文章编号：1674-8085（2022）02-0015-07

# 基于时间序列与随机森林模型的江西省空气质量影响因素分析

*王新长，殷振宇

（井冈山大学数理学院，江西，吉安 343009）

**摘 要：**空气质量状况直接影响着人们的身心健康，空气污染治理一直是一个广受争论的热点问题。本文基于2015~2020 年江西省各地级市主要污染物浓度和气象数据，采用时间序列与随机森林模型，深入分析江西省各地级市的空气质量状况及其影响因素，得到以下结果：(1)从整体角度来看，2015~2020 年间江西省城市的空气质量一直处于优良状态，且呈现出不断提高趋势。(2)从季节角度来看，各城市每年的 AQI 值呈现倒“山”形分布，各污染物浓度变化特征明显，除 $O_3$ 外其他污染物浓度大致呈现“冬秋高、夏秋低”的特点，其中萍乡市的变化最为突出。(3)从气象因子的角度看，平均气温、平均气压和平均水气压对空气中主要污染物影响最大，其他气象因子对不同污染物浓度影响程度有明显差异。

**关键词：**空气质量指数；数据挖掘；随机森林；气象因素

**中图分类号：**X823 **文献标识码：**A DOI：10.3969/j.issn.1674-8085.2022.02.003

# ANALYSIS OF INFLUENCING FACTORS OF AIR QUALITY IN JIANGXI PROVINCE BASED ON TIME SERIES AND RANDOM FOREST MODEL

*WANG Xin-chang, YIN Zheng-yu

（School of Mathematics and Physics, Jinggangshan University, Ji’an, Jiangxi 343009, China）

**Abstract:** Air quality has a direct impact on people’s physical and mental health, and air pollution treatment always is a widely debated hot issue. Based on the concentration of major pollutants and meteorological data of prefecture-level cities in Jiangxi Province from 2015 to 2020, this paper uses time series and random forest model to deeply analyze the air quality of prefecture-level cities in Jiangxi Province and its influencing factors, and the following results are obtained: (1) From the overall perspective, the urban air quality of Jiangxi Province is in a good state from 2015 to 2020, and shows a trend of continuous improvement. (2) From the perspective of season, the annual AQI value of each city presents an inverted “mountain” distribution, and the variation of pollutant concentration is obvious. Except O3, the concentrations of other pollutants are generally “high in winter and autumn, low in summer and autumn”, among which the change of Pingxiang City is the most prominent. (3) From the perspective of meteorological factors, average temperature, average pressure and average water pressure have the greatest impact on the main pollutants in the air, while other meteorological factors have significant differences in the impact of pollutants with different concentrations.

**Key words：** air quality index; data mining; random forest; meteorological factors

收稿日期：2021-08-01；修改日期：2021-10-22
基金项目：江西省教育厅科技计划项目(GJJ180565)；井冈山大学校级课题(JZB1824)
作者简介：*王新长(1971-)，男，江西于都人，副教授，博士，主要从事管理决策和优化研究(E-mail:wangxinchang11@163.com).

# 0 引言

空气质量直接影响着人们的身心健康，自 1978 年的改革开放以来，我国的经济高速发展。与此同时，能源不断地被消耗，空气污染问题已经迫在眉睫。从倡导低碳生活到垃圾分类，环境问题已经得到了全社会的广泛关注。

到目前为止，很多学者对空气质量变化及其影响因素进行过大量的研究。在空气中主要污染物分布情况上这一方面，刘艳艳研究表明供暖季的空气污染程度明显高于非供暖季[1]；$SO_2$、$NO_2$、$PM_{10}$ 月均浓度呈 V 型变化，各空气污染物浓度夏季最低，冬季最高[2]；$PM_{2.5}$、$PM_{10}$、$NO_2$、$SO_2$ 和 CO 在冬季空气中的浓度最高，$O_3$ 在夏季空气中的浓度最高；AQI 与平均降水量和气温与呈负相关，并且受到工业企业数的影响较大[3-4]；中国空气中的 $PM_{2.5}$ 平均浓度具有“冬秋高、春夏低”的“U”型逐月变化规律[5-7]。在气象因子对主要污染物影响程度上这一方面，赵东宇研究发现日均气温、气压、日照时数和降水量均与 API 呈现负相关关系，日最高温、日均风速和日均相对湿度均与 API 呈现正相关关系，气温、气压、降水量和相对湿度与部分城市的空气质量关系密切[8-10]。污染源的排放时间与大气扩散条件的时间是空气质量变化的主要原因[10-12]；第二产业占 GDP 的百分比、年平均饱和水气压、城区海拔落差、城市建成区面积是影响中国城市空气质量的主要因素[13-15]；

