中央财经大学
Central University of Finance and Economics

# 本科生毕业论文（设计）

## 我国 PM2.5 浓度时空变化特征及其影响因素研究—基于时空自回归模型


学生姓名：　　刘拓臻

学　　号：　　2017310846

学　　院：　　统计与数学学院

专　　业：　　应用统计学

指导教师：　　潘蕊


日　　期：　　2021 年 4 月 15 日

# 内 容 摘 要

本文针对大气数据，提出了一种时空污染物自回归模型 (MSAR) 刻画 PM2.5 浓度变化规律，在考虑了 PM2.5 浓度的时间相关性、空间相关性的基础上，创新性地利用 PM2.5 与其他污染物协变量（PM10、SO2 等）的相关性建模，并选取我国石家庄市为例实证分析，结果表明，PM2.5 浓度有很强的时间相关性、空间相关性、污染物相关性，并且变化规律会受到其他气象变量的影响。此外，不同季节的模型拟合结果存在一定差异，反映出 PM2.5 变化规律的季节特征。研究结果有助于我国开展大气污染防治治理工作，在大气污染物浓度预测、大气污染物浓度异常值识别、大气污染物缺失值插补等领域有一定的参考价值。

**关键词：** PM2.5 数据　时空模型　线性回归

# ABSTRACT

The purpose of this paper is to construct a spatio-temporal-pollutant autoregression model (MSAR) to describe the trend of PM2.5 concentration based on the atmospheric data. Considering the temporal and spatial correlation of PM2.5 concentration, this paper innovatively use the correlation between PM2.5 and other pollutants covariates (PM10, SO2, etc.). We select Shijiazhuang City as an example for empirical analysis. The results show that PM2.5 concentration has strong temporal correlation, spatial correlation and pollutants correlation, and the trend of PM2.5 concentration is affected by other meteorological variables. In addition, the modelling results of different seasons are different, which reflects the seasonal pattern of PM2.5 concentration. The research results are useful to the prevention and control of air pollution in China, and have certain reference value in the fields of air pollutant concentration prediction, abnormal value detection and missing value interpolation.

**KEY WORDS:** PM2.5 Data　Spatio-temporal Model　Liner Regression

# 目　录

# 我国 PM2.5 浓度时空变化特征及其影响因素研究
# —基于时空自回归模型

## 一、研究背景

### （一）选题背景

随着中国经济的飞速发展，工业化进程不断加快，环境污染问题也越来越严重，2013 年以来几次严重的雾霾污染事件引起了公众的广泛关注。据耶鲁大学发布的《2020 年全球环境绩效指数报告》显示，在全球 180 个国家的空气质量排名中，中国以 37.3 分位居第 120 位。其中，PM2.5 指数超标（超过世卫组织标准）的分数全球排名倒数第三。此外，在环保部发布的《2019 中国生态环境状况公报》中提及，2019 年，全国 337 个地级及以上城市中，有 180 个城市的空气质量超标，占比超过 50%，且在这 338 个地级及以上城市中，全年平均超标天数比例接近 20%，以 PM2.5 为首要污染物的超标天数占总超标天数的 45%。据估计，每年全国 17%的死亡是由严重的空气污染造成的，环境户外空气污染致死的代价超过国内 GDP 的 8%。由此可见，中国的大气污染问题异常严重。

PM2.5 污染问题尤其受人关注。PM2.5 是指直径在 2.5 微米或更小的颗粒物，会对人民的健康、国家的经济造成巨大影响。首先，长时间生活在高浓度 PM2.5 的环境下尤其会增加心血管和呼吸系统疾病以及癌症的风险，PM2.5 由于其粒径小、表面积大，为一些细菌、病毒、重金属及致癌物质提供了良好的载体；其次，PM2.5 等微粒物会散射和吸收阳光，使得可见性恶化并进一步导致雾霾，同时，累积的微粒物会阻碍光合作用，导致农业减产；此外，PM2.5 污染会导致当地气候条件发生变化。研究表明，颗粒物可能改变气溶胶和云的性质，增强城市地区的热岛现象；最后，为了控制空气污染恶化而采取的极端环境保护措施会导致高能耗第二产业短期内减产、停产，企业面临生存危机。综合以上所有负面影响，PM2.5 污染会对人民对健康带来极大危害，并造成巨大的直接或间接经济损失。

习近平总书记在十九大报告中明确指出要“满足人民日益增长的优美生态环境需要”，要高度重视解决损害群众健康的突出环境问题，持续实施大气污染防治行动，明显改善环境质量。监测和控制 PM2.5 的浓度水平成为一个至关重要的问题。为了应对高浓度的 PM2.5 污染，我国政府已经采取了一系列强有力的

行动。早在 2013 年，中国国务院发布了《大气污染防治行动计划》。在这个计划的指导下，我国已在全国所有地级市建立了共 1700 余个监测站，并配备了远程质量控制系统。自 2015 年起，各站点 PM2.5 实时数据在互联网上公开，中国已建成发展中国家中最大的大气质量监测网络。此外，PM2.5 污染控制还被纳入政府绩效考核体系，污染水平的高低直接影响到地方政府资金与官员晋升，有效促进了各地方政府采取行动治理大气污染。

然而，我国目前的大气质量监测体系还存在很多问题，例如站点的分布不均，数据的缺失率过高，数据的异常值过多。更令人堪忧的是，到目前为止，我国已至少发生了两起严重的 PM2.5 数据欺诈案件。2016 年，西安发生首例环境空气自动监测数据造假案。车站工作人员故意堵塞检测设备，因此与监控系统发生冲突。7 名被告都被判处至少 1 年监禁。2018 年，临汾市某工作人员篡改数据系统，伪造数据报告，收到生态环境部通报。因此，建立强有力的数据检测体系是十分必要的，为建立有效的数据检测体系，就需要对大气污染物的变化规律及影响因素进行分析。

（二）文献综述

在大气污染物变化规律的研究中，国内外已有部分学者尝试利用统计方法建模分析，主要的研究方法可归为三类：基于时间序列分析、基于空间建模分析、基于其他气象解释变量分析。

首先，在时间序列分析领域，Kübra Ağaç等人采用季节性自回归综合移动平均模型（SARIMA）的时间序列方法，对伊斯坦布尔 PM10 日浓度进行了模拟和预测，采用奇异谱分析（SSA）的间隙填充法对浓度缺失值进行插补。Wang 等人利用向量自回归模型（VAR）对洛杉矶长滩地区月度 PM2.5 浓度时间序列进行了建模，探讨了 PM2.5 浓度随时间变化规律以及当月 PM2.5 浓度和交通量以及包括风速、温度、土壤温度、露点在内的气象协变量之间的关联。

其次，在空间建模分析领域，Michel Bobbia 等人利用增量克拉金方法研究法国诺曼底 PM10 浓度的空间变化规律，并应用于异常点识别。Hone-Jay Chu 等人有限混合分布模型（FMDM）对台湾省 PM10、PM2.5 浓度变化进行建模，并指出变化规律取决于天气状况和局部地区土地利用和排放格局的空间分布。张亮林等人利用 2017 年京津冀地区气溶胶光学厚度(AOD)遥感数据、GEOF 气象格网数据以及 PM2.5 污染物实测站点监测数据,构建多因子的地理加权回归模型（GWR）,对 PM2.5 污染物浓度进行了模拟估算,并对 PM2.5 污染物浓度的空间

