基于郑州市的 PM2.5 浓度与空气质量预测

王静怡 2017111550 数据科学与商务统计班

摘要

近年来，随着经济、工业和科技的发展，空气污染情况日益严峻。越来越多的“雾霾天”，给大众的生活和身体健康造成了巨大的影响，同时也为我国的生态文明建设带来了很多的难题。郑州市作为我国重要的交通枢纽和河南省的经济重心，空气污染程度在全国来看非常严重。同时，有关郑州市的 PM2.5 浓度以及空气质量预测的科学定量研究的文献比较少，因此建立郑州市 PM2.5 浓度和空气质量预测模型非常重要，以便为政府的决策和居民的日常生活防护提供指导和帮助。

以郑州市当天的 PM2.5 浓度为因变量，以郑州市前一天的 PM2.5、二氧化硫、一氧化碳等污染物浓度，和郑州市前一天的气象因素，以及郑州市相邻城市前一天的 PM2.5 浓度为自变量。通过从总体样本中随机抽取 20 天做测试集，其他样本为训练集。使用训练集来构建多元线性回归模型和随机森林回归模型，并使用测试集进行预测。通过预测折线图、RMSE（均方根误差）和 MAE（平均绝对误差）等指标对两个预测模型进行评估。

将空气质量等级中“优”和“良”两类定义为空气较好的一类，其余作为较差的一类，以该二分类变量作为因变量。郑州市前一天的 AQI 指数、各种污染物浓度和气象因素以及郑州市相邻城市前一天的 PM2.5 浓度作为自变量，以 7：3 的比例随机划分训练集和测试集，通过支持向量机模型和随机森林分类模型，来预测郑州市当天的空气质量等级，并以测试集预测的准确度和 ROC 曲线等指标来评价模型的拟合效果和泛化能力。

对 PM2.5 浓度进行预测，两个模型的训练集拟合优度接近 0.6，在测试集中，两个模型预测值的 MAE 值均为 20 微克/立方米左右，随机森林回归的效果稍好于线性回归模型。对空气质量等级进行二分类预测，两个模型的测试集预测准确率均为 0.8 左右，AUC 值均为 0.85 左右，分类效果较好。同时，二者的精确率较高，说明两个模型对正样本（即空气质量较差的一类）预测准确率较高。最终发现自变量中，对 PM2.5 浓度预测和空气质量等级预测影响较大的自变量有：郑州市前一天的 PM2.5 浓度，平均气温，相对湿度，平均风速，二氧化硫浓度和开封等城市前一天的 PM2.5 浓度。

关键词：多元线性回归 随机森林回归与分类 支持向量机 郑州市 PM2.5 浓度 空气质量等级

一、 引言

1. 研究背景

细颗粒物又称细粒、细颗粒、PM2.5，是导致雾霾的主要原因之一。细颗粒物是指环境空气中空气动力学当量直径小于等于 2.5 微米的颗粒物。与较粗的大气颗粒物相比，PM2.5 粒径小，面积大，活性强，易附带有毒、有害物质（例如，重金属、微生物等），且在大气中的停留时间长、输送距离远，因而对人体健康和大气环境质量的影响更大。2013 年 2 月，全国科学技术名词审定委员会将 PM2.5 中文名称命名为细颗粒物，主要化学成分包括有机碳 (OC) 、元素碳 (EC) 、硝酸盐 (NO3-) 、硫酸盐 (SO42-) 、铵盐 (NH[…])等[1]

相关研究表明，严重的 PM2.5 污染，对人的心肺功能健康有着极大的危害，严重时可危及生命。2012，年在北京、上海、广州、西安四城市因 PM2.5 等空气污染造成的早死人数将达 8000 多人。而在 2010 年，北京上海因空气污染导致早死的人数已经接近同期交通意外死亡人数的三倍。因此，通过对未来 PM2.5 的浓度以及空气质量等级的预测，来提醒群众注意相关方面的自我保护以及出行情况十分重要。

郑州市是河南省省会，是我国中部地区重要的中心城市，国家重要的综合交通枢纽。但是近些年来，随着城市工业化的发展，郑州市的空气质量状况越来越不容乐观，空气污染情况非常严重，严重影响着市民的健康和生活。在河南省发布的《2017 年全省环境监察执法工作要点》中，将“以减轻雾霾改善空气质量为重点，继续开展大气污染防治攻坚战环境监管专项执法检查，”列为第一条要点。由此可见河南省空气污染的严重程度。

2. 研究目的

目前，对郑州市未来的 PM2.5 浓度和空气质量等级的研究的相关文献比较少，仅有一些描述性的文章。郑州市作为河南省的经济中心，其发展也与空气质量息息相关。因此建立对郑州市空气状况的预测模型，对政府在环境治理中的决策和市民的日常出行及保护身体健康有着十分重要的作用。

3. 研究思路

考虑到气象因素对空气中的污染物浓度的影响，以及空气污染物的扩散作用导致的郑州市相邻城市的空气污染物扩散对郑州市的空气质量的影响。本文使用郑州市前一天的温度、湿度、风速、降水等气象因素、洛阳市、焦作市等六个与郑州市相邻的城市前一天的 PM2.5 浓度，以及郑州市前一天的 PM2.5、二氧化硫、二氧化氮等污染物的浓度，通过多元线性回归模型和随机森林回归模型，来建立对后一天郑州市 PM2.5 浓度的预测模型，并在测试集中评价预测值的拟合效果。之后，将“优”、“良”、“轻度污染”到“严重污染”等空气质量等级分为好和坏

两类，使用上述因素，通过支持向量机模型和随机森林分类模型，预测后一天的空气质量等级，并通过测试集上的预测准确度和 ROC 曲线来评价模型效果

二、 模型原理与方法

1. 多元线性回归模型

研究在线性关系相关性条件下，两个或者两个以上自变量对一个因变量的影响，为多元线性回归分析，表现这一数量关系的数学公式，称为多元线性回归模型。计算公式如下：

设随机 y 与一般变量 $x_1, x_2, \mathrm{L}\ x_k$ 的线性回归模型为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

其中 $\beta_0, \beta_1, \mathrm{L}\ \beta_k$ 是 $k+1$ 个未知参数， β_0 称为回归常数， $\beta_1, \mathrm{L}\ \beta_k$ 称为回归系数； y 称为被解释变量； $x_1, x_2, \mathrm{L}\ x_k$ 是 k 个可以精确可控制的一般变量，称为解释变量。

$k \geq 2$ 时，上式就叫做多元形多元回归模型。 ε 是随机误差，通常假设：

$$\begin{cases} E(\varepsilon) = 0 \\ Var(\varepsilon) = \sigma^2 \end{cases}$$

同样，多元线性总体回归方程为 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \mathrm{L} + \beta_k x_k$

系数 β_1 表示在其他自变量不变的情况下，自变量 x_1 变动到一个单位时引起的因变量 y 的平均单位。其他回归系数的含义相似，从集合意义上来说，多元回归是多维空间上的一个平面。

多元线性样本回归方程为： $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \mathrm{L} + \hat{\beta}_k x_k$

2. 随机森林回归模型和分类模型

1) 随机森林回归模型原理：

随机森林由 Leo Breiman(2001)提出．它通过自助法(bootstrap)重抽样技术，由随机向量θ(回归树)构成组合模型{h(X, θ_k)，k = 1．…，P}。预测变量为数值型变量,生成的随机森林为多元非线性同归分析模型。随机森林预测的形成是通过求 k 棵树的{h(X, θ_k)}的平均值，形成随机森林的训练集各自独立，选自随机向量 Y，X。数值型预测向量h(X)的推广误差均方为：

$$E_{X,Y}\big(Y-h(X)\big)$$

2) 模型特征

① 当森林中树的个数趋于无穷大时，有：

$$E_{X,Y}\big(Y-\alpha v_k h(X,\theta_k)\big)^2 \to E_{X,Y}\big(Y-E_\theta h(X,\theta)\big)^2$$

② 如果对于所有的θ，$(\mathrm{E}(\mathrm{Y})=E_X h(X,\theta)$，则：

$$\mathrm{PE}^*(forest)\le \bar{\rho}\mathrm{PE}^*(tree)$$

其中：$\mathrm{PE}^*(tree)$ $=E_\theta E_{X,Y}(Y-h(X,\theta))^2$

$\bar{\rho}$为剩余$Y-h(X,\theta)$和$Y-h(X,\theta')$间的权重相关，θ是独立的。

3) 算法实现：

①原始数据样本含量为 N，应用 bootstrap 有放回 地随机抽取 b 个自助样本集，并由此构建 b 棵回归树，每次 bootstrap 抽样未被抽到的样本组成了 b 个袋外数据(out—of-bag，OOB)，作为随机森林的测试样本；