本文通过网络爬虫获取 2015~2020 年间江西省每月大气污染物浓度数据，并通过气象数据网获取同一时间的气象数据，分析空气质量指数与空气中主要污染物浓度变化的时序图、季节效应图和趋势图，在此基础上通过随机森林模型度量各气象因子对于大气中主要污染物的影响程度，以期为进一步治理江西省大气污染提供科学依据。

# 1 资料与方法

## 1.1 数据来源

江西省 11 个地级市的 AQI 值和 6 项主要污染物浓度数据来源于中国空气质量在线监测分析平台发布的每日数据（https://www.aqistudy.cn/historydata），并通过网络爬虫技术获得南昌、吉安、赣县、南城和景德镇的气象因子数据，主要来源于中国气象数据网(http://data.cma.cn 2015-2020 年)。以上数据均通过质量检验。

## 1.2 研究技术与方法

根据中国空气质量在线监测分析平台所获取的每日历史数据，通过 R 软件计算出各地级市的月均值，在季节划分上本文按照气象的四个季节时间(春季：3-5 月，夏季：6-8 月，秋季：9-11 月，冬季：12-2 月)来分季节。空气质量指数(AQI)及 6 种大气主要污染物分级根据《环境空气质量标准》（GB 3095-2012）（中国环境科学研究院，2012)将空气质量划分为 6 个等级（见表 1）。各主要空气污染物质量浓度划分标准按照《环境空气质量评价技术规范（试行）》（HJ 663—2013)(中华人民共和国生态环境部，2013）执行。

AQI 是对空气污染指数 API 的一种修正，它综合考虑了 $PM_{2.5}$、$PM_{10}$、$NO_2$、$SO_2$、CO 和 $O_3$ 对人体健康的影响，其计算公式如下：

$$AQI = \max\{IAQI_1，IAQI_2，\cdots，IAQI_6\} \quad (1)$$

(1)式中 $IAQI_1$ 表示空气质量分指数，其计算公式如下：

$$IAQI_i = \frac{IAOI_{high} - IAOI_{low}}{B_{high} - B_{low}}(C_i - B_{low}) + IAQI_{low} \quad (2)$$

式中 $IAQI_i$ 是指污染物 $i$ 的空气质量分指数；$C_i$表示污染物$i$的质量浓度；$B_{high}$ 与 $B_{low}$ 分别对应与污染物 $C_i$ 相近的污染物浓度的最高和最低限值；$IAQI_{high}$ 与 $IAQI_{low}$ 分别为污染物$B$的质量分数。

**表 1 AQI 取值范围及相应的空气质量类别**

Table 1 Value range of AQI and corresponding air quality category

| AQI 取值范围 | 空气质量状况 |
|---|---|
| 0-50 | 优 |
| 51-100 | 良 |
| 101-150 | 轻度污染 |
| 151-200 | 中度污染 |
| 201-250 | 重度污染 |
| 251-300 | |
| 301-500 | 严重污染 |

**表 2 AQI 及各项污染物浓度限值**

Table 2 AQI and pollutant concentration limits

| 空气质量分指数 | 污染物浓度限值 | | | | | |
|---|---|---|---|---|---|---|
| （IAQI） | $SO_2$ (24h) | $NO_2$ (24h) | $PM_{10}$ (24h) | CO (24h) | $O_3$ (8h) | $PM_{2.5}$ (24h) |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50 | 50 | 40 | 50 | 2 | 100 | 35 |
| 100 | 150 | 80 | 150 | 4 | 160 | 75 |
| 150 | 475 | 180 | 250 | 14 | 215 | 115 |
| 200 | 800 | 280 | 350 | 24 | 265 | 150 |
| 300 | 1600 | 565 | 420 | 36 | 800 | 250 |
| 400 | 2100 | 750 | 500 | 48 | - | 350 |
| 500 | 2620 | 940 | 600 | 60 | - | 500 |