布局及季节演化特征进行分析。

最后，在基于其他气象解释变量分析领域，Batterman 等人利用带有气象变量的自回归模型对北京市 2013–2015 年 PM2.5 日浓度变化规律进行拟合，结果表明，PM2.5 日浓度与大气压力、相对湿度、日照时数、地表温度和环境温度、降水量和清除系数、风向等气象变量具有自相关关系，气象变量和自回归项共解释了 PM2.5 水平 60%以上的变化。陈兵红等人采用随机森林模型分析浙江省 2014—2019 年 PM2.5 浓度变化的影响因子,结果表明日最低地表气温(MI-GST)、日最低气压（MI-PRS）、日降水量（PRE）等 15 个因子对 PM2.5 浓度影响显著。

（三）研究价值及创新点

本文在前人的研究基础上，尝试对 PM2.5 浓度建立多元自回归模型，在考虑了 PM2.5 浓度的时间相关性、空间相关性的基础上，创新性地加入了污染物相关性，即利用了 PM2.5 与其他污染物（PM10、SO2 等）协变量的相关性建模，最终提出了一种时空污染物自回归模型（MSAR）刻画 PM2.5 浓度变化规律，并比较不同季节的模型拟合差异。

石家庄市是我国北部的重要工业城市，也是我国受大气污染最严重的城市之一。本文选取石家庄市大气数据为例，研究我国 PM2.5 浓度时空变化特征及其影响因素，研究结果有助于我国开展大气污染防治治理工作，在大气污染物浓度预测、大气污染物浓度异常值识别、大气污染物缺失值插补等领域有一定的参考价值。

## 二、数据说明

### （一）站点说明

本文使用的数据为河北省石家庄市 6 个国控气象站点(西北水源、人民会堂、职工医院、西南高教、世纪公园、高新区）记录的大气数据，6 个站点的位置分布如图 1 所示，西北水源站点位于石家庄市的西北部，其他 5 个站点分布在石家庄市城区中心。

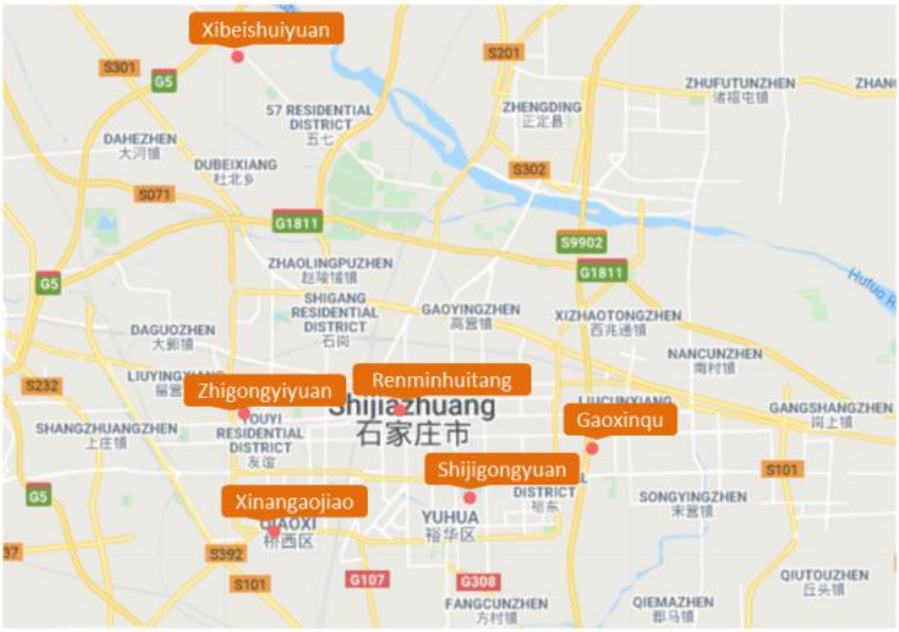


图 1：石家庄市 6 站点位置分布图

6 个站点之间两两的距离矩阵如表 1 所示，可以看出西北水源与其他站点之间的距离较远，均在 10km 以上，人民会堂处于中心位置，与其他站点距离相对较小。

表 1：石家庄市 6 站点距离矩阵

| | 高新区 | 人民会堂 | 世纪公园 | 西北水源 | 西南高教 | 职工医院 |
|---|---|---|---|---|---|---|
| 高新区 | 0 | 7.41 | 5.00 | 19.97 | 12.40 | 13.20 |
| 人民会堂 | 7.41 | 0 | 4.21 | 14.71 | 6.56 | 5.87 |
| 世纪公园 | 5.00 | 4.21 | 0 | 18.83 | 7.47 | 9.07 |
| 西北水源 | 19.97 | 14.71 | 18.83 | 0 | 17.98 | 13.50 |
| 西南高教 | 12.40 | 6.56 | 7.47 | 17.98 | 0 | 4.57 |
| 职工医院 | 13.20 | 5.87 | 9.07 | 13.50 | 4.57 | 0 |

## （二）变量说明

数据集共包括 6 个污染物变量、3 个气象变量，时间跨度为 2014 年 1 月至 2017 年 2 月，6 个站点共计 166,320 条数据，具体变量信息如表 2 所示。

表 2：变量说明表

| | 变量名称 | 单位 | 取值范围 | 缺失率(%) |
|---|---|---|---|---|
| 污染物变量 | PM2.5 浓度 | $\mu g/m^3$ | 2 ~ 990 | 2.00 |
| | PM10 浓度 | $\mu g/m^3$ | 2 ~ 1168 | 2.19 |
| | CO 浓度 | $\mu g/m^3$ | 0 ~ 30 | 2.04 |
| | NO2 浓度 | $\mu g/m^3$ | 2 ~ 300 | 3.45 |
| | SO2 浓度 | $\mu g/m^3$ | 3 ~ 689 | 5.63 |
| | O3 浓度 | $\mu g/m^3$ | 1 ~ 860 | 3.49 |
| 气象变量 | 气压 | $hPa$ | 981 ~ 1033 | 0.76 |
| | 气温 | ℃ | -14 ~ 42 | 0.83 |
| | 风速 | $m/s$ | 0 ~ 182 | 0.63 |

从表 2 可以看出，原始数据集中存在一定的缺失情况，其中变量 SO2 浓度的缺失率最高，高达 5.63%，变量风速的缺失率最低，仅为 0.63%。以 PM2.5 浓度为例具体查看各年各站点数据缺失率，从图 2 可以看出，2014 年到 2017 年，PM2.5 浓度的缺失率大体上有着递增的趋势，尤其是世纪公园站点，2017 年的缺失率（9.75%）明显高于其他。

