PM_{2.5}$ 网络区块分析

本节我们基于分区块图模型[21] 构建 31 个省会城市的 $PM_{2.5}$ 污染所形成的网络结构, 从而研究各城市 $PM_{2.5}$ 污染之间的关系以及 $PM_{2.5}$ 的分区块结构. 我们首先简单介绍所使用的分块对角图模型, 然后给出将该方法应用于 $PM_{2.5}$ 数据的结果并进行结果分析.

4.1　分块对角图模型

第 3 节, 我们利用网络无标度的特征构建了 $PM_{2.5}$ 污染的网络, 发现了网络的中心点即污染影响力大的城市. 对于 $PM_{2.5}$ 网络, 除了具有中心点外, 由于地理结构、经济结构、气象等因素影响, 各城市间的 $PM_{2.5}$ 互相影响具有分区特征, 同一区块内部的城市 $PM_{2.5}$ 污染的相互影响强于不同区块的城市 $PM_{2.5}$ 污染的相互影响. 为了研究各城市的 $PM_{2.5}$ 污染所形成的网络的分区结构, 我们采用 Devijer 与 Gallopin[21] 所提的分块对角图模型, 其

所得结果对应的网络有明显的分区块特征, 即网络的边仅存在于区块内部, 而区块与区块之间无边相连.

下面我们叙述 Devijer 与 Gallopin 的方法. 简单来说, 其基本思想为两个步骤: 第一步基于正则化方法来寻找最佳划分; 第二步基于第一步划分结果限制在每一区块上, 分别使用 Graphical Lasso[13] 估计网络. 为了便于说明, 我们所使用符号与 Devijer 和 Gallopin[21] 保持一致. 假设 $\{\boldsymbol{x}_i\}_{i=1}^n$ 是从 $\mathrm{N}(\mathbf{0},\boldsymbol{\Sigma})$中抽取的 n 个样本. 设其样本协方差矩阵为 S. 对于阈值 λ, 定义如下由阈值 λ 截断之后的样本协方差矩阵,

$$\boldsymbol{E}_\lambda = [\mathbf{1}_{\{|S_{j,j'}|>\lambda\}}]_{1\leqslant j,j'\leqslant p}.$$

考虑划分集合

$$\mathscr{B}_\Lambda = (\boldsymbol{B}_\lambda)_{\lambda\in\Lambda},$$

其中, Λ 为 λ 的取值集合, $\boldsymbol{B}_\lambda$ 表示由 $\boldsymbol{E}_\lambda$ 所决定的划分. 这里需要说明的是对于一个矩阵, 划分描述了矩阵的分块对角结构, 即在块之外, 矩阵的所有元素为 0, 非零元素仅存在于块内部. 另外, 置换块内部的变量和重新排序块所对应的划分均不变. Devijver 与 Gallopin[21] 定义了如下正则化模型, 实现从划分集合 $\mathscr{B}_\Lambda$ 中选择最佳的划分,

$$\widehat{\boldsymbol{B}} = \mathop{\arg\min}_{\boldsymbol{B}\in\mathscr{B}_\Lambda}\Big\{ -\frac{1}{n}\sum_{i=1}^n \ln[\widehat{f}_{\boldsymbol{B}}(\boldsymbol{x}_i)] + \kappa\frac{\boldsymbol{D}_{\boldsymbol{B}}}{n}\Big\}, \tag{2}$$

其中, $\widehat{f}_{\boldsymbol{B}} = \mathrm{N}(\mathbf{0},\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{B}})$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{B}}$ 为真实划分为 $\boldsymbol{B}$ 时, $\boldsymbol{\Sigma}_{\boldsymbol{B}}$ 的最大似然估计, 也就是样本协方差矩阵限制在划分 $\boldsymbol{B}$ 之上的矩阵. $\boldsymbol{D}_{\boldsymbol{B}} = \sum\limits_{k=1}^K p_k(p_k-1)/2$ 为模型的维数, 即未知参数的个数, 其中, K 为划分所对应的块的个数, p_k 为第 k 个子块中变量个数. 以上即为 Devijver 与 Gallopin 针对 Graphical Lasso 的步骤一提出的改进方法. 在确定划分之后, 第二步与 Graphical Lasso 类似, 对于每个子块, 分别使用 Graphical Lasso 估计网络. 不同之处在于对于每个子块, Graphical Lasso 所对应的调控参数相同, 而 Devijver 与 Gallopin 的方法可以不同, 增加了模型的灵活性.