1.2.1 网络爬虫技术

网络爬虫技术是通过计算机模拟人的手动操作，通过批处理的手段从网站上获取数据，从而减少人力。本文所涉及的步骤如下：

a.设定 headers 以及需要访问的 url；

b.使用 requests 包中的 get 和 content 函数获取网页信息；

c.选择需要爬取城市和日期范围；

d.通过 selenium 包中的 PhantomJS 函数模拟浏览器获取选定的城市和日期范围内的数据；

e.通过 pandas 包中的 read_html 函数获取网页中的表格；

f.将所爬取的数据通过 to_csv 函数保存至本地（csv）；

g.将收集到的数据使用 R 软件进行清洗。

1.2.2 时间序列确定性分解

通过 Cramer 分解定理，对于任何一个时间序列$\{Y_t\}$都可以分解成两部分的叠加，其中一部分为多项式决定的确定性趋势序列，另一部分是平稳的零均值误差序列，即：

$$x_t = \mu_t + \varepsilon_t = \sum_{j=0}^{d} \beta_j t^j + \psi(B)a_t \tag{3}$$

式中 $d<\infty$，$\beta_1,\cdots,\beta_d$ 为常数系数，$\{a_t\}$ 为一个零均值白噪声序列，$B$为延迟算子。本文将时间序列简单假定为相加模型的时间序列：

$$x_t = Seasonal_t + Trend_t + Irregular_t$$

其中 $x_t$ 为在时刻 $t$ 下的观测值，$Seasonal_t$、$Trend_t$ 和 $Irregular_t$ 分别为在时刻 $t$ 下的季节效应、趋势值与随机影响之和，下文主要对 $x_t$、$Seasonal_t$、$Trend_t$ 和 $Irregular_t$ 进行分析。

1.2.3 随机森林变量重要性度量

随机森林通过同时生成多个模型，通过计算各个变量在不同参数模型下的贡献度，取平均得变量的重要性，本文使用的计算方法是平均不纯度的减少。

假设有特征分别为 $x_1,x_2,\cdots,x_m$，以计算 Gini 指数为例，用 $GI_m$ 来表示，则：

$$GI_m = \sum_{k=1}^{|K|}\sum_{k'\neq k} P_{mk}P_{mk'} = 1-\sum_{k=1}^{|K|} P_{mk}^2 \tag{4}$$

其中，$K$ 表示类别，$P_{mk}$ 表示树的节点 $m$ 中类别 $k$ 所占的比例，则特征 $X_j$ 在节点$m$的重要性为：

$$VIM_{jm}^{Gini} = GI_m - GI_l - GI_r \tag{5}$$

其中，$VIM$ 表示变量重要性评分，$GI_l$ 和 $GI_r$ 表示分枝后两个新节点的 Gini 指数。假设特征 $X_j$ 在决策树 $i$ 中出现的节点在集合 $M$ 中，那么 $X_j$ 在第 $i$ 颗树的重要性为：