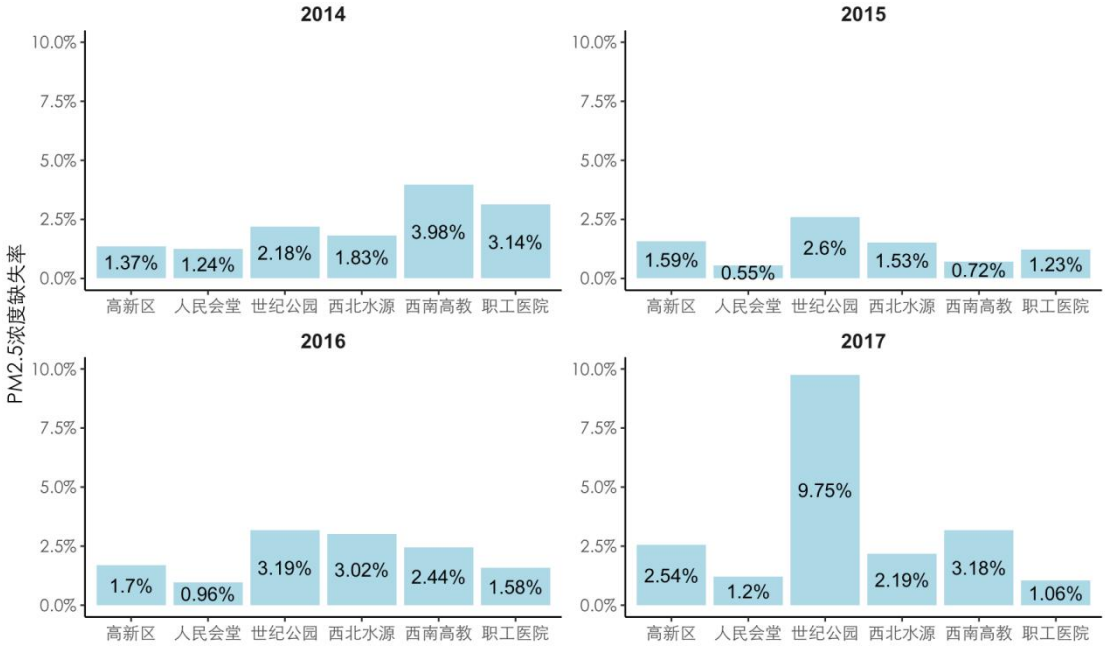


图 2：石家庄市各站点分年份 PM2.5 浓度缺失率柱状图

# 三、描述性分析

本文试图通过回归模型，研究 PM2.5 浓度的变化规律及影响因素，首先进行描述性分析，初步判断 PM2.5 浓度与各潜在因素之间的关联，为后续建模做铺垫。

（一）PM2.5 浓度总体情况

图 3 为 2014-2017 每一年各个站点在各个月份的 PM2.5 平均浓度, 可以看出, PM2.5 的浓度变化规律呈现出明显的季节特征，冬季（12 月、1 月、2 月）的平均浓度最高，夏季（6 月、7 月、8 月）的平均浓度最低。此外，不同站点之间平均浓度差异较小。

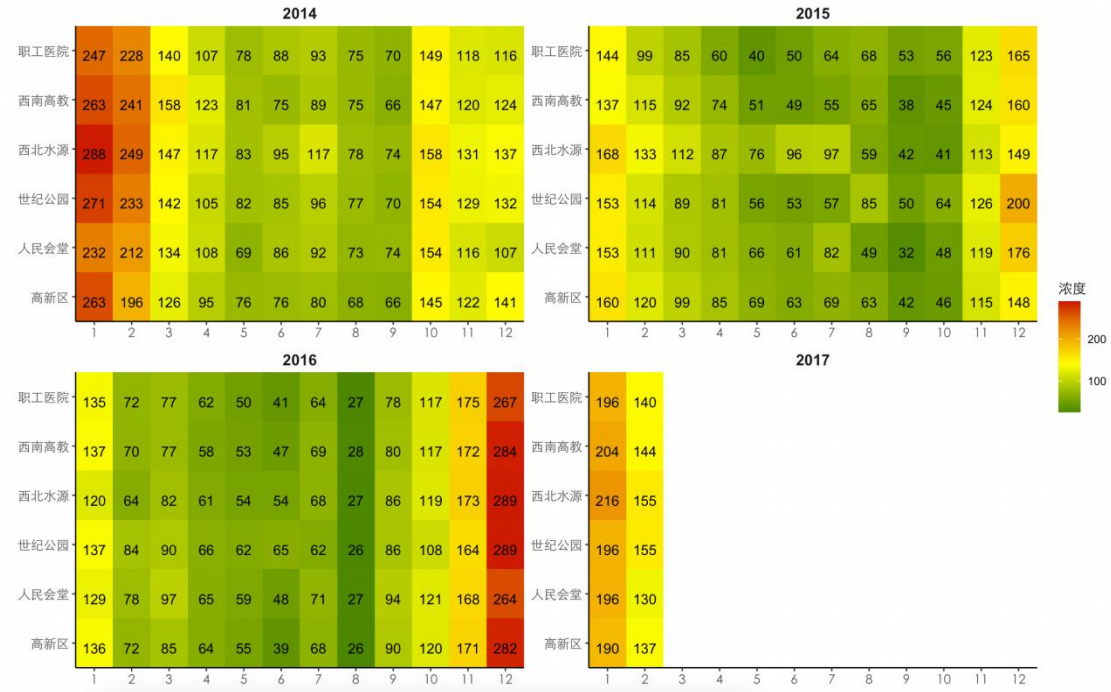


图 3：石家庄市各站点分月份 PM2.5 浓度热图

（二）PM2.5 浓度的时间相关性

图 4（a）为 2014-2017 每一年各个站点当期 PM2.5 浓度与自身上一期浓度的散点图，可以看出，当期 PM2.5 浓度与上一期 PM2.5 浓度呈现很高的正相关性，经计算得知，两者的相关系数高达 0.97，说明 PM2.5 浓度有很强的时间相关性。此外，以站点人民会堂为例，计算 PM2.5 浓度的自相关系数和偏自相关系数，如图 4（b）、图 4（c）所示，可以看出自相关系数呈现拖尾特征，偏自