②设原始数据的变量个数为 p，则在每一棵回归 树的每个节点处随机抽取m_{try}(m_{try}<<p)个变量作为备选分枝变量，然后在其中根据分枝优度准则选取最优分枝。在随机森林回归中,参数m_{try}=p/3；

③每棵回归树开始自顶向下的递归分枝，设定叶节点的最小尺寸 nodesize，以此作为回归树生长的终止条件；

④将生成的 b 棵回归树组成随机森林回归模型，回归的效果评价采用袋外数据(OOB)预测的残差均方：

$$MSE_{OOB}=n^{-1}\sum_{1}^{n}\{y_i-\hat{y}_i^{\,OOB}\}^2$$

$$R_{RF}^2=1-\frac{MSE_{OOB}}{\hat{\sigma}_y^2}$$

其中，y_i为袋外数据中因变量的实际值，$\hat{y}$为随机森林对袋外数据的预测值，$\hat{\sigma}_y^2$为随机森林对袋外数据预测值的方差。

4) 随机森林分类模型

如果 cart 树是分类数，那么采用的计算原则就是 gini 指数。随机森林基于每棵树的分类结果，采用多数表决的手段进行分类。

定义：基尼指数（基尼不纯度），表示在样本集合中一个随机选中的样本被分错的概率。Gini 指数越小表示集合中被选中的样本被分错的概率越小，也就是说集合的纯度越高，反之，集合越不纯。即基尼指数（基尼不纯度）= 样本被选中的概率 * 样本被分错的概率

其余的算法步骤类似于随机森林回归模型，这里不再赘述。

3. 支持向量机模型

支持向量机，主要用于解决二分类问题，可以支持线性和非线性的分类。与二维空间类似，超平面的方程也可以写成一下形式：：

$$\omega^T x + b = 0$$

其中$\omega = (\omega_1, \omega_2, \cdots, \omega_n)$为法向量，b 为位移项。样本空间中任意点 x 到超平面的距离为:

$$\mathrm{d} = \frac{|\omega^T x + b|}{\sqrt{||\omega||_2^2}} = \frac{|\omega_1 x_1 + \cdots + \omega_n x_n + b|}{\sqrt{\sum_{i=1}^{n} \omega_i^2}}$$

假设该超平面能够将训练样本正确分类，即对于（x_i, y_i） ∈ D,

令

$$y(\omega^T x_i + b) - 1 \geq 0$$

使该不等式成立的点被称为“支持向量”，两个异类支持向量到超平面的距离之和，即间隔为$\mathrm{m} = \frac{2}{||\omega||}$

即求如下有约束条件的最大间隔：

$$\max \frac{2}{||\omega||} \qquad s.t.\ y(\omega^T x + b) - 1 \geq 0$$

为方便计算，可将目标函数等价替换为：

$$\min \frac{||\omega||^2}{2} \qquad s.t.\ \mathrm{y}(\omega^T x + b) - 1 \geq 0$$

这是一个有约束条件的优化问题，通常可以用拉格朗日乘子法来求解：

$$\mathrm{L}(\omega, \mathrm{b}, \alpha) = \frac{1}{2}||\omega||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\omega^T x_i + b))$$

令L对ω和 b 求偏导,等于零，可化简得

$$\omega = \sum_{i=1}^{m} \alpha_i y_i x_i$$

$$0 = \sum_{i=1}^{m} \alpha_i y_i$$

化简上式和约束，可得原问题的对偶问题：

$$\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

s. t.

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\alpha_i \geq 0, \mathrm{i} = 1,2,\cdots,\mathrm{m}$$

最后可得到模型

$$\mathrm{f(x)} = \omega^T x + b = \sum_{i=1}^{m} \alpha_i y_i x_i^T x + b$$

而对于非线性的情况，SVM 的处理方法是选择一个核函数，通过将数据映射到高维空间，来解决在原始空间中线性不可分的问题。

在线性不可分的情况下，支持向量机首先在低维空间中完成计算，然后通过核函数将输入空间映射到高维特征空间，最终在高维特征空间中构造出最优分离超平面，从而把平面上本身不好分的非线性数据分开

本文采用高斯核函数。

三、 数据预处理与描述性统计

1. 数据来源

(1)空气污染基本信息

使用八爪鱼爬虫软件从中国空气质量在线监测分析平台历史数据查询网站爬取了郑州市、许昌市、焦作市、洛阳市、开封市、平顶山市、和新乡市从 2013 年 12 月 1 日到 2017 年 12 月 31 日每天的相关空气污染基本信息。其中信息包括：PM2.5,PM10,CO,SO2,NO2,O3 浓度、AQI 以及空气质量分等级的数据。 （网址为 https://www.aqistudy.cn/historydata/）

(2)气象基本信息

从全国温室数据系统下载了 2013 年 12 月 1 日到 2017 年 12 月 31 日郑州市每一天的平均气温、平均相对湿度湿度、平均风速等气象数据。
（网址为 http://data.sheshiyuanyi.com/WeatherData/ ）

2. 数据预处理

(1) 变量处理

每条样本的气象和污染物相关的自变量，为前一天的气象和污染物的观测值。因变量为但当天的 PM2.5 浓度。同时，按 PM2.5 浓度和 AQI 划分的空气质量等级[2]，如表 1 所示。

表 1：空气质量等级划分表

空气质量等级	PM2.5 浓 度 $/\mu g \cdot m^{-3}$	AQI
优	0~35	0~50
良	35~75	51~100
轻度污染	75~115	101~150
中度污染	115~150	151~200
重度污染	150~250	201~300
严重污染	＞250	＞300

将PM2.5浓度和AQI的取值范围划分为两类:优和良划为一类,标记为好类;轻度污染、中度污染、重度污染和严重污染划为另一类，标记为坏类。坏类的PM2.5浓度值和AQI值属于空气质量异常范畴，需要通过预报以提请人们注意和重点防范。[3]

最终变量如下所示

表2：变量展示表

序号	变量名	变量类型	变量解释
1	PM2.5	Numeric	郑州市当天PM2.5浓度
2	AQI	Numeric	郑州市当AQI指数
3	质量指数	Nominal	郑州市当天空气质量等级
4	ye_pm2.5	Numeric	郑州市前一天PM2.5
5	ye_wendu	Numeric	郑州市前一天平均温度
6	ye_shidu	Numeric	郑州市前一天平均湿度
7	ye_fengsu	Numeric	郑州市前一天平均风速
9	ye_jiangshui	Numeric	郑州市前一天累计降水量
10	ye_qiya	Numeric	郑州市前一天平均气压
11	ye_rizhao	Numeric	郑州市前一天平均日照时数5
12	ye_pm10	Numeric	郑州市前一天PM10浓度
13	ye_SO2	Numeric	郑州市前一天二氧化硫浓度
14	ye_CO	Numeric	郑州市前一天一氧化碳浓度
15	ye_NO2	Numeric	郑州市前一天二氧化氮浓度
16	ye_O3	Numeric	郑州市前一天臭氧浓度
17	ye_luoyang_pm	Numeric	洛阳市前一天PM2.5浓度
18	ye_xuchang_pm	Numeric	许昌市前一天PM2.5浓度
19	ye_jiaozuo_pm	Numeric	焦作市前一天PM2.5浓度
20	ye_xinxiang_pm	Numeric	新乡市前一天PM2.5浓度
21	ye_kaifeng_pm	Numeric	开封市前一天PM2.5浓度

22	ye_pingding_pm	Numeric	平顶山市前一天 PM2.5 浓度
23	ye_aqi	Numeric	郑州市前一天 AQI 指数
24	t	Nominal	郑州市当天空气质量好坏分类
25	ye_t	Nominal	郑州市前一天空气质量好坏分类