$$VIM_{ij}^{Gini} = \sum_{m\in M} VIM_{jm}^{Gini} \tag{6}$$

并且，假设随机森林中共有 $n$ 颗树，则：

$$VIM_{j}^{Gini} = \sum_{i=1}^{n} VIM_{ij}^{Gini} \tag{7}$$

最后，将上述所求得的变量重要性评分进行归一化处理：

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i} \tag{8}$$

# 2 江西省空气污染物时间分布情况

## 2.1 时间变化特征

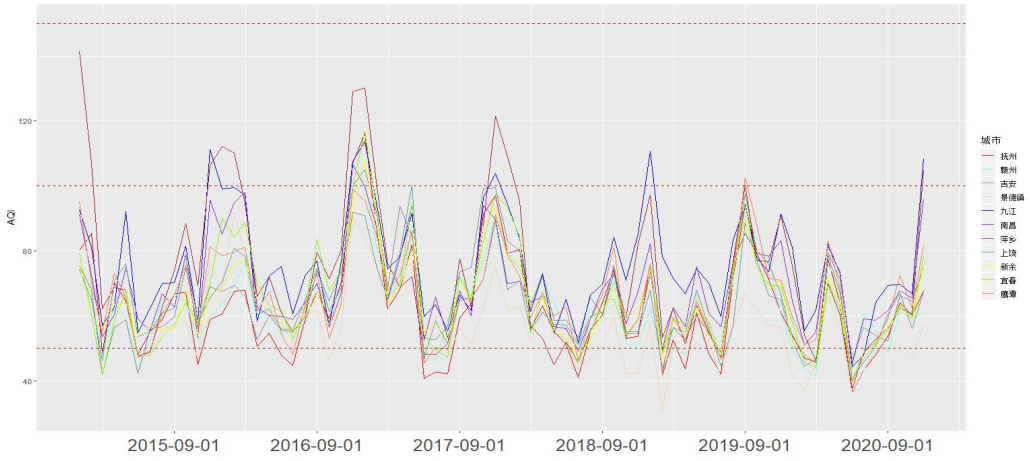
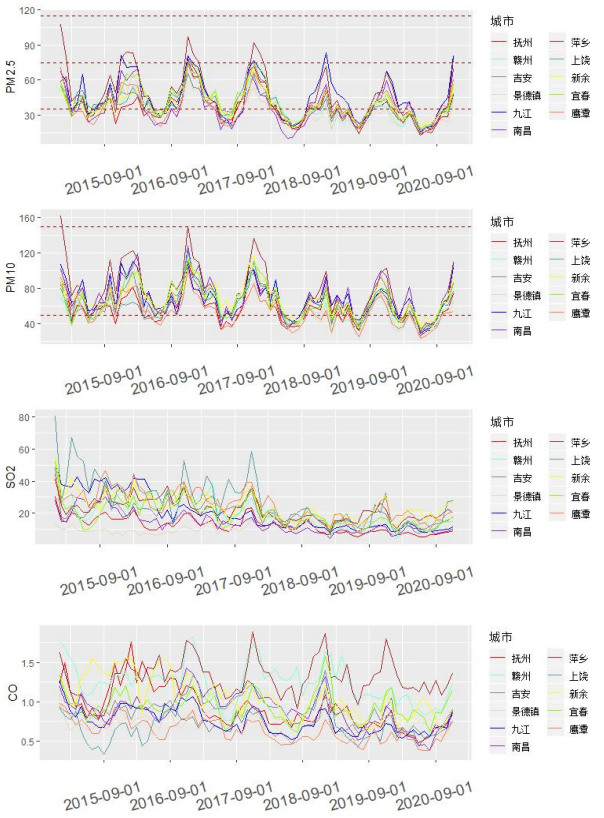



图 1 AQI 指数值的逐年变化图

Fig. 1 Chart of AQI index values changing year by year

图 1 可知，从整体角度来看，AQI 均值在 60-70 左右，说明江西省空气质量整体处于优良。仅萍乡市在 2016~2017 年间的 AQI 略有提升，不过空气质量依然属于良，并且 2018 年之后空气质量不断提升。



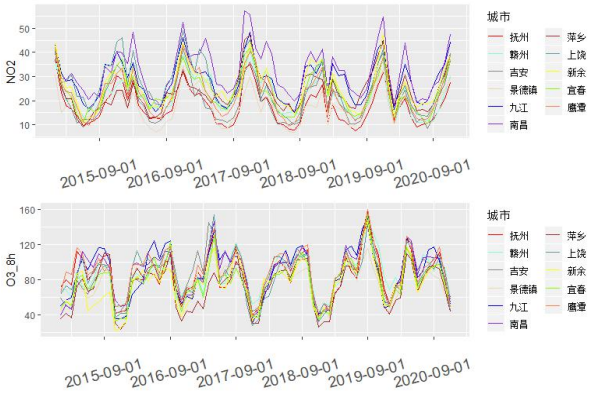




图 2 各污染物浓度逐年变化图

Fig. 2 Chart of the variation of pollutant concentration year by year

由图 2 可知，只有鹰潭市和萍乡市的 $PM_{2.5}$ 值超过限定的二级标准，说明两市有轻度 $PM_{2.5}$ 污染；萍乡市 $PM_{10}$ 污染程度相对较高，但仍属于良；上饶市的 $SO_2$ 污染在 2015~2017 年间相对较高，但仍属于良，并且在之后处于优的水平；各城市 CO 和 $NO_2$ 治理一直较好，均属于优；各城市的 $O_3$ 污染程度差距不大，均属于优良。