相关系数呈现一阶截尾特征。

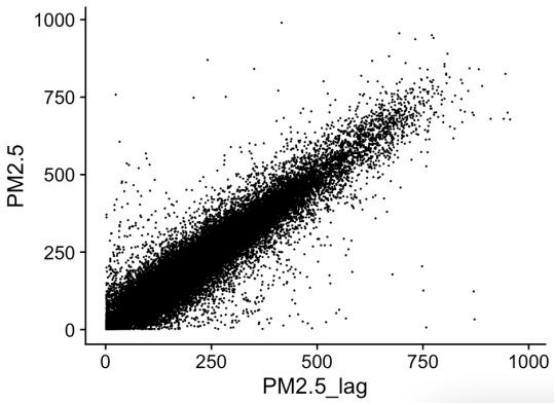


图 4 (a)：当期 PM2.5 浓度与滞后一期 PM2.5 浓度的散点图

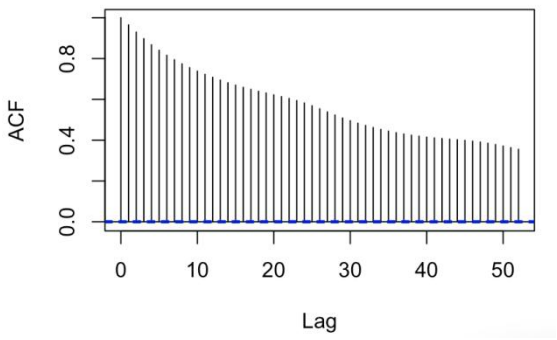


图 4 (b)：PM2.5 浓度自相关系数图

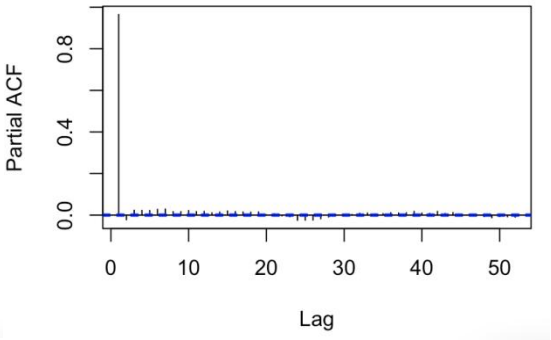


图 4 (c)：PM2.5 浓度偏自相关系数图

### (三) PM2.5 浓度的空间相关性

对于每个确定的目标站点 i，定义 t 时刻周围“邻居”站点的平均浓度为各个站点浓度的加权平均，权重与站点之间的距离成反比，即

$$PM2.5\_neighbor_{i,t} = \sum_{i \neq j} PM2.5_{j,t} \times weight_{i,j}\ ，其中\ weight_{i,j} = \frac{1/D_{i,j}}{\sum_{i \neq j}(1/D_{i,j})} \quad (1).$$

图 5 为 2014-2017 年每一年各个站点当期 PM2.5 浓度与当期周围“邻居”站点 PM2.5 浓度的散点图。如图 5 所示，PM2.5 浓度与其他邻居站点 PM2.5 浓度之间呈现较强的正相关性，经计算得知，两者的相关系数为 0.95，说明 PM2.5 浓度有很强的空间相关性。

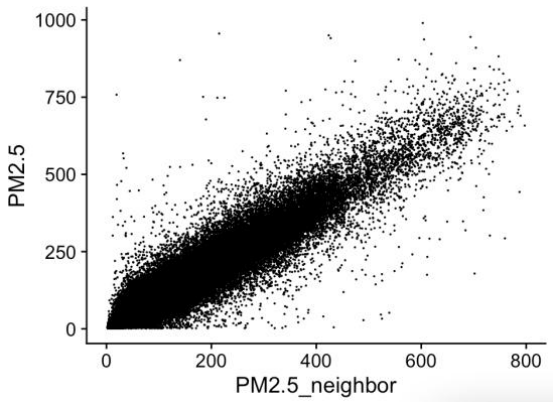


图 5：当期 PM2.5 浓度与当期周围“邻居”站点 PM2.5 浓度的散点图

（四）PM2.5 浓度的污染物相关性

图 6 为 2014-2017 年每一年各个站点当期 PM2.5 浓度与当期其他污染物浓度的散点图，如图 6 所示，PM2.5 浓度与当期 PM10 浓度呈现很强的正相关性，两者的相关系数高达 0.93，PM2.5 浓度与当期 SO2、NO2 之间也有一定的正相关性，相关系数分别为 0.52 和 0.65。与其他污染物相反，O3 与 PM2.5 之间呈现出较弱的负相关性，经计算，两者的相关系数为-0.23。

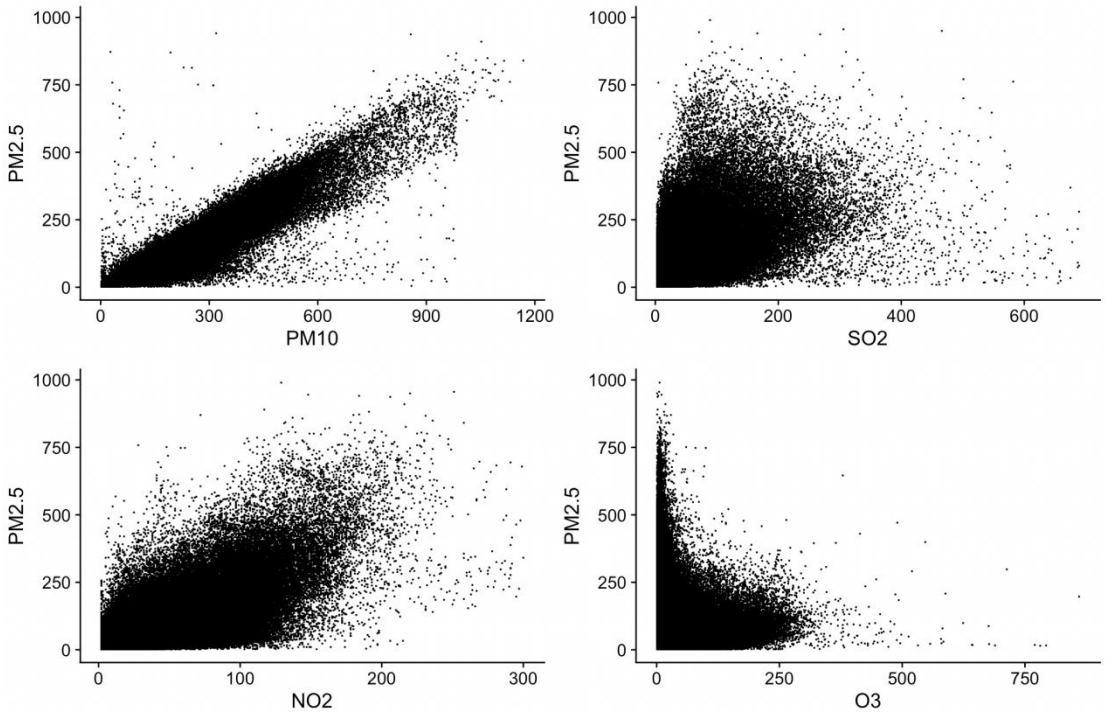


图 6：当期 PM2.5 浓度与当期其他污染物浓度的散点图

（五）PM2.5 与其他气象变量的关系

首先，对每个气象变量按照样本分位数划分档位，定义数据小于 25%分位数为“低”档位，大于 25%分位数但小于中位数为“中低”档位，大于中位数但小于 75%分位数为“中高”档位，大于 75%分位数为“高”档位。

图 7（左上）为不同气压档位下 PM2.5 浓度的分布图，可以看出，不同气压档位下，PM2.5 浓度呈现相似的分布，随着气压的升高，PM2.5 浓度的平均水平有上升趋势。图 7（右上）为不同气温档位下 PM2.5 浓度的分布图，可以看出，随着气温的升高，PM2.5 浓度的平均水平有下降趋势。图 7（下）为不同风速档位下 PM2.5 浓度的分布图，可以看出，随着风速的升高，PM2.5 浓度的平均水平有下降趋势。