4.2 结果分析

各城市间 $\mathrm{PM}_{2.5}$ 污染的网络结构并不是恒定不变的, 而是随着时间动态变化的. 注意到我们这里研究 31 个城市, 因此考虑到 $\mathrm{PM}_{2.5}$ 污染数据的实时性, 以及避免样本太少导致图模型拟合结果精度不高, 我们选取 2014 年至 2017 年的每 31 天作为一个数据单元, 即样本集. 具体地, 2014 年 1 月 1 日至 1 月 31 日为第一个样本集, 2014 年 1 月 2 日至 2 月 1 日为第二个样本集, 依次类推. 这样, 2014 至 2017 年 1 450 天 (其中 11 天数据缺失) 共产生 1 420 个样本集. 对于每一个样本集, 我们使用上述分块对角模型拟合网络, 其中, 第一步中使用 Slope 选择调控参数, 第二步中使用 BIC 准则选择调控参数[21]. 最终得到 1 420 个分

区块网络, 其中, 块与块之间的城市无边相连, 边仅存在于区块内部城市之间. 为了节省空间, 我们将其中前 4 个网络作为示例呈现于图 3. 由于天气、工业等因素影响, 1 420 个网络中区块不尽相同, 每个网络的社区结构可以看作真实社区结构的一个带误差的观测. 为了得到 $PM_{2.5}$ 污染的区块的整体信息, 我们把这 1 420 个网络简单集成为一个网络, 即对全部网络的邻接矩阵求和得到集成后的加权的邻接矩阵, 所得矩阵对应一个无向图. 需要说明的是, 我们将每 31 天作为一个样本集并将每 31 天对应网络集成来源于样本平均的思想, 既利用了局部信息, 又可以得到区块的整体信息. 所得结果比单纯分析每 31 天对应网络和利用全部数据估计一个分区块网络更可靠、实用. 为了发现 31 个城市 $PM_{2.5}$ 污染的社区结构, 我们将 Modularity 方法 [22] 应用于集成后的邻接矩阵. 需要说明的是, 社区个数与社区发现的尺度相关, 一个大社区可能有几个小社区构成. 为了便于分析, 我们进行了 3 次社区发现, 得到以下社区结果, 并呈现于图 4.