(2) 数据清洗

1) 缺失值与无效值的检测与基于 Knn 近邻算法的插补

从数据的基本情况检测，数据没有缺失值，但是空气污染数据的来源网站有些天的污染物数据显示为“无”和 0，导致郑州、洛阳等城市的数据是无效的。郑州市的空气污染数据作为因变量，有 58 个无效值。这些无效值的分布比较分散，为随机缺失，因此考虑将其删除。同时，考虑到本文需用前一天的污染物情况预测后一天的，因此有无效值的样本的后一天的自变量数据也无效，也需要删除。删除后，对剩余的自变量的无效值进行过诊断，结果是如下

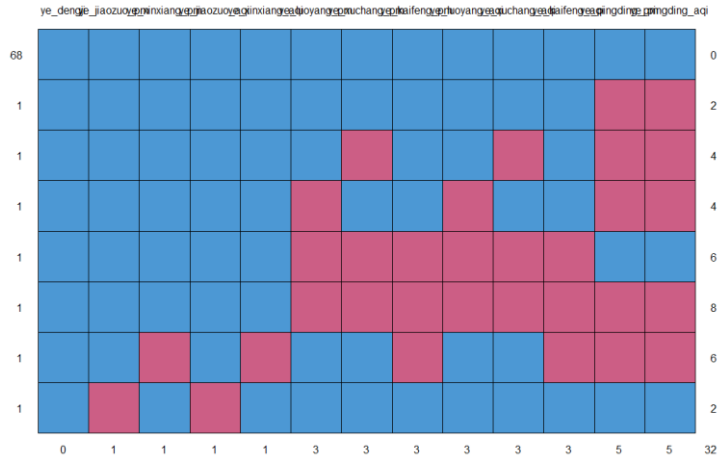

图 1：数据无效值诊断图

可以看到，洛阳、开封等其他城市仍有无效数据，但每个变量的无效数据的条数均不超过 5 条。

因此使用 KNN 近邻算法。先将郑州市当天 PM2.5 浓度、空气质量等级、AQI 等可以代表因变量的变量去掉，再对其他城市的 PM2.5 浓度缺失值进行插补，并按照插补后的 PM2.5 浓度划分其空气质量等级，防止影响后面监督学习模型的建立。KNN 算法首先计算目标数据Y_{mis}与所有完全观测数据$Y_{obs,i}$的距离d_i（常用欧式距离），所有完全观测数据中选取与目标数据距离最短的k个作为目标数据的最近邻，k个最近邻数据的加权平均值即为填补值，最终样本数为 1375 条。

2) 异常值点检测

做出所有变量的箱线图并分析，其中，部分异常值较多的变量箱线图如下所示：

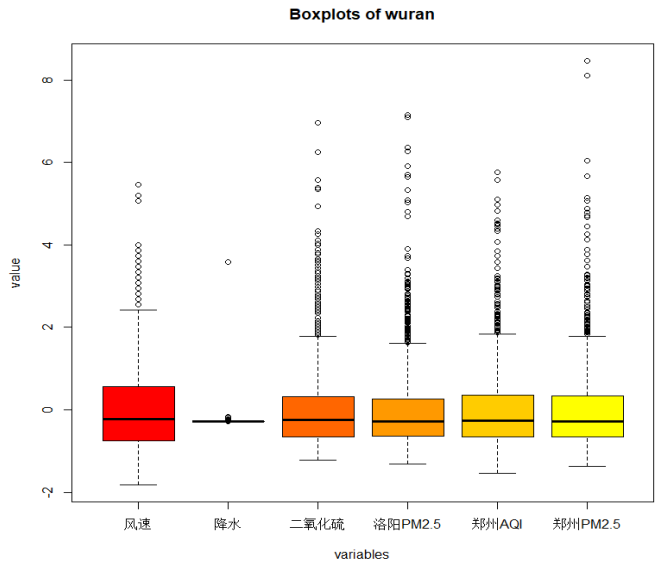

图 2：部分变量箱线图

可以看出洛阳等其他城市（未显示）的 PM2.5 浓度、AQI 指数以及二氧化硫等其他污染物浓度均有很多大的离群值。但是由于这些离群值可能是由重度污染或严重污染的天气导致的，有着重要的信息和价值，因此不处理，保留这些离群值。但是郑州市当天 PM2.5 作为因变量有过多的离群值可能会导致不满足多元线性回归正态性假设，因此在回归之前可能需要截取部分离群值点。同理，风速的离群值可能为极端大风天导致的，因此不做处理。而降水的箱线图四分位数几乎挤在了一起，有一些离群值，因为郑州市位于我国中北部地区，一年之间大部分天数是不下雨的，大多数降水量为 0，这是比较合理的，因此不做处理。

(3) 描述性统计分析

1) 线性回归的目标变量——郑州市当天的 PM2.5 浓度

a) 2017 年污染日历图

用 RStudio 做出 2017 郑州市污染严重程度日历图，可以看出严重的 PM2.5 污染大多集中在冬天的十二月、一月和二月，而春天和秋天的严重污染和重度污染的天数比较少。可能因为秋末冬初，郑州市开始集中供暖，燃烧产生的污染物较多，导致 PM 浓度升高，集中出现严重污染和重度污染的情况。同时，冬天地面夜间的辐射降温较明显，大气中低空比较容易出现“逆温层”，空气的水平、垂直方向流通能力变弱，排放的污染物被限制在浅层大气中，并逐渐集聚成霾，导致空气污染。

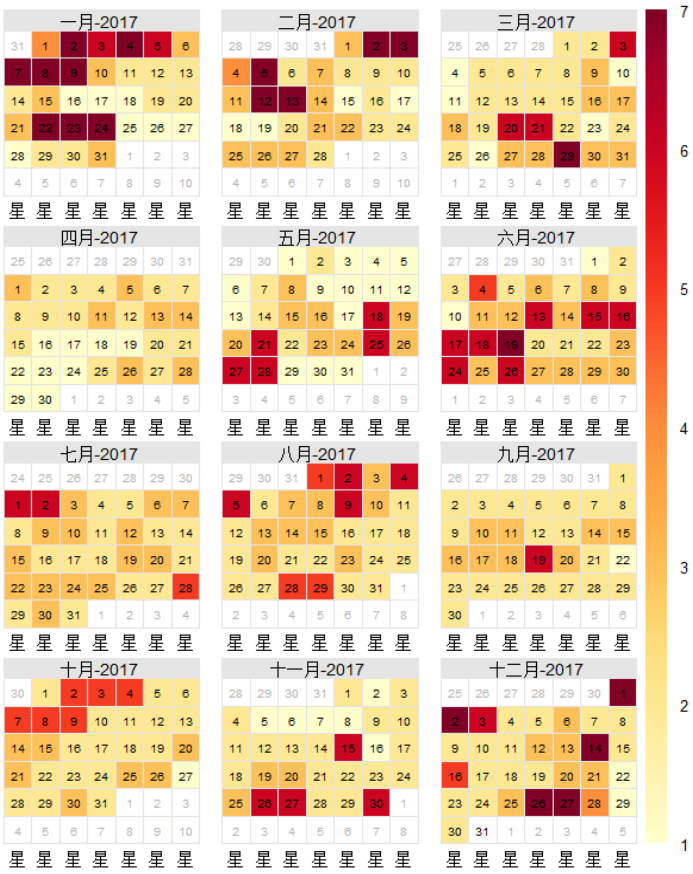

图 3：郑州市空气污染程度日历图

b)PM2.5 浓度时间序列图

进一步按不同年份画出从 2014 年初到 2017 年郑州市 PM2.5 浓度的时间序列图，如下所示。可以看出，这是一个有周期性规律的时间序列[4]，这四年间均有夏天 PM2.5 浓度较低，冬天 PM 浓度较高，且有较多极端值的特点。其中，这四年每年的 PM2.5 走势和平均浓度大致相同，且极端值并没有减少，在 2016 年年末，PM2.5 的浓度甚至超过了 600 微克/立方米，说明郑州市的治霾之路仍然任重而道远。