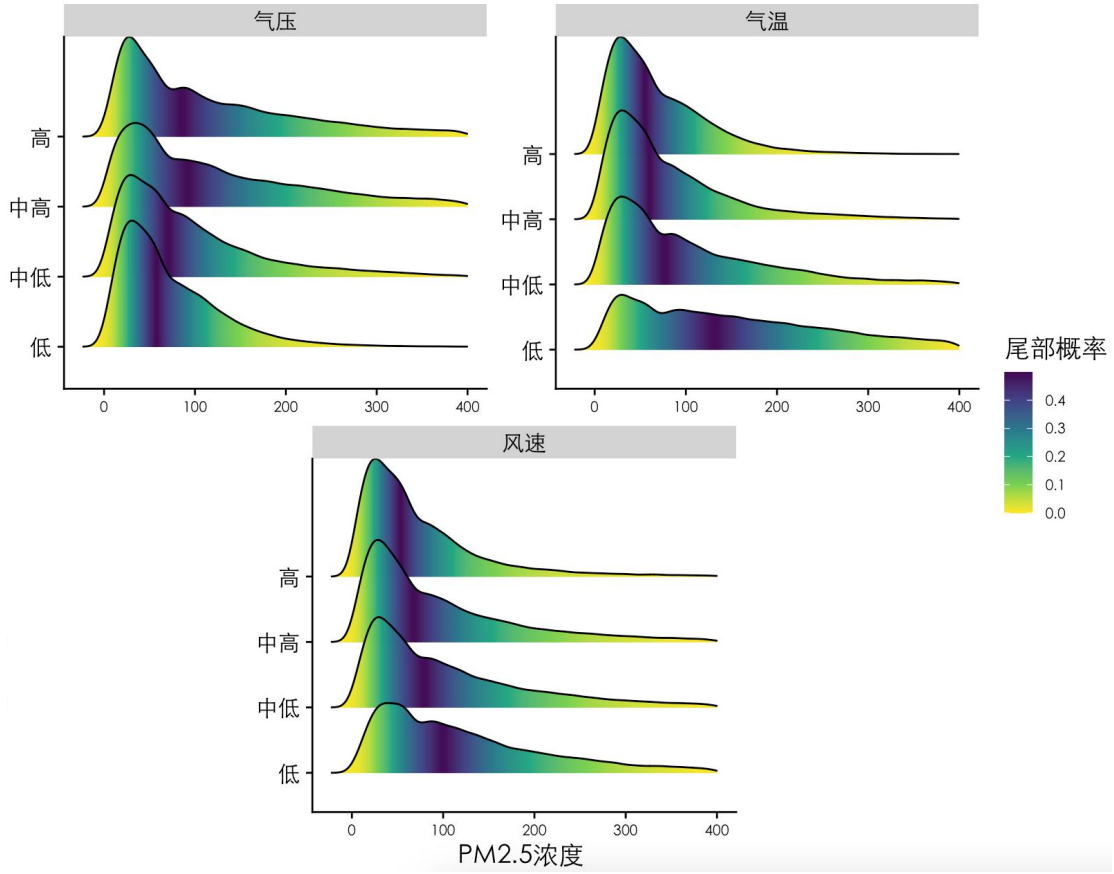


图 7：不同档位的气象变量下 PM2.5 浓度的分布图

此外，以每个站点每一年为组，计算在不同风速档位下的 PM2.5 一阶自相关系数，如图 8 所示，随着风速的增加，PM2.5 的时间相关性逐渐减弱。

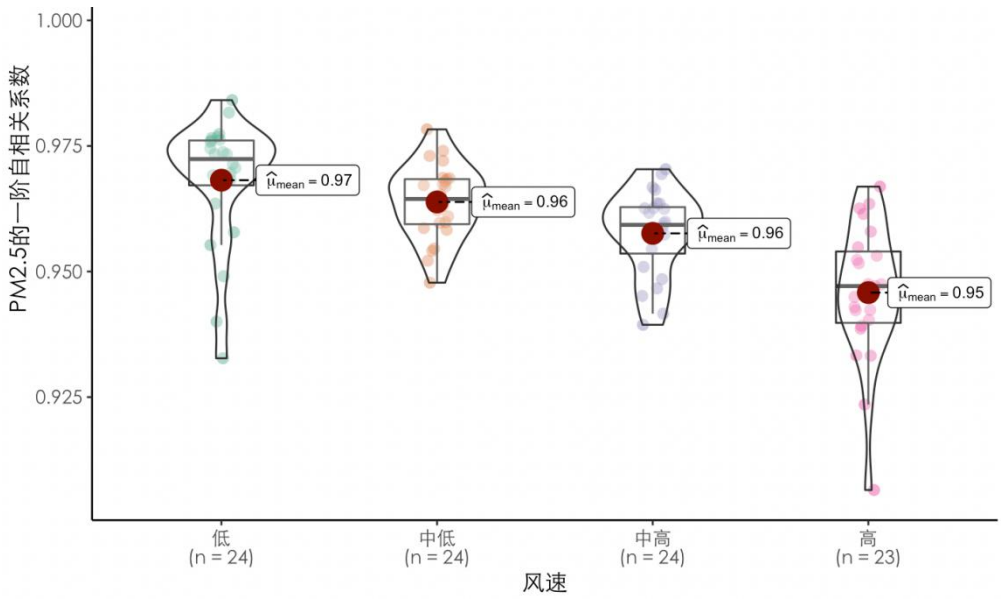


图 8：不同档位的风速下 PM2.5 浓度一阶自相关系数的箱线图

综上，通过对数据的描述性分析，可以发现 PM2.5 浓度有着很强的时间相关性、空间相关性以及污染物相关性。除此之外，PM2.5 浓度的变化规律还受到一些其他气象变量的影响，描述性分析的结果为后续建模分析打下基础。

## 四、模型建立

为了深入挖掘 PM2.5 浓度的变化规律及影响因素，参考描述性分析中的结果，本文将建立时间自回归模型、时空自回归模型、时空污染物自回归模型对 PM2.5 浓度变化规律进行拟合，并不同季节下时空污染物回归模型结果的差异。

（一）时间自回归模型（AR）

参考描述性分析中 PM2.5 浓度时间序列的自相关系数、偏自相关系数特征，定义时间自回归模型为

$$PM2.5_{s,t} = \beta_0 + \beta_1 \times PM2.5_{s,t-1} + \beta_2 \times Pres_{s,t} + \beta_3 \times Temp_{s,t} + \beta_4 \times Iws_{s,t} + \beta_5 \times Iws \times PM2.5_{s,t-1} + \varepsilon_{s,t} \quad (2)\ .$$

其中，$PM2.5_{s,t}$为站点 s 在 t 时刻记录的 PM2.5 浓度，$PM2.5_{s,t-1}$为站点 s 在 t-1 时刻记录的 PM2.5 浓度，气压（*Pres*）、气温（*Temp*）和风速（*Iws*）为控制变量。此外，根据描述性分析中不同风速对 PM2.5 浓度一阶自相关系数的影响，向回归模型中引入交互变量 $Iws \times PM2.5_{s,t-1}$。$\beta_0, \beta_1 \ldots \beta_5$为回归模型的系数，与 s、t 无关，$\varepsilon_{s,t} \sim N(0, \sigma^2)$为独立同分布的随机误差项。

**（二）时空自回归模型（SAR）**

参考描述性分析中 PM2.5 浓度的空间相关性特征，在时间自回归模型（AR）中进一步加入变量周围“邻居”站点的平均浓度，定义时空自回归模型为

$$PM2.5_{s,t} = \beta_0 + \beta_1 \times PM2.5_{s,t-1} + \beta_2 \times PM2.5\_neighbor_{s,t} + \beta_3 \times Pres_{s,t} + \beta_4 \times Temp_{s,t} + \beta_5 \times Iws_{s,t} + \beta_6 \times Iws \times PM2.5_{s,t-1} + \varepsilon_{s,t} \quad (3).$$