## 2.2 季节变化特征

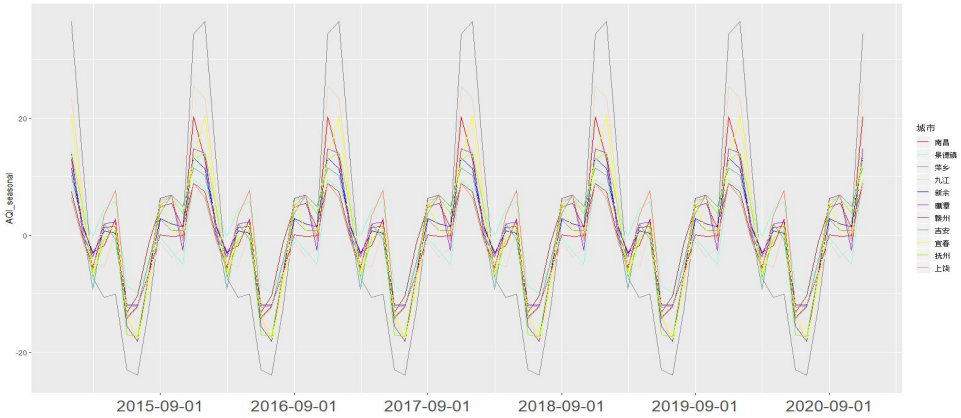
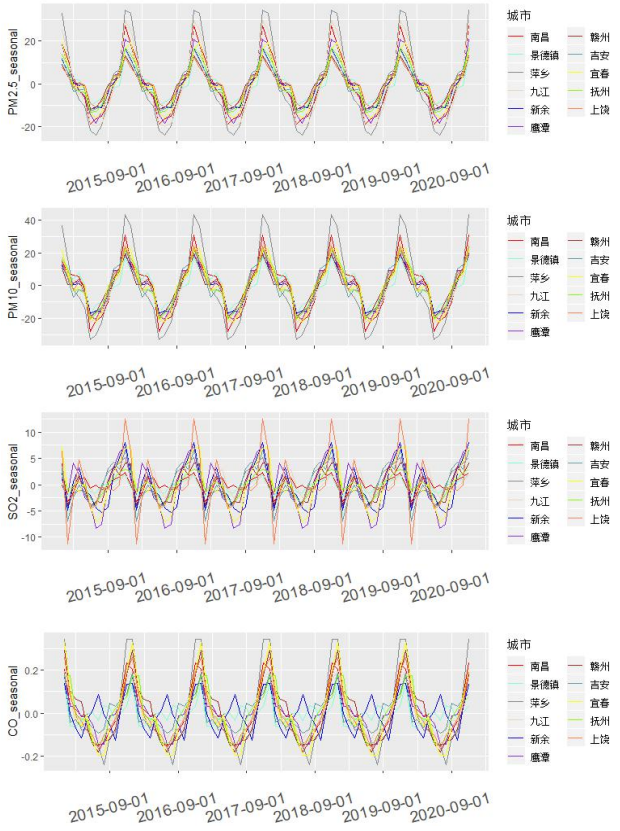



图 3 各市 AQI 值的季节效应图

Fig. 3 Seasonal variation diagram of AQI concentrations

由图 3 可知，在去除趋势因子和随机因子的前提下，萍乡市和九江市的 AQI 值变化幅度最为明显；其中上饶市的 AQI 值变化幅度最小。总体上来看，各城市的 AQI 值具有明显的季节性特征，即“冬季高，夏季低”。