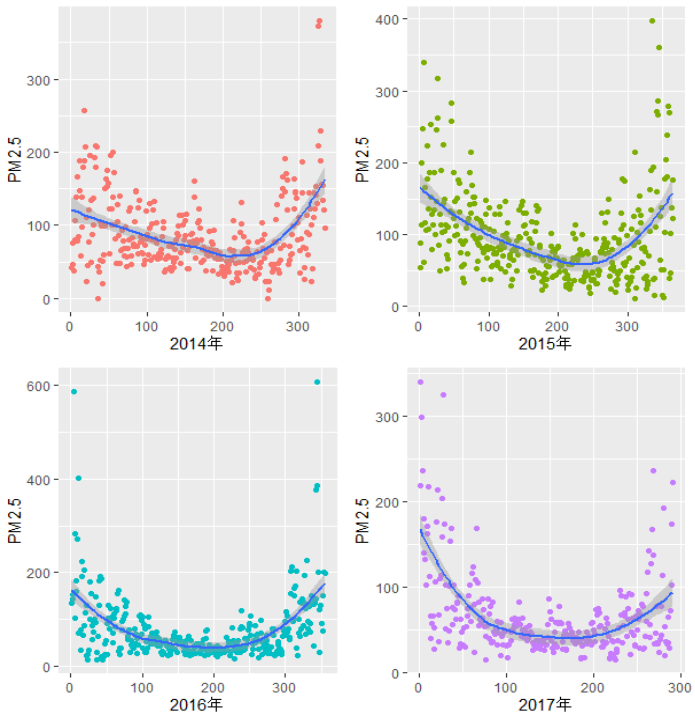

图 4：郑州市 PM2.5 浓度时间序列图

c） 测试集和训练集的划分

从样本中随机抽取 20 条样本作为测试集，剩余的样本作为训练集进行多元线性回归和随机森林回归。使用训练后的两种模型预测测试集中的样本，并建立相应的评价指标，将预测值与测试集中的目标变量进行比较，考察模型的泛化能力。

d） 目标变量正态性检验

画出郑州市当天的 PM2.5 浓度直方图，可以看出，其呈明显的右偏趋势，不符合正态分布。进一步做 Shapiro-Wilk 假设检验，检验其正态性。W 值为 0.78473，p-value < 2.2e-16，因此可以拒绝原假设，认为目标变量不符合标准正态分布。同时每隔 0.1%求因变量分位数，发现在 PM2.5 浓度高于 400 后，分位数变化巨大，影响正态性。因此，在做多元线性回归之前，考虑先去除前两个因变量为 0 的样本和后十个严重影响正态性的离群值点，此时所剩样本总数为 1355 条。

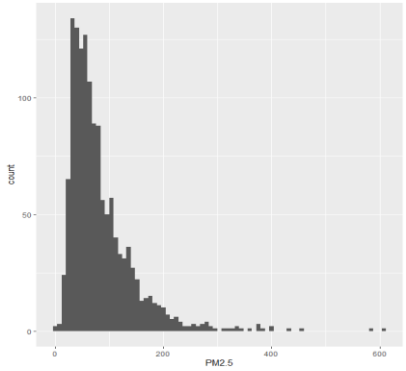

图 5：Box-cox 变换前因变量直方图

之后再对随机抽取的训练集做 Box-cox 变换，以满足多元回归分析中的正态性假设。以 λ 的步长为 0.01 作图，由图可得取 λ =0.14，正态性可以得到极大改善。做出经截取少部分离群值点和 Box-cox 变换过后的训练集目标变量的直方图，可以看出，训练集的目标变量的正态性得到了极大的改善。

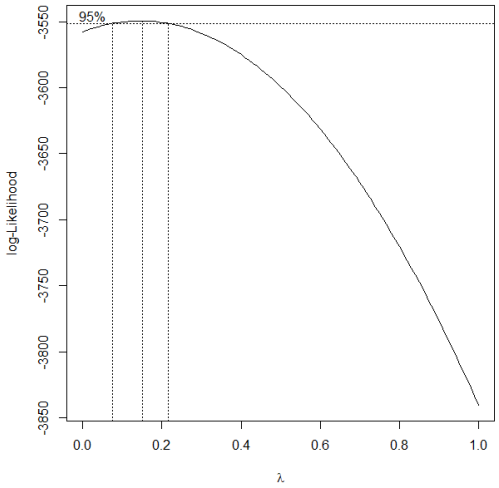

图 6：Box-Cox 变换参数 λ 选择图

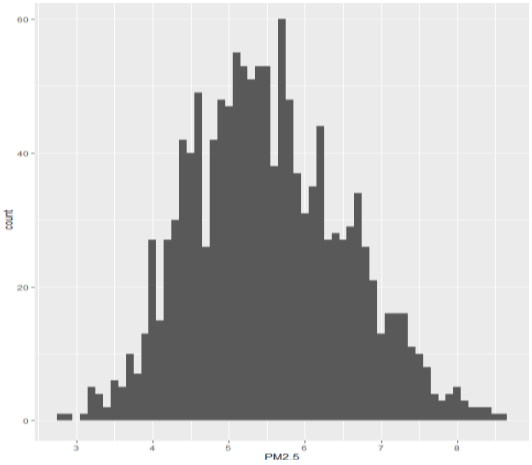

图 7：Box-cox 变换后因变量直方图

2) 关于空气质量状况的二分类预测的描述性统计

b)数值型变量直方图

用不同的颜色代表不同类别的因变量 t,其中红色代表当天的空气质量属于差类，做出如下所示部分比较有代表性的数值型自变量的直方图。

对于平均气温，可以看出当前一天郑州市平均气温处于较低水平时，郑州市当天的空气质量较差的概率远大较好的概率，当前一天气温适中时，当天空气质

量较好和较差的概率几乎相当，而当前一天的气温较高（大于 30 摄氏度时）郑州市当天空气质量较好的概率远大于较差的概率。就湿度而言，当前一天的湿度较大时，郑州市当天的空气质量较好的概率要远大于较差的概率，当前一天的湿度为中等或中等偏低时，郑州当天空气质量较好的概率稍低于较差的概率。就平均风速而言，当前一天风速较大，当天郑州市空气质量较好的概率要远大于较差的概率，当前一天风速较小时，当天郑州市空气质量较差的概率远大于较好的概率。

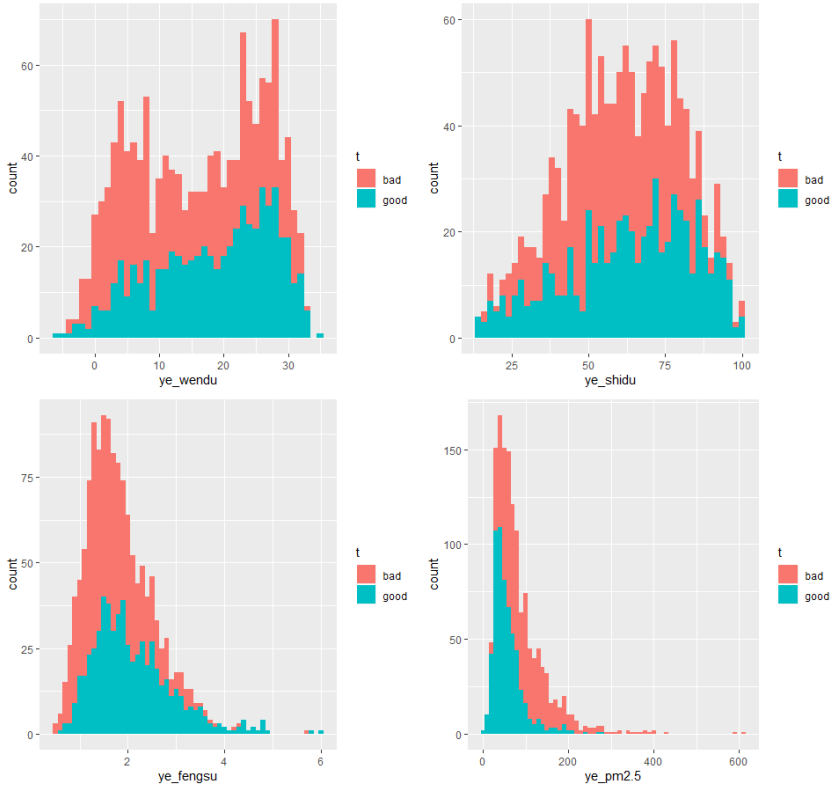

图 8：前一天气象因素和前一天 PM2.5 浓度直方图

就郑州市前一天 PM2.5、PM10、二氧化硫等污染物浓度和开封等其他城市前一天的 PM2.5 浓度而言，分别来看，当前一天这些污染物浓度偏高时，郑州市当天空气质量较差的概率要远高于较好的概率，当前一天这些污染物浓度偏低时，郑州市当天空气质量较好的概率要远高于较差的概率。