其中，$PM2.5_{s,t}$、$PM2.5_{s,t-1}$、*Pres*、*Temp*、*Iws*的定义与时间自回归模型（AR）相同，$PM2.5_neighbor_{s,t}$为 t 时刻站点 s 周围“邻居”站点的平均浓度，具体定义如公式（1）所示。$\beta_0, \beta_1 \ldots \beta_6$为回归模型的系数，与 s、t 无关，$\varepsilon_{s,t} \sim N(0, \sigma^2)$为独立同分布的随机误差项。

（三）时空污染物自回归模型（MSAR）

参考描述性分析中 PM2.5 浓度的污染物相关性特征，在时空自回归模型（SAR）中进一步加入其他污染物变量，定义时空污染物自回归模型为

$$PM2.5_{s,t} = \beta_0 + \beta_1 \times PM2.5_{s,t-1} + \beta_2 \times PM2.5\_neighbor_{s,t} + \beta_3 \times PM10_{s,t} + \beta_4 \times CO_{s,t} + \beta_5 \times SO2_{s,t} + \beta_6 \times NO2_{s,t} + \beta_7 \times O3_{s,t} + \beta_8 \times Pres_{s,t} + \beta_9 \times Temp_{s,t} + \beta_{10} \times Iws_{s,t} + \beta_{11} \times Iws \times PM2.5_{s,t-1} + \varepsilon_{s,t} \quad (4).$$

其中，$PM2.5_{s,t}$、$PM2.5_{s,t-1}$、$PM2.5\_neighbor_{s,t}$、*Pres*、*Temp*、*Iws*的定义与时

空自回归模型（SAR）相同，CO、SO2、NO2、O3 为 t 时刻站点 s 记录的其他污染物浓度。$\beta_0, \beta_1 \cdots \beta_{11}$为回归模型的系数，与 s、t 无关，$\varepsilon_{s,t} \sim N(0, \sigma^2)$为独立同分布的随机误差项。

## 五、实证分析

### （一）三种自回归模型比较

为比较三种自回归模型的拟合能力，首先将样本数据划分为训练集和测试集，训练集为各个站点 2014-2016 年全部数据，共计 157824 条观测（包含部分存在变量有缺失值的观测），测试集为各个站点 2017 年 1 月至 2 月数据，共计 8496 条观测（包含部分存在变量有缺失值的观测）。根据上述建立的三种 PM2.5 浓度自回归模型，将训练集带入估计方程，三组回归的系数估计、训练集决定系数（R^2）、测试集均方误差（RMSE）如表 3 所示，

表 3：三种自回归模型回归系数表

| | AR | SAR | MSAR |
|---|---|---|---|
| Intercept | 352.9 *** | 129.6 *** | 18.95 |
| PM2.5_lag | 0.967 *** | 0.645 *** | 0.523 *** |
| PM2.5_neighbor | | 0.360 *** | 0.239 *** |
| PM10 | | | 0.155 *** |
| CO | | | 2.719 *** |
| SO2 | | | -0.023 *** |
| NO2 | | | 0.045 *** |
| O3 | | | 0.028 *** |
| Pres | -0.340 *** | -0.127 *** | -0.026 * |
| Temp | -0.434 *** | -0.120 *** | -0.045 *** |
| Iws | 0.071 *** | 0.064 *** | 0.064 *** |
| Iws*PM2.5_lag | -0.003 *** | -0.002 *** | -0.002 *** |
| $R^2$ | 93.39% | 94.58% | 95.55% |
| *RMSE* | 36.62 | 31.41 | 28.04 |

注：***表示 0.001 显著，**代表 0.005 显著，*代表 0.01 显著，.代表 0.05 显著

对比三个模型的系数估计以及模型拟合能力可以看出，除截距项、PM2.5 一

阶滞后项外，其余变量在不同模型中系数估计的结果较为稳定。PM2.5 浓度的一阶滞后项对当期 PM2.5 浓度的影响随着变量周围“邻居”PM2.5 浓度、其他污染物浓度的引入逐渐减弱。

AR 模型、SAR 模型、MSAR 模型训练集上的$R^2$依次增加，并且三者的$R^2$均超过了 90%，说明三个模型在样本内都有着很强的拟合能力。其中，MSAR 模型的$R^2$高达 95.55%，说明自变量可解释因变量波动的 95.55%,

AR 模型、SAR 模型、MSAR 模型测试集上的 ***RMSE*** 依次减小，并且减小幅度较为明显，MSAR 模型的 ***RMSE*** 仅有 28.04，明显低于其他两个模型，说明虽然三个模型的$R^2$差别较小，但 MSAR 模型在样本外有着更强的预测能力。

（二）MSAR 模型诊断

在建立回归模型后，需对回归模型进行诊断，以 MSAR 模型为例，检验是否存在违背线性回归基本假设的问题，使得模型估计结果有误。

1. 多重共线性分析

回归模型的多重共线性问题会导致系数估计的结果不稳定，使得系数估计的结果与预期相反，或是导致对因变量有重要的变量反而不显著。多重共线性问题常用方差膨胀因子 (VIF) 判断，第 i 个自变量的方差膨胀因子定义为$VIF_i = \frac{1}{1-R_i^2}$，其中$R_i^2$为以第 i 个自变量为被解释变量、以其余自变量为解释变量建立回归方程得到的决定系数。

表 4 计算了 MSAR 模型中，各个变量的方差膨胀因子，可以看出所有变量的方差膨胀因子均小于 10，可以认为模型不存在明显的多重共线性问题。

表 4：MSAR 模型各变量方差膨胀因子

| PM2.5_lag | PM2.5_neighbor | PM10 | CO | SO2 | NO2 | O3 | PRES | TEMP | Iws | Iws*PM2.5_lag |
|---|---|---|---|---|---|---|---|---|---|---|
| 9.44 | 9.17 | 6.71 | 3.01 | 2.15 | 2.44 | 1.8 | 3.89 | 5.48 | 2.2 | 2.19 |