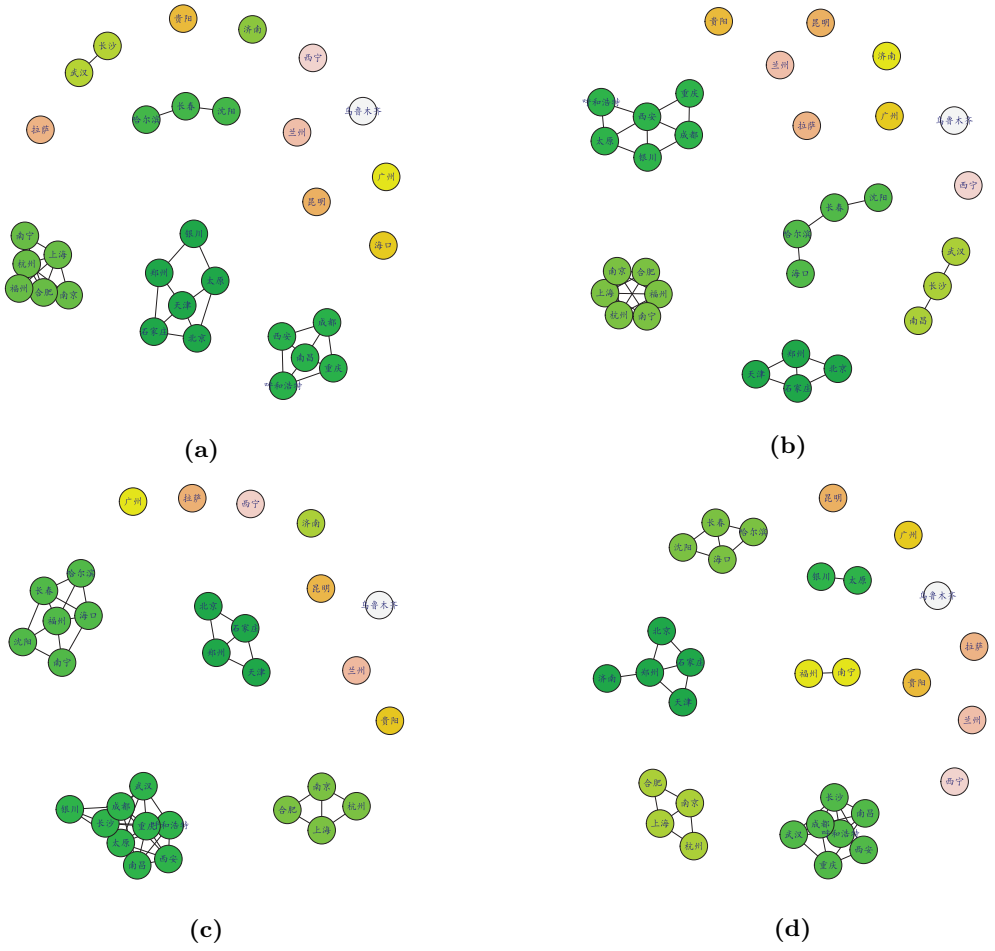


图 3　$PM_{2.5}$ 分区块网络. (a)、(b)、(c)、(d) 分别对应 1 420 个网络中的前 4 个网络. 对于每个网络, 用同种颜色表示处在同一社区的城市.

图 4　社区发现示意图. 其中, 用同种颜色表示同一社区内的城市.

北京、天津、石家庄、太原组成一个社区, 这 4 个城市地处中国北部. 该社区中, 北京和天津之间的边最多, 二者之间有边的网络占所有网络的 84.6%; 其次是北京和石家庄、天津和石家庄, 二者之间有边的比例分别为 79.5%、73.7%. 这可以解释为北京、天津、石家庄三者之间距离相近, 故 $PM_{2.5}$ 相互影响程度较大. 和太原连接最多的城市是石家庄, 二者之间有边的网络占所有网络的 61.5%. 另外, 石家庄是该社区中度最大的城市, 这与石家庄处于该社区地理中心的事实吻合.

郑州、济南、西安组成一个社区, 这 3 个城市地处中国中部. 该社区中郑州和济南之间的边最多, 二者之间有边的网络占所有网络的 46.3%; 其次是郑州和西安, 二者之间有边的网络占所有网络的比例是 28.0%. 另外, 郑州为该社区中度最大的城市, 而郑州在地理位置上正好位于济南和西安之间.

哈尔滨、长春、沈阳组成一个社区, 这 3 个城市地处中国东北部. 该社区中, 长春和哈尔滨之间的边最多, 二者之间有边的网络占所有网络的 75.6%; 其次是长春和沈阳, 二者之间有边的网络占所有网络的比例是 69.3%. 此外, 长春为该社区中度最大的城市, 而长春在地理位置上正好位于哈尔滨和沈阳之间.

兰州、西宁、银川、呼和浩特组成一个社区, 这 4 个城市地处中国西北部. 该社区中, 西宁和兰州之间的边最多, 二者之间有边的网络占所有网络的 31.3%; 其次是银川和乌鲁木齐, 二者之间有边的网络占所有网络的 29.7%.

上海、南京、杭州、合肥组成一个社区, 这 4 个城市地处中国中部和东部. 该社区中, 南京和合肥之间的边最多, 二者之间有边的网络占所有网络的 93.2%; 其次是南京和杭州, 二者之间有边的比例是 72.0%.

南昌、武汉、长沙组成一个社区, 这 3 个城市两两相邻, 地处中国中部. 该社区中, 长

沙和武汉之间的边最多, 二者之间有边的网络占所有网络的 73.8%; 其次是长沙和南昌, 二者之间有边的比例是 69.3%.

成都、重庆、贵阳、昆明、拉萨组成一个社区, 这 5 个城市地处中国西南部. 该社区中, 重庆和成都之间的边最多, 二者之间有边的网络占所有网络的 44.8%; 其次是重庆与贵阳, 二者之间有边比例是 28.2%. 这 5 个城市中, 和拉萨连接最多的城市是重庆, 但仅有 45 条边, 占比 3.2%, 因此将拉萨视为独立的社区更合理.