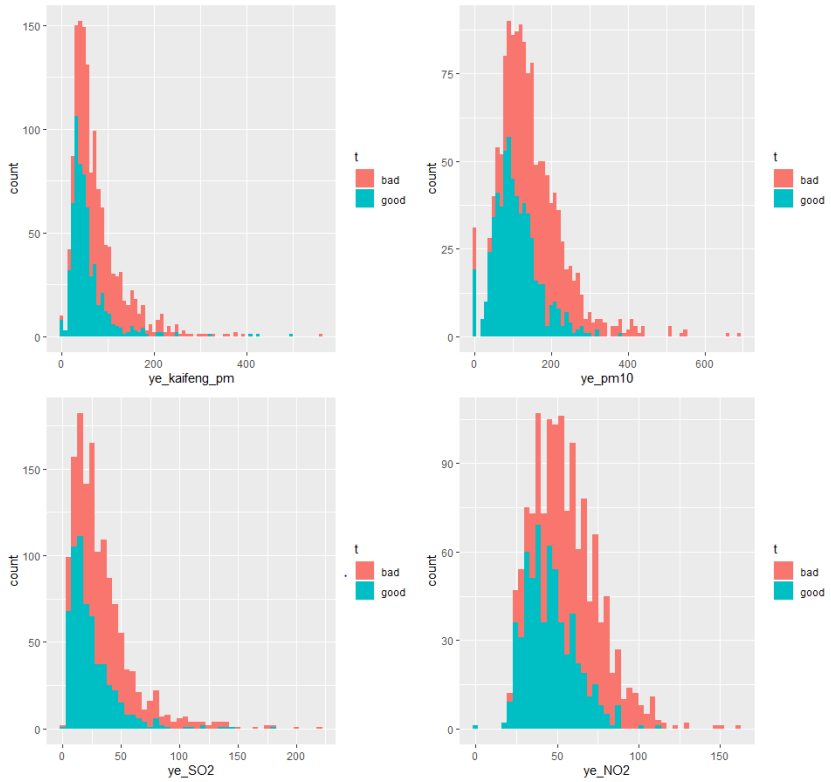

图 9：其他污染物前一天浓度和其他城市前一天 PM2.5 浓度直方图

四、 PM2.5 浓度预测模型的构建与结果分析

1. 多元线性回归

(一) 回归结果

多元线性回归模型显著性检验，F 统计量为 87.71，p-value<0.05，说明模型是显著的。对每个自变量的系数做 t 检验，其 p-value<0.05,显著的自变量为 ye_pm2.5、ye_wendu、ye_shidu、ye_fengsu、ye_qiya、ye_rizhao、ye_kaifeng_pm、ye_pingdi_pm、ye_xinxiang_pm。模型的拟合优度 R 方为 0.5492，调整后 R 方为 0.5429。

(二) 回归诊断

画出如下所示的残差图、正态 QQ 图等

1) 正态性检验

由正态 QQ 图可以看出，大部分点都分布在 45 度角的直线上，只有少部分点有所偏离。因此，认为经去掉了 12 个离群值点并做 Box-cox 变换后的训练集的样本，基本上符合模型正态性假设。

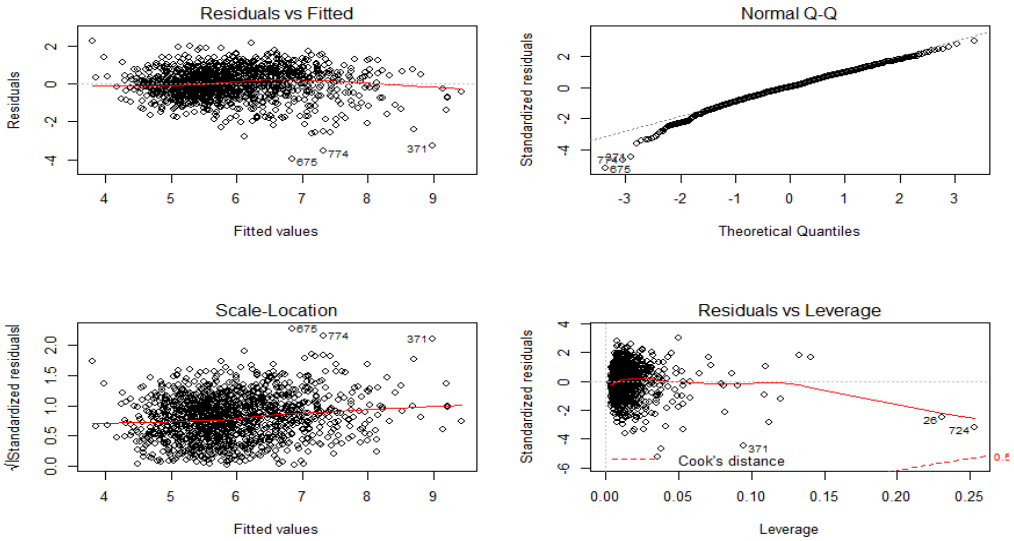

图 10：多元线性回归诊断图

2）模型线性

由模型的残差和拟合值的散点图可以看出，散点的分布没有明显的形状特征，且红线基本呈水平直线状没有其他有规律的形状特征，因此认为模型线性性基本满足。

进一步做出自变量的散点图，部分变量的散点图如下所示。可以看出平顶山市、焦作市和新乡市前一天的 PM2.5 浓度与因变量呈明显的非线性关系，因此考虑加上这三个变量的二次方项，再做回归。但是加上二次方项后，回归方程的拟合优度仅增加约 0.01，同时，出现多个方差扩大因子大于 10。

因此，加入二次方项后的模型效果并没有较大提升，反而出现了非常严重的多重共线性，因此不考虑在最终模型中增加二次方项。

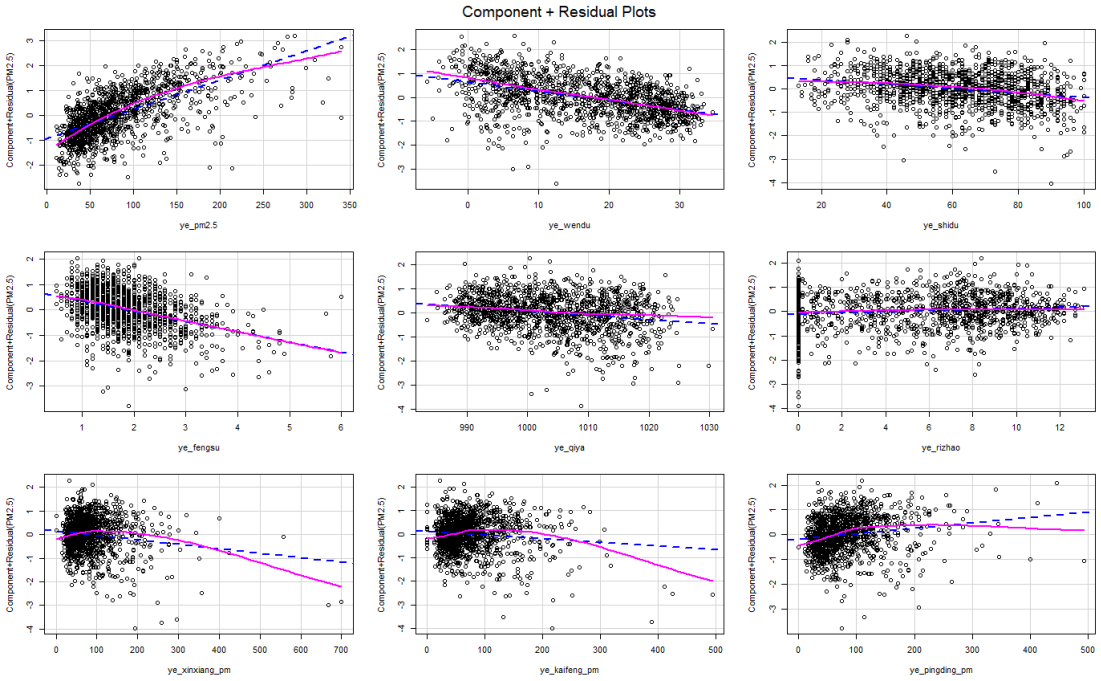

图 11：单变量残差散点图

3) 方差齐性

由残差和拟合值的散点图可以看出，散点基本上均匀地分布在 x 轴水平直线两边，没有特别呈喇叭口状，残差没有随拟合值的变化而有明显的增大或减小，因此可以认为模型的误差方差齐性是满足的。