2. 残差分析

回归模型的异方差问题会导致虽然系数估计仍是线性无偏的，但不再是最优线性无偏估计量（BLUE）。异方差通常可以通过残差图判断。

图 9 为 MSAR 模型残差关于拟合值的散点图，可以看出散点随机分布在水平带状区域，可以认为模型不存在异方差的问题。

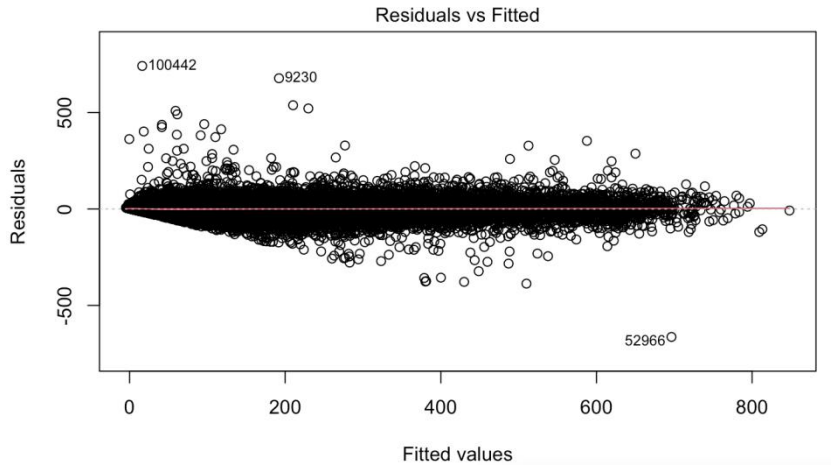


图 9：MSAR 模型残差关于拟合值的散点图

综上所述，MSAR 模型不存在违背模型基本假设的问题，并且 MSAR 模型的拟合及预测能力优于 AR 和 SAR 模型，最终选取 MSAR 模型作为 PM2.5 浓度的拟合模型。

**（三）MSAR 模型解读**

从表 3 中 MSAR 模型的系数估计中可以解读出，在控制其他变量不变时，上一期 PM2.5 浓度每增加 1μg/m$^3$，当期 PM2.5 浓度平均增加 0.523μg/m$^3$；周围站点 PM2.5 的加权平均浓度每增加 1μg/m$^3$，中心站点 PM2.5 浓度平均增加 0.239μg/m$^3$；PM10 浓度每增加 1μg/m$^3$，PM2.5 浓度平均增加 0.155μg/m$^3$；CO 浓度每增加 1μg/m$^3$, PM2.5 浓度平均增加 2.719μg/m$^3$; SO2 浓度每增加 1μg/m$^3$，PM2.5 浓度平均增加-0.023μg/m$^3$；NO2 浓度每增加 1μg/m$^3$，PM2.5 浓度平均增加 0.045μg/m$^3$；O3 浓度每增加 1μg/m$^3$，PM2.5 浓度平均增加 0.028μg/m$^3$；气压每增加 1hPa，PM2.5 浓度平均降低 0.026μg/m$^3$；气温每增加 1℃，PM2.5 浓度平均降低 0.045μg/m$^3$；风速每增加 1m/s，PM2.5 浓度平均增加 0.064μg/m$^3$；风速每增加 1m/s，上一期 PM2.5 浓度对当期 PM2.5 浓度的边际影响（系数估计）降低 0.002；

可以看出，回归模型系数估计定量反映出了 PM2.5 浓度的时间相关性、空间相关性以及污染物相关性。部分气象变量的系数估计与描述性分析的结论相反，可能是由于存在中间效应，受到了其他自变量的影响。

（四）考虑季节效应的 MSAR 模型

PM2.5 浓度的分布呈现出明显的季节特征，进一步地分季节拟合 MSAR 模型，比较不同季节下各个变量的系数估计和模型拟合能力的差异，结果如表 5

所示。可以看出，冬季模型（12～2 月）的决定系数最高，达到 96.13%，秋季模型（6～8 月）的决定系数最低，只有 88.37%；四个季节模型 PM2.5 滞后一期浓度（PM2.5_lag）、“邻居”PM2.5 浓度（PM2.5_neighbor）的系数估计总体差异较小，夏秋季节 PM2.5 滞后一期浓度的系数估计略高于春冬，说明这两个季节下 PM2.5 浓度的时间相关性更高，夏冬季节“邻居”PM2.5 浓度的系数估计略高于春秋，说明这两个季节下 PM2.5 浓度的空间相关性更高。

表 5：各季节 MSAR 模型回归系数表

| | 春季（3～5 月） | 夏季（6～8 月） | 秋季（9～11 月） | 冬季（12～2 月） |
|---|---|---|---|---|
| (Intercept) | -132.9*** | 178.57*** | 121.3*** | -163.53*** |
| PM2.5_lag | 0.47*** | 0.55*** | 0.54*** | 0.51*** |
| PM2.5_neighbor | 0.17*** | 0.32*** | 0.21*** | 0.27*** |
| PM10 | 0.23*** | 0.08*** | 0.19*** | 0.16*** |
| CO | 2.70*** | 1.3*** | 0.54** | 1.00*** |
| SO2 | -0.05*** | -0.005 | -0.03*** | 0.001 |
| NO2 | 0.16*** | 0.04*** | -0.02*** | 0.01 . |
| O3 | 0.02** | 0.02*** | 0.04*** | 0.03*** |
| PRES | 0.12*** | -0.18*** | -0.11*** | 0.16*** |
| TEMP | 0.02 | -0.24*** | -0.54*** | -0.03 |
| Iws | 0.01 | 0.11*** | 0.07*** | 0.08*** |
| I(Iws * PM2.5_lag) | -0.003*** | -0.002*** | -0.001*** | -0.001*** |
| $R^2$ | 95.93% | 92.31% | 88.37% | 96.13% |

注：***表示 0.001 显著，**代表 0.005 显著，*代表 0.01 显著，.代表 0.05 显著

将 2017 年 1 至 2 月的测试集数据带入冬季 MSAR 模型，并与 MSAR 模型的预测结果进行对比，结果表明冬季 MSAR 模型在测试集的预测均方误差更低（27.51），小于不考虑季节因素的 MSAR 模型的预测结果（28.04）。从图 10 中可以看出，冬季 MSAR 模型的预测值和实际值的时间序列折线图基本重合，不考虑季节效应的 MSAR 模型存在较多的偏离值（图中圈中部分），综上所述，分季节建立 MSAR 模型有助于提高 PM2.5 浓度变化规律的拟合能力。