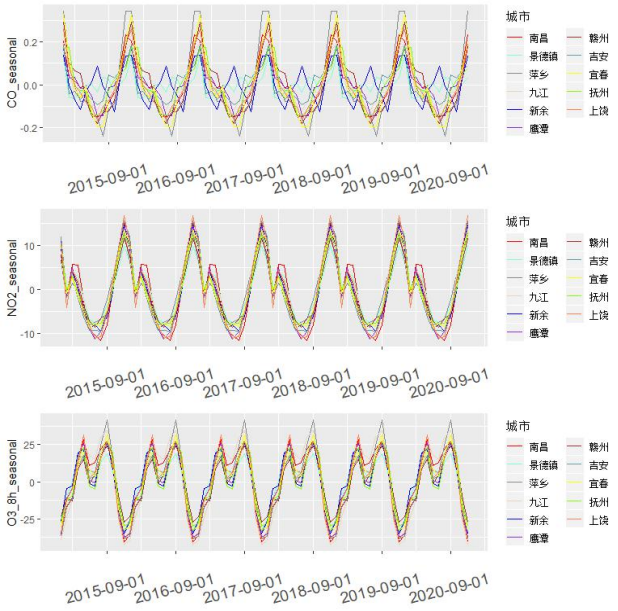




图 4 各污染物浓度的季节效应图

Fig. 4 Seasonal variation diagram of pollutant concentration

由图 4 可知，大气中的主要污染物浓度变化均呈现明显的季节性特征并在 3 月、4 月均有一个折点。萍乡市、九江市和南昌市 $PM_{2.5}$ 和 $PM_{10}$ 季节性特征最为明显；上饶市 $SO_2$ 的峰值与谷值均较为突出，鹰潭市和宜春市 $SO_2$ 浓度在 4 月、5 月的谷值相比其他城市较低；景德镇市 CO 浓度变化无明显的季节性特征，新余市 CO 浓度变化呈现“春秋高，冬夏低”的特点，其余城市 CO 浓度均呈现“冬季高，夏季低”的特点。$O_3$ 浓度变化与上述污染物的变化情况相反，呈现“夏季高，冬季低”的特点。

## 2.3 趋势变化特征

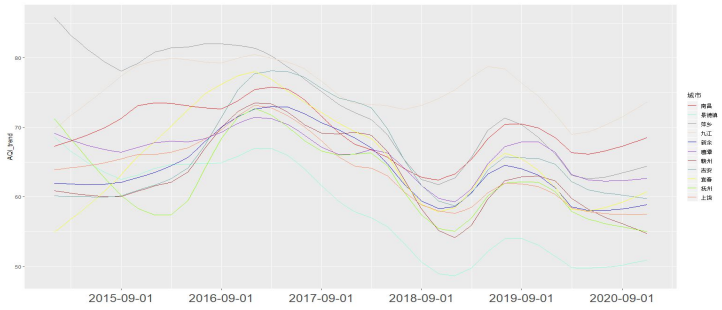
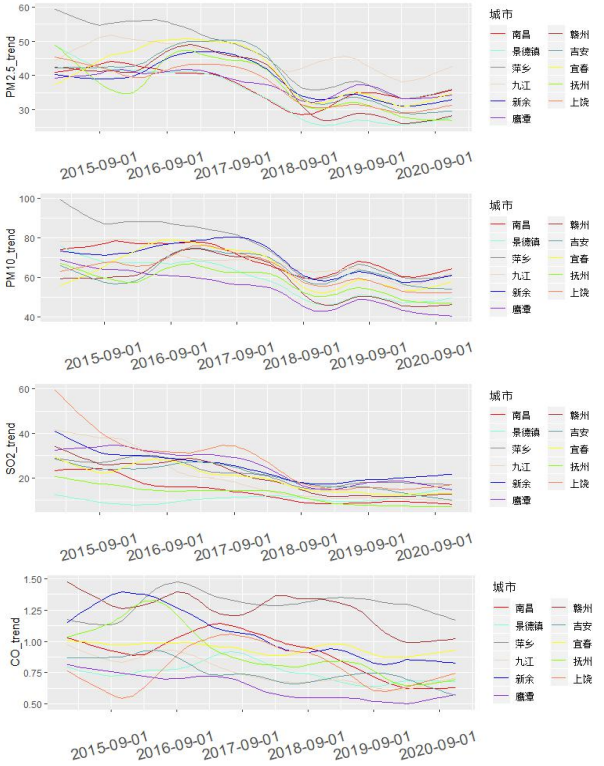