4) 残差独立性检验

通过 DW 检验，来检验模型残差独立性。得到 DW 统计量为 2.038243，P-value 为 0.504>0.05，因此可以接受原假设，认为残差独立。

5) 异常值点和强影响点

由模型回归诊断图可以看出，存在 371、675、774 三个样本点的标准化残差大于 2，为异常值点。存在 26、371、724 三个样本点为强影响点，这些强影响点对模型的回归系数的估计具有很大的影响，考虑先不做处理。

6) 多重共线性检验

求出方差扩大因子，发现有一个方差扩大因子为 12.21567，超过了 10，因此模型的多重共线性是比较严重的。为解决多重共线性问题，考虑先通过进行逐步回归，做变量的筛选，考察能否解决模型的多重共线性。

2. 逐步回归

为筛选显著变量并解决模型的多重共线性问题，采用逐步回归进行变量筛选，逐步回归分析以 AIC 信息统计量为准则，来达到删除或增加变量的目的。最终得到的显著的变量为 ye_pm2.5、ye_wendu、ye_shidu、ye_fengsu、ye_qiya、

ye_rizhao、ye_kaifeng_pm、ye_pingdi_pm、ye_xinxiang_pm。模型 R 方为 0.5476 和调整后的 R 方为 0.5445。

最终得到回归方程为：

$$\begin{aligned} PM2.5 = {} & 0.0123ye\_pm2.5 - 0.0403ye\_wendu - 0.00996ye\_shidu \\ & - 0.42886ye\_fengsu - 0.01871ye\_qiya + 0.02311ye\_rizhao \\ & - 0.00186ye\_kaifeng\_pm - 0.00175ye\_pingdi\_pm \\ & + 0.00234ye\_xinxiang\_pm + 25.77665 \end{aligned}$$

1）回归诊断

经逐步回归筛选变量后，求出方差扩大因子，只有一个方差扩大因子为 6.0634，略微超过了 5，其他方差扩大因子均小于 5，因此模型的多重共线性有了极大的改善。

由整体的残差和拟合值的散点图可以看出，散点的分布没有明显的形状特征，且红线基本呈水平直线状没有其他有规律的形状特征，因此认为模型线性性基本满足。

由残差和拟合值的散点图可以看出，散点基本上均匀地分布在 x 轴水平直线两边，没有特别呈喇叭口状，残差没有随拟合值的变化而有明显的增大或减小，因此可以认为模型的误差方差齐性是满足的。

通过 DW 检验，来检验模型残差独立性。得到 P-value >0.05，因此可以接受原假设，认为残差独立。

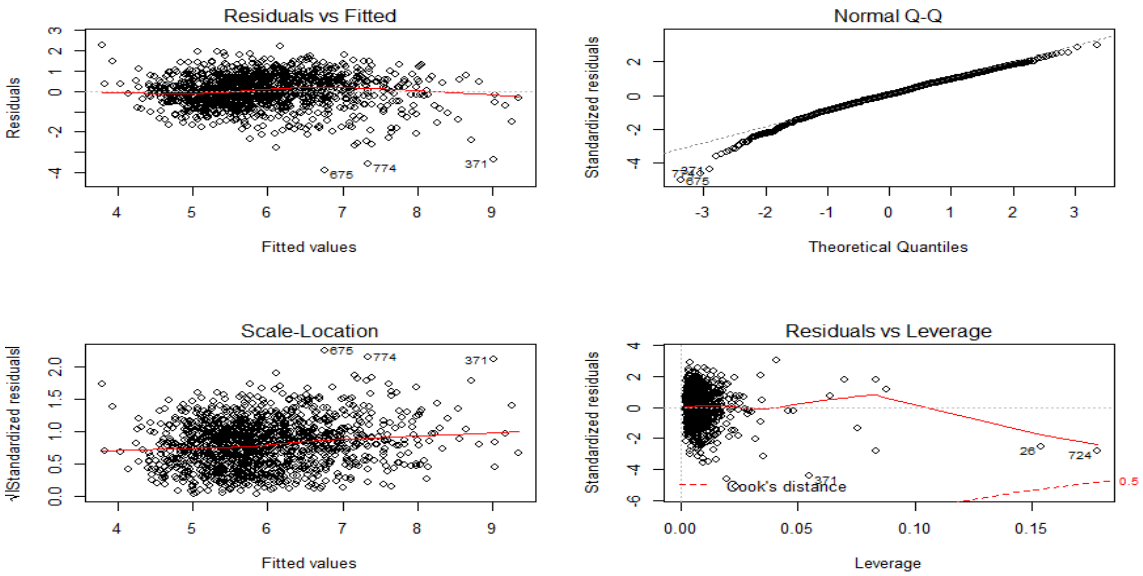

图 12：逐步回归诊断图

2）参数估计的结果实际含义分析

郑州市前一天 PM2.5 浓度的回归系数为正，该变量对郑州市当天 PM2.5 浓度

影响较大，PM2.5 颗粒物在空气中有运动、沉降等过程，短时间内难以完全扩散或消失,因此前一天 PM2.5 浓度较高会导致当天 PM2.5 浓度较高。郑州市前一天温度的回归系数为负。近地面气温较高时，大气对流作用加剧，可以降低 PM2.5 浓度；反之，大气出现逆温层时，PM2.5 不易扩散。同时冬天北方会集中供暖，燃烧可能导致 PM2.5 浓度的增加。郑州市前一天湿度的回归系数为负，相对湿度较高时，空气中的水分子遇到微粒物，会加重微粒物的自身重量，起到降沉的作用。因此，当空气湿度越大，空气中所含的水分子相对较多，从而可能使 PM2.5 浓度降低。郑州市前一天风速的回归系数为负。在风的作用下，PM2.5 浓度容易得到稀释，空气质量也会提升，尤其是大风天，加快了 PM2.5 的扩散。郑州市前一天气压的回归系数为负，郑州市前一天日照时数的回归系数为正，说明气压的升高可能会使 PM2.5 浓度降低,而日照时数的升高可能会导致 PM2.5 浓度的增大。而开封市、新乡市、平顶山市前一天 PM2.5 浓度也是显著的，说明在预测模型中相邻城市的 PM2.5 浓度对郑州市有影响,同时也说明不同城市间雾霾的扩散是不可忽视的，但是由于山脉等不同地貌和风速等因素的影响，不同城市间的影响是不同的。

3. 随机森林回归

设置不同大小的回归树棵树，做随机森林回归，当回归树棵树大于 400 后，模型比较稳定。因此设置回归树棵树为 600，每次选入回归树的特征数为 6。以郑州市当天 PM2.5 浓度为因变量，其他的变量为自变量，使用随机森林回归训练训练集。得到拟合优度为 55.5，稍高于多元线性回归。

得到以下特征重要性散点图,可以看出对郑州市当天 PM2.5 浓度影响最大的因素是郑州市前一天的 PM2.5 浓度。影响最大的三个气象因素为前一天的温度、前一天的湿度和前一天的气压。而由于地形和地貌的差异，对郑州市当天 PM2.5 浓度影响最大的是许昌市、开封市和平顶山市前一天的 PM2.5 浓度。影响最大的其他污染物为二氧化硫。