广州、南宁、海口、福州、乌鲁木齐组成一个社区, 其中前 4 个城市地处中国南部. 该社区中, 广州和南宁之间的边最多, 二者之间有边的网络占所有网络的 41.8%; 其次是南宁与海口, 二者之间有边比例是 32.5%. 值得注意的是, 这 5 个城市中, 乌鲁木齐连接最多的城市是南宁, 但仅有 55 条边, 占比仅有 3.9%, 即在大部分网络中, 乌鲁木齐都是该社区中的孤立点. 结合乌鲁木齐的地理位置和污染数据, 将乌鲁木齐视为独立的社区更为合理.

基于以上分析, 我们得出 $PM_{2.5}$ 污染存在明显的社区特征, 从统计意义来讲, 社区内部城市之间 $PM_{2.5}$ 污染的相互影响远大于不同社区城市之间 $PM_{2.5}$ 污染的相互影响. 而且城市所处的社区与城市地理位置高度一致, 地理位置相近的城市更容易被分到同一社区, 社区内部位于地理中心的城市容易成为与其余社区内城市 $PM_{2.5}$ 相互影响最多的城市. 基于本节结论, 我们建议治理 $PM_{2.5}$ 污染需要分区治理, 按区研究 $PM_{2.5}$ 形成机理, 制定区域治理特色方案.

§5. 结　　论

本文关注于中国 31 省会城市 $PM_{2.5}$ 污染的网络结构学习问题, 利用图模型研究成果分别进行了 $PM_{2.5}$ 污染网络的中心点分析、区块分析. 结果表明: $PM_{2.5}$ 污染严重的城市同时也是 $PM_{2.5}$ 污染网络的中心点, 即网络中影响力最大的点; $PM_{2.5}$ 污染网络存在明显的区块特征, 地理位置相近的城市常常处于 $PM_{2.5}$ 污染网络的同一区块. 基于网络分析结果, 我们给出了要重点关注 $PM_{2.5}$ 污染严重区域, 同时要分区分块治理的建议. 另外, 应重点开展西部 $PM_{2.5}$ 污染治理和原因分析.

沿本文工作可进一步进行如下研究:

1) 我们只考虑了中国 31 个省会城市, 而非省会城市的 $PM_{2.5}$ 污染同样严重, 而且它们的加入可能会影响整个网络的结构. 本文的研究方法及模型可进一步应用于更多城市 $PM_{2.5}$ 网络结构的研究. 另外, 我们只考虑了日平均 $PM_{2.5}$ 浓度, 未来可将其细化为小时浓度;

2) 在研究 $PM_{2.5}$ 网络的中心点和社区结构时, 我们仅考虑了各城市的 $PM_{2.5}$ 浓度. 为了进一步用数据分析探究中心点和社区结构产生的原因, 可引入各城市的其他属性信息, 如地理位置、GDP、人口等等. 近期关于图模型的研究成果 (如文献 [23]) 为引入属性信息提供了可操作性.

3) 本文关注了 $PM_{2.5}$ 污染的无向网络. 为了进一步确定城市之间 $PM_{2.5}$ 污染的因果关系, 可利用有向图模型 (如文献 [24]) 等相关因果分析方法开展研究.

4) $PM_{2.5}$ 浓度会受到气象、温度、湿度等环境因素的影响. 进一步可利用经属性调整的图模型 (如文献 [25]) 构建 $PM_{2.5}$ 污染网络结构, 研究去掉气象、温度、湿度等因素影响之后是否会大大减少 $PM_{2.5}$ 污染网络中的边.

以上均在我们目前研究中.

参 考 文 献

[1] 陈迪, 张永文, 铁学熙. 我国 $PM_{2.5}$ 空间关联性的探讨 [J]. 中国科学: 物理学 力学 天文学, 2017, **47(2)**: 020501(1–8).

[2] 周雪明, 郑乃嘉, 李英红, 等. 2011∼2012 北京大气 $PM_{2.5}$ 中重金属的污染特征与来源分析 [J]. 环境科学, 2017, **38(10)**: 4054–4060.

[3] 刘华军, 杜广杰. 中国雾霾污染的空间关联研究 [J]. 统计研究, 2018, **35(4)**: 3–15.

[4] LIANG X, ZOU T, GUO B, et al. Assessing Beijing's $PM_{2.5}$ pollution: severity, weather impact, APEC and winter heating [J]. *Proc R Soc A*, 2015, **471(2182)**: 20150257(1–20).

[5] LIANG X, LI S, ZHANG S Y, et al. $PM_{2.5}$ data reliability, consistency, and air quality assessment in five Chinese cities [J]. *JGR: Atmospheres*, 2016, **121(17)**: 10220–10236.