图 5 AQI 值趋势变化图

Fig. 5 The trend chart of AQI values

从图 5 可知，从总体来看，江西省空气质量整体处于优良，并且呈现不断提高的趋势。在 2015 年间萍乡市和抚州市的空气质量有明显改善，但是在 2016 年间均有轻微污染，其他城市正好与之相反，在 2015-2016 年间空气质量有轻度污染；各城市在 2017-2018 年间的空气质量均有所改善，说明这期间江西省的空气治理有显著成效，但是均在 2019 年中旬有轻度污染。

从图 6 可知，除 $O_3$ 外各污染物浓度总体呈现正在减少的趋势，并且所有污染物浓度均处于优的水平。说明近些年江西省的空气治理问题上，有显著成效。



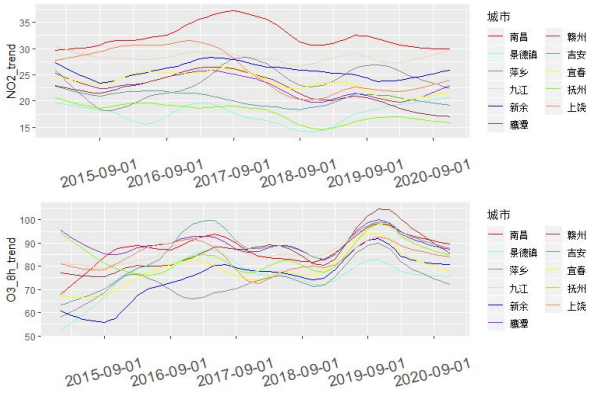
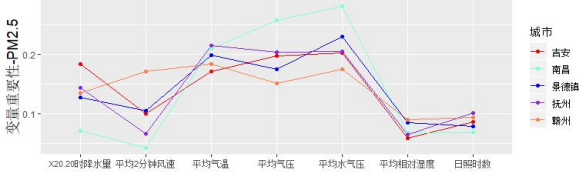




图 6 各污染物浓度趋势变化图

Fig. 6 The trend chart of variation values

# 3 气象因子对江西省空气质量的影响

多项研究表明，气象因素与空气污染密不可分。可以认为在固定的时间内，一个地区的污染物释放量是固定的，所以此地的空气质量主要由气象因素所决定，其影响着污染物的稀释、扩散、迁移以及沉降。考虑到江西省只有五个气象观测站，所以本文仅以此为例，考虑到的地区有：吉安市、南昌市、景德镇市、抚州市和赣州市。本文采用的气象因子有：20-20 时降水量、平均气压、平均 2 分钟风速、平均气温、平均水气压、平均相对湿度和日照时数。



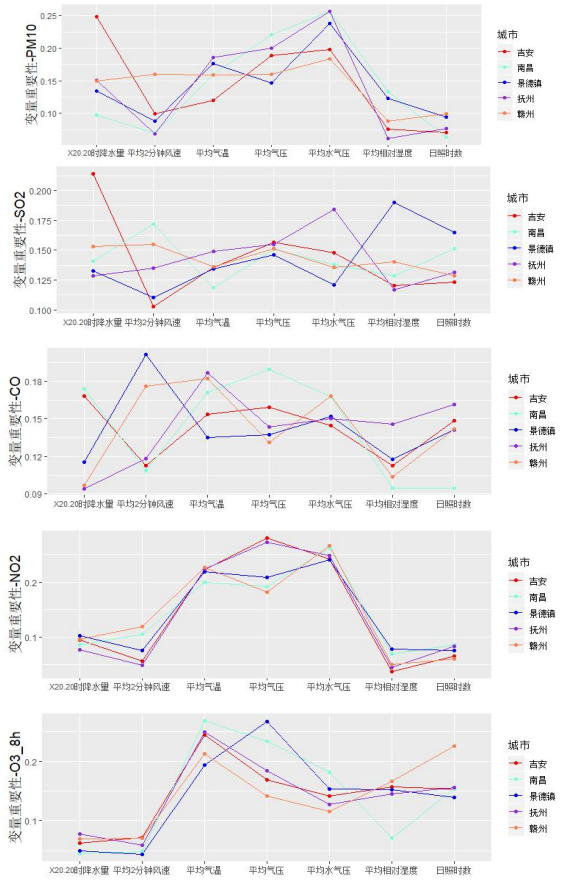



图 7　各市空气污染物的随机森林变量重要性图

Fig. 7　Map of the importance of random forest variables for air pollutants in each city