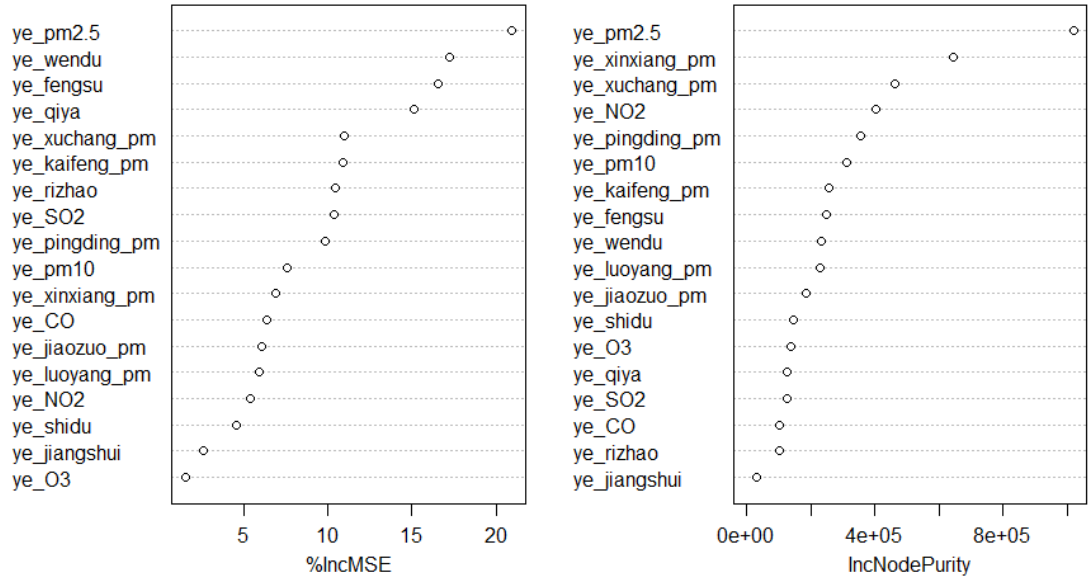

图 13：随机森林回归变量重要性图

4. 两种模型预测结果评估与比较

将上述使用训练集得到的多元线性回归模型和随机森林回归模型，分别用来预测测试集中的样本，考察模型的泛化能力，用测试集中郑州市当天 PM 浓度与分别使用两个模型拟合出的预测值做折线图，并使用平均绝对误差（MAE）和均方根误（RMSE）来衡量预测精度。

可以看出随机森林模型的预测结果，MAE 和 RMSE 均较小，要稍微优于多元线性回归模型。

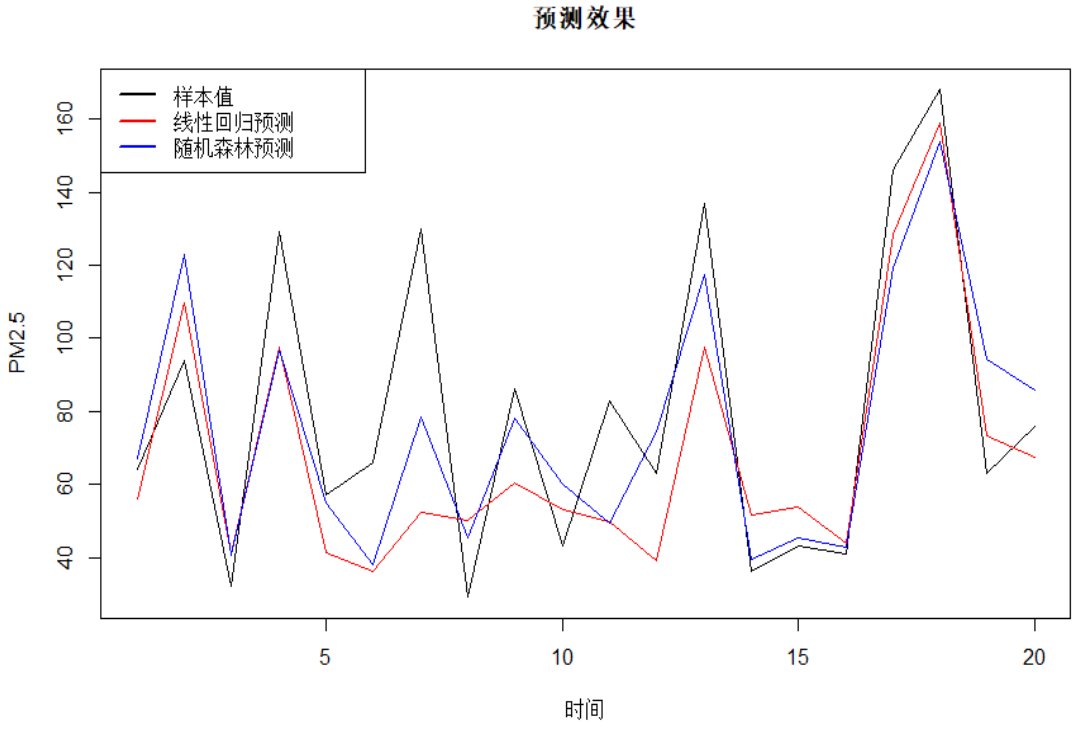

图 14：两种模型测试集预测图

表 3：两种模型测试集拟合评价表

	RMSE	MAE
线性回归	26.3962	20.7854
随机森林回归	21.9775	17.4915

五、空气质量等级预测结果

在两种 PM2.5 具体浓度预测的模型中，通过多次抽样并预测，发现当 PM2.5 浓度较高（大于 250 微克/立方米）的情况下，两个模型的对 PM2.5 浓度的预测效果均有较大幅度的下降，不能达到令人满意的效果。并且在多元线性回归中，为了满足模型正态性，截取掉了部分污染较严重的离群值点。因此为弥补这些缺点，考虑将空气质量等级分为两类，一类为属于空气质量异常的范畴，需要提醒人们注意防护，另一类属于空气质量良好的情况，来进行空气质量的二分类预测。

本文样本数较少，只有 1375 条，且两种类别比例较均衡，因此两种模型均把样本总体以 7：3 的比例随机抽样分为训练集与测试集，不做交叉验证。

1. 支持向量机模型

（1） 参数调整

本文使用高斯核函数，使损失惩罚参数 C 在 0.1,1,10,50,100 之间变换，KSVM 函数的高斯核函数参数 sigma 从 0.005 开始，以 0.004 为步长，到 0.3 来训练模型，并在测试集上进行预测，以 AUC 为指标评价模型。发现在参数 C 不变的情况下，随着 sigma 的增大，支持向量个数越来越多，训练集的分类准确率越高，但是测试集上的分类准确度越来越低，模型的泛化能力变差，模型会存在过拟合问题，且模型精确率与召回率的差异先增大后减小再增大。因此最终选择参数 sigma 为 0.05，C 为 10。

（2） SVM 模型结果

得到如图所示的 ROC 曲线和混淆矩阵，AUC 值为 0.84，精确率为 0.824，召回率为 0.738，测试集中准确率为 0.7912621。模型的分类效果比较好，尤其是精确率比较高，说明对污染较严重的坏类空气质量预测较准确，符合我们想要弥补多元线性回归缺点的需求，从而达到提醒群众注意防护的目的。

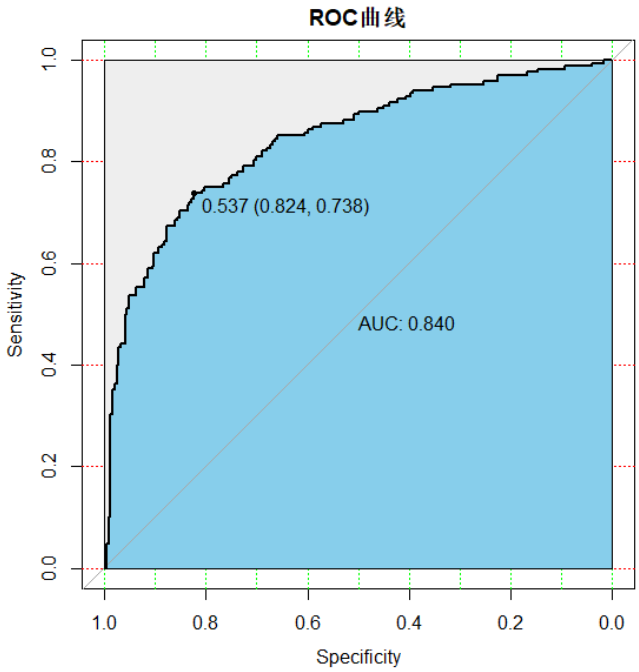

图 15：支持向量机模型 ROC 曲线

预测值	真实值 bad	good
bad	208	50
good	36	118

图 16：支持向量机模型混淆矩阵

2. 随机森林分类模型

(1) 参数调整

做出模型误判率与每次被选入的特征个数的折线图，可以看出当 mtry=14 和 24 时，模型误判率较小。同时做出模型误判率与决策树棵树之间的的折线图，可以看出当决策树棵树大于 1000 后，模型较稳定，因此这里选择参数 mtry 为 14，决策树棵树 ntree 为 1100，来构建随机森林。