[6] EDWARDS D. *Introduction to Graphical Modelling* [M]. 2nd ed. New York: Springer, 2000.

[7] LAURITZEN S L. *Graphical Models* [M]. New York: Oxford University Press, 1996.

[8] PEARL J. *Causality: Models, Reasoning, and Inference* [M]. Cambridge, UK: Cambridge University Press, 2000.

[9] FRIEDMAN N. Inferring cellular networks using probabilistic graphical models [J]. *Science*, 2004, **303(5659)**: 799–805.

[10] MEINSHAUSEN N, BÜHLMANN P. High-dimensional graphs and variable selection with the Lasso [J]. *Ann Statist*, 2006, **34(3)**: 1436–1462.

[11] YUAN M, LIN Y. Model selection and estimation in the Gaussian graphical model [J]. *Biometrika*, 2007, **94(1)**: 19–35.

[12] PENG J, WANG P, ZHOU N F, et al. Partial correlation estimation by joint sparse regression models [J]. *J Amer Statist Assoc*, 2009, **104(486)**: 735–746.

[13] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Sparse inverse covariance estimation with the graphical lasso [J]. *Biostatistics*, 2008, **9(3)**: 432–441.

[14] WITTEN D M, FRIEDMAN J H, SIMON N. New insights and faster computations for the graphical lasso [J]. *J Comput Graph Statist*, 2011, **20(4)**: 892–900.

[15] MAZUMDER R, HASTIE T. Exact covariance thresholding into connected components for large-scale graphical lasso [J]. *J Mach Learn Res*, 2012, **13**: 781–794.

[16] ZHAO T, LIU H, ROEDER K, et al. The huge package for high-dimensional undirected graph estimation in R [J]. *J Mach Learn Res*, 2012, **13**: 1059–1062.

[17] DANAHER P, WANG P, WITTEN D M. The joint graphical lasso for inverse covariance estimation across multiple classes [J]. *J R Stat Soc Ser B Stat Methodol*, 2014, **76(2)**: 373–397.

[18] 郭骁, 张海, 吴奖伦. 基于无标度先验的图模型结构学习 [J]. 中国科学: 信息科学, 2016, **46(7)**: 870–882.

[19] ALBERT R, BARABÁSI A L. Statistical mechanics of complex networks [J]. *Rev Modern Phys*, 2002, **74(1)**: 47–97.

[20] XU Z B, ZHANG H, WANG Y, et al. $L_{1/2}$ regularization [J]. *Sci China Inf Sci*, 2010, **53(6)**: 1159–1169.

[21] DEVIJVER E, GALLOPIN M. Block-diagonal covariance selection for high-dimensional Gaussian graphical models [J]. *J Amer Statist Assoc*, 2018, **113(521)**: 306–314.

[22] BRANDES U, DELLING D, GAERTLER M, et al. On modularity clustering [J]. *IEEE T Knowl Data En*, 2008, **20(2)**: 172–188.

[23] CHENG J, LEVINA E, WANG P, et al. A sparse ising model with covariates [J]. *Biometrics*, 2014, **70(4)**: 943–953.

[24] SHOJAIE A, MICHAILIDIS G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs [J]. *Biometrika*, 2010, **97(3)**: 519–538.

[25] CHEN M J, REN Z, ZHAO H Y, et al. Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model [J]. *J Amer Statist Assoc*, 2016, **111(513)**: 394–406.

Structure Learning of $PM_{2.5}$ Distribution Using Sparse Graphical Models

ZHANG Hai GUO Xiao REN Sa DENG Yajing

(*School of Mathematics, Northwest University, Xi'an, 710127, China*)

Abstract: We consider the structure learning problem of the $PM_{2.5}$ pollution data over 31 provincial capitals in China. Specifically, we make use of the graphical model tools to study the hubs and the community structures of the $PM_{2.5}$ pollution networks. The results show that the hubs in the $PM_{2.5}$ pollution networks are always seriously polluted cities, and the $PM_{2.5}$ pollution networks have significant community structures which consist of cities which in some sense can be regarded as blocks with similar cause of pollution. In view of the results, we suggest that the government should strengthen the effort to treat the seriously polluted areas and western China areas. Moreover, the management of the $PM_{2.5}$ pollution should be region-dependent.

Keywords: graphical model; network; community; scale-free; $PM_{2.5}$

2010 Mathematics Subject Classification: 62-07; 62H12