由图 7 可知，平均气温、平均气压和平均水气压对大气主要污染物浓度均有较大影响，20-20 时降水量对吉安市的 $PM_{2.5}$、$PM_{10}$、$SO_2$ 和 CO 浓度影响较大，平均 2 分钟风速对南昌市的 $SO_2$ 浓度与景德镇市和赣州市的 CO 浓度影响较为突出，平均相对湿度对景德镇市的 $SO_2$ 浓度影响较大，日照时数对赣州市的 $O_3$ 浓度影响较为突出。

## 4　结论

本文基于 2015-2020 年江西省 11 个地级市的空气质量数据、5 个气象观测站的气象数据和江西省统计年鉴，利用 R 软件进行时间序列分析和随机森林模型探索江西省空气质量时间分布特征及其与气象因子的联系，得出以下结论：

1）江西省整体空气质量较优良，并且呈现出不断提高的趋势，但是也有部分城市的空气质量问题有待改善。萍乡市和鹰潭市有轻度 $PM_{2.5}$ 污染，各城市 CO 和 $NO_2$ 治理一直较好，均属于优，其余污染物浓度均处于优良水平。

2）污染物浓度分布特征明显。大部分城市的空气主要污染物浓度除 $O_3$ 外，均呈现“冬季高，夏季低”的特点。所有城市 $O_3$ 的浓度分布差异不大，均处于优良，但是 $O_3$ 浓度有增加的趋势，说明未来江西省 $O_3$ 污染的治理是值得注意的。

3）通过随机森林模型可知，平均气温、平均气压和平均水气压对空气中主要污染物浓度影响最大，尤其对 $PM_{2.5}$、$PM_{10}$ 和 $NO_2$ 的影响；20-20 时降水量、平均 2 分钟风速、平均相对湿度和日照时数均对不同城市的污染物有着不同程度的影响。

## 参考文献：

[1] 刘艳艳.庆阳市供暖期空气质量及影响因素分析[J].宁夏师范学院学报, 2018, 39(1):60-65.
[2] 刘贺,李雪铭.中国城市空气质量时空演变及影响因素研究[J].生态经济, 2021, 37(9):91-101.
[3] 刘昕,辛存林.陕甘宁地区城市空气质量特征及影响因素分析[J].环境科学研究,2019, 32(12):2065-2074.
[4] 谢昆,陈博明.长沙市气象因素对大气污染的影响分析[J].矿冶工程, 2021, 41(4):170-175.
[5] 董铮.襄阳市大气污染物 时空分布特征及影响因素分析[J].农村经济与科技, 2020, 31(3):44-46.
[6] 张宸赫.沈阳大气污染物浓度变化及气象因素影响分析[J].环境科学与技术, 2020, 43(S2):39-46.
[7] 朱玲慧.新疆某高校宿舍空气质量变化规律及影响因素[J].中国学校卫生, 2020.8, 41(8):1264-1266.
[8] 王敏.黄河流域空气质量时空分布及影响因素分析[J].环境保护, 2019, 47(24):56-61.
[9] 赵东宇.安徽省城市空气污染的时空分布特征及影响因素分析[D].合肥:安徽财经大学, 2021.
[10] 廖秋虹,梁杰珍. 环境空气质量变化及其影响因素探讨[J].资源节约与环保, 2021(9):42-43.
[11] 王乐,田东方. 基于灰色关联分析法的宜昌市空气质量影响因素分析[J]. 能源环境保护, 2019, 23(5):60-64.
[12] 肖稚颖,丁俊伊. 北京市空气质量的影响因素分析——基于灰色关联[J].内蒙古科技与经济,2018, 408(14):50-51.
[13] 柳晓燕.城市环境空气质量问题及影响因素分析 [J].资源节约与环保, 2020,30(6):202-213.
[14] 秦聪.山西省空气质量时空格局及影响因素分析[J].环境与发展, 2019, 31(9):19-21.
[15] 赵英.京津冀大气污染的影响因素研究[J].资源节约与环保, 2020(11):39-40.