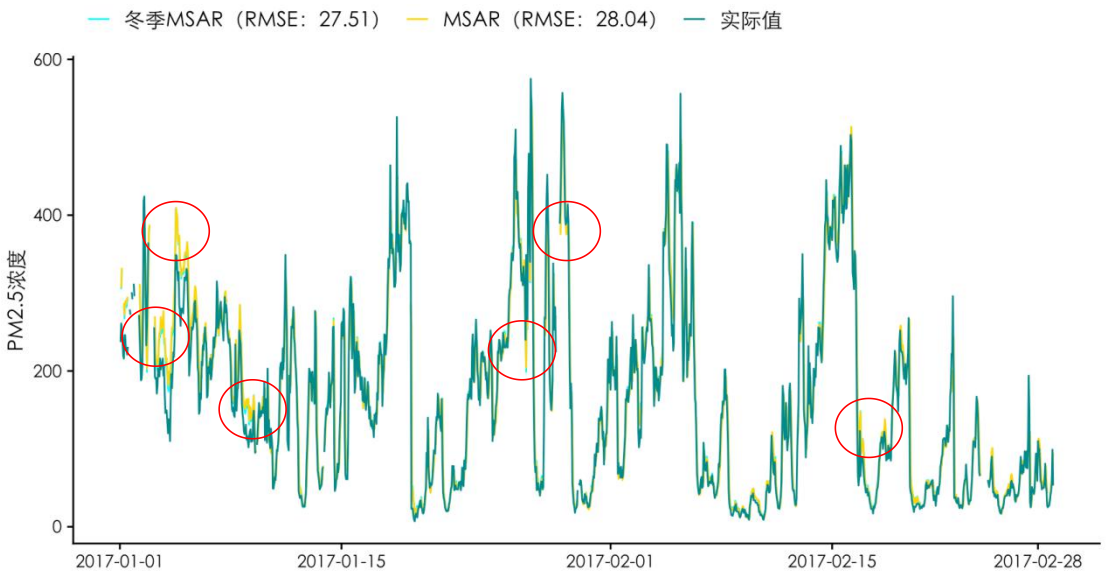


图 10：MSAR 模型、冬季 MSAR 模型预测折线图

## 六、模型应用

（一）异常值检测

对实时大气污染检测数据的进行异常标注是空气质量监测的重要工作之一，快速、高效的异常值标签能够有效建立检测预警机制，帮助监测人员及时发现异常。然而，气象观测站点现行的打标签模式多为人工模式，即每个站点各自挑选多位工作人员，根据自身经验，对污染物数据进行异常标注。上述打标签模式可能存在以下问题：标准不统一，一方面，不同工作人员的评判标准无法完全一致；另一方面，不同站点自身情况不同、所处地理环境不同，对标签异常的判定标准也不统一；漏打标签，由于精力和注意力有限，当工作人员面对大量的污染物浓度观测时很有可能出现漏打标签的情况，并且人工很难同时考虑到时间异常（相较滞后期变动较大）、空间异常（相较周围站点变动较大）、污染物异常（相较其他污染物变动较大）；人力成本高，每个站点都需要有多位工作人员全天候对数据进行核查检测，人力成本较高。基于人工标签存在的上述问题，运用本文建立的 PM2.5 时空污染物自回归模型（MSAR），可以实现异常值实时检测、自动化打标签以代替人工标签。

（二）缺失值插补

目前的大气污染检测数据受技术因素或人为因素的影响，存在一定的缺失情

况。建立及时、合理的缺失值插补机制有助于大气污染情况的实时播报预警，也有助于其他数据挖掘工作。运用本文建立的 PM2.5 时空污染物自回归模型（MSAR），可以实现缺失值的插补工作。

## 七、研究总结与展望

本文提出了一种时空污染物自回归模型（MSAR）用于刻画 PM2.5 浓度变化规律，并将模型应用到石家庄市 2014-2017 年六站点大气检测数据，结果表明本文提出的 MSAR 模型相比传统时间自回归、时空自回归有着更好的拟合能力。此外，通过比较不同季节的 MSAR 模型拟合差异，发现冬季模型 PM2.5 浓度的拟合效果更好。研究结果可用于大气污染物成因分析、大气污染物浓度异常值识别、大气污染物浓度缺失值插补以及大气污染物浓度预测。

未来研究可以扩展数据的广度，将模型应用到全国站点进行分析。此外，可以扩展模型的深度，尝试更多的距离加权方式、将多种污染物一起建立向量自回归模型（VAR）。

## 参考文献

[1]郝春旭、邵超峰、董战峰、赵元浩. 2020 年全球环境绩效指数报告分析[J]. 环境保护, 2020, v.48;No.687(16):68-72.

[2]中华人民共和国生态环境部. 2019 中国生态环境状况公报[M]. 2020.

[3]Rohde R A , Muller R A . Air Pollution in China: Mapping of Concentrations and Sources[J]. Plos One, 2015, 10(8):e0135749.

[4]Roy R , Braathen N A . The Rising Cost of Ambient Air Pollution thus far in the 21st Century: Results from the BRIICS and the OECD Countries[J]. OECD Environment Working Papers, 2017.

[5] Hoek G , Krishnan R M , Beelen R , et al. Long-term air pollution exposure and cardio- Respiratory mortality: A review[J]. Environmental Health, 2013, 12(1):43.

[6]Gu X, Wang L, Zhuang W, Han L. Reduction of wheat photosynthesis by fine particulate (PM2.5) pollution over the North China Plain. Int J Environ Health Res. 2018 Dec;28(6):635-641.

[7]Cao C , Lee X , Liu S , et al. Urban heat islands in China enhanced by haze pollution[J]. Nature Communications, 2016, 7:12509.

[8]国务院. 国务院关于印发大气污染防治行动计划的通知[EB/OL]. http://www.gov.cn/zhengce/content/2013-09/13/content_4561.htm, 2013-09-13.

[10]新华社. 环保部回应西安环境数据造假案：将建立监测数据造假防范和惩治机制[EB/OL]. http://www.xinhuanet.com/politics/2017-06/17/c_1121162461.htm, 2017-06-17.

[11]人民网. 生态环境部通报山西临汾空气监测数据造假案[EB/OL]. http://env.people.com.cn/n1/2018/0830/c1010-30262201.html, 2018-08-30.

[12] Bobbia M , Misiti M , Misiti Y , et al. Spatial outlier detection in the PM10 monitoring network of Normandy (France)[J]. Atmospheric Pollution Research, 2015, 6(3).

[13] Chu H J , Yu H L , Kuo Y M . Identifying spatial mixture distributions of PM2.5 and PM10 in Taiwan during and after a dust storm[J]. Atmospheric Environment, 2012, 54(Jul.):728-737.

[14]张亮林, 张大弘, 潘竟虎,等. 基于 GWR 降尺度的京津冀地区 PMsub2.5/sub 质量浓度空间分布估算[J]. 环境科学学报, 2019.

[15]Kübra Ağaç, Kasım Koçak, Ali Deniz. Simulation And Forecasting of Daily Pm10 Concentrations Using Autoregressive Models In Kagithane Creek Valley, Istanbul[J]. Geophysical Research Abstracts, 2015,Vol. 17, EGU2015-5737-7

[16]Batterman, Stuart, Lizhong, et al. Characteristics of PM2.5 concentrations across Beijing during 2013-2015[J]. Atmospheric environment. 2016, 145(Nov.):104-114.

[17]Wang C, Jiang H , D Pan, et al. A key study on spatial source distribution of PM2.5 based on the airflow trajectory model[J]. International Journal of Remote Sensing, 2016, 37(23-24):5864-5883.

[18]Wang W , Niu Z . VAR Model of PM2.5, Weather and Traffic in Los Angeles-Long Beach Area[C]// 2009 International Conference on Environmental Science and Information Application Technology. IEEE Computer Society, 2009.

[19]陈兵红,靳全锋,柴红玲,郭福涛.浙江省大气 PM_(2.5)时空分布及相关因子分析[J].环境科学学报,2021,41(03):817-829.

# 中央财经大学本科毕业论文（设计）原创性声明

本人郑重声明：所提交的毕业论文（设计）《我国 PM2.5 浓度时空变化特征及其影响因素研究—基于时空自回归模型》，是本人在指导老师的指导下独立进行研究工作所取得的成果。除文中已经注明引用的内容外，不含任何其他个人或集体已经发表或撰写过的作品成果，不存在购买、由他人代写、剽窃和伪造数据等作假行为。对本文研究/设计做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果，如违反有关规定或上述声明，愿意承担由此产生的一切后果。

作者签名：  2021 年 4 月 15 日

# 致谢

感谢父母的养育之恩，感谢导师潘蕊在我大学时给予我的帮助，感谢身边朋友一直对我的包容和体谅，感谢一直坚持努力的自己。