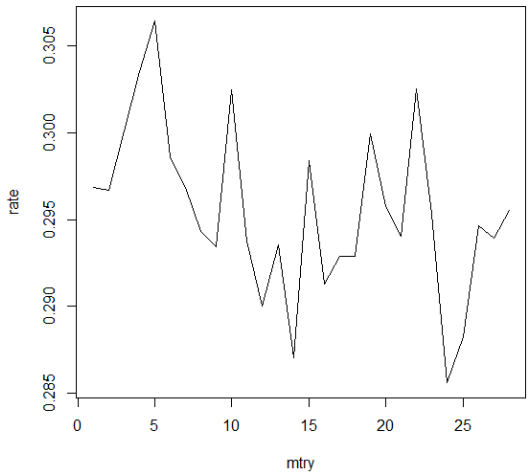

图 17：参数 mtry 选择图

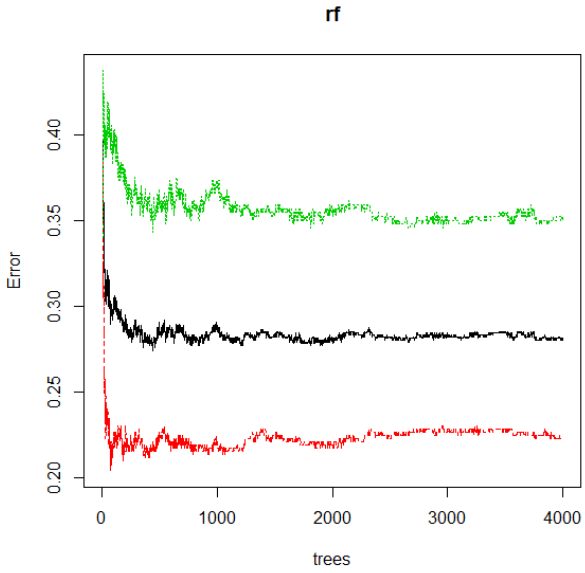

图 18：参数 ntree 选择图

(2) 随机森林模型结果

将训练集的分类效果降维，投影到二维空间中来，红色和蓝色分别代表空气质量的好类和坏类，可以看出两个分类效果比较清晰明显。由变量重要性散点图，对空气质量等级分类最重要的五个自变量分别为：郑州市前一天的 AQI 指数，郑州市前一天的二氧化氮浓度，郑州市前一天的平均风速，新乡市前一天的 AQI 指数，焦作市前一天的 AQI 指数。

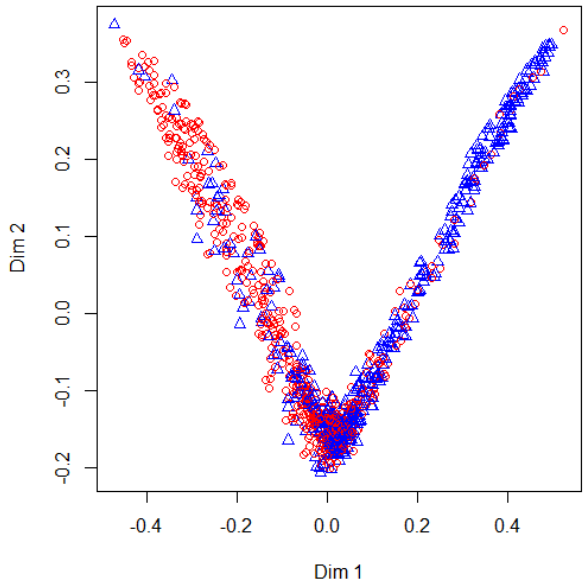

图 19：训练集分类情况降维图

输入变量重要性测度散点图

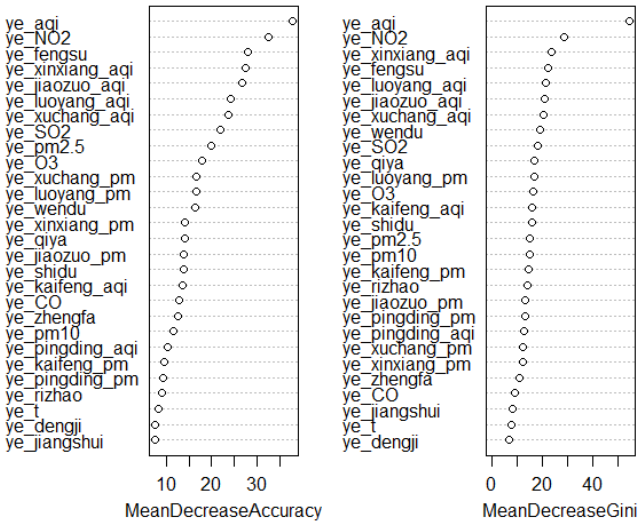

图 20：随机森林分类模型变量重要性图

做出如下所示的 ROC 曲线和混淆矩阵。模型的 AUC 值为 0.859，精确率为 0.828，召回率为 0.732，最终测试集上预测准确率为 0.788835。模型分类效果较好。

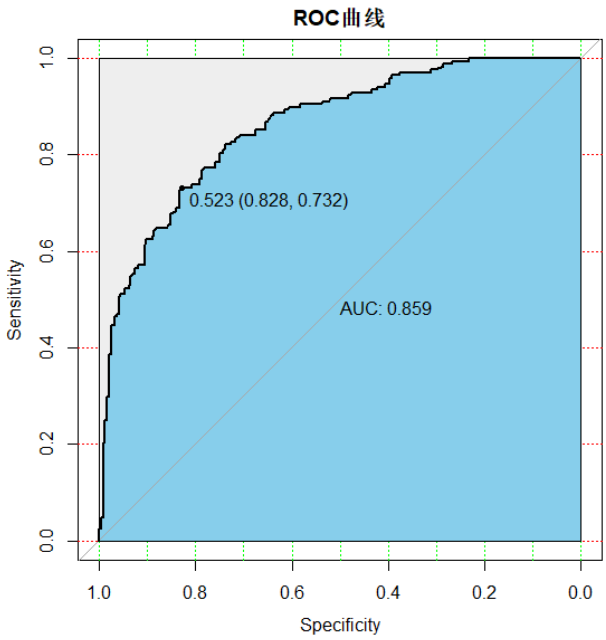

图 21：随机森林模型 ROC 曲线

预测值	真实值 bad	good
bad	202	45
good	42	123

图 22：随机森林模型混淆矩阵

3. 两种模型结果比较

可以看出两个模型的预测效果并没有太大差异，在 7：3 的情况下划分训练集和测试集后，在测试集上的预测准确率均接近 0.8。AUC 值均接近 0.85。同时，二者的精确率均比较高，说明在预测坏类空气质量（即轻度污染至严重污染的情况）时准确率相对较高，符合我们为了弥补多元线性回归缺点的需求，同时符合我们希望在污染较严重的情况下提醒群众加强防护的需求。

表 4：两种模型测试集分类效果评价表

	准确率	AUC	精确率	召回率
SVM	0.7912621	0.84	0.824	0.738
随机森林	0.788835	0.859	0.828	0.732

六、 模型缺点分析

1. 本文没有考虑到郑州市一些对 PM2.5 浓度和空气质量等级影响非常大的政

策性因素，如单双号限行等治理雾霾的措施。

2. 两种模型对 PM2.5 浓度的预测精度相对来说不是非常高，MAE 和 RMSE 均为 20 左右。因此还需进一步完善模型，来提高预测准确度。

七、启示和建议

1. 可以通过建立合理的多元线性回归与随机森林回归的 PM2.5 浓度预测模型，支持向量机与随机森林分类空气质量等级预测模型，通过前一天的气象因素和污染物浓度以及相邻城市前一天的污染物浓度来预测郑州市是当天的 PM2.5 浓度和空气质量做出预测，来为政府等部门的决策做出参考。
2. 相关部门在治理雾霾时，可以更加重视科学的定量的分析，建立更加精确的预测模型，找出对污染物浓度影响较大的因素，从而对症下药，制定合理的应对严重空气污染的政策，并加大对污染物排放的监管力度。
3. 可以适当的加强对群众应对雾霾的健康教育，并通过合理的空气质量等级预测模型，来给普通群众对于空气污染的预警，以达到保护人民群众身体健康的目的。

参考文献：

[1]白微静. 环境空气 PM2.5 危害分析与防护. [J]低碳世界, 2017

[2] 《环境空气质量指数（AQI）技术规定（试行）》（HJ 633—2012）

[3] 夏润，张晓龙. 基于改进集成学习算法的在线空气质量预测. [J]武汉科技大学学报, 2019

[4] 贾春光. 深度学习在 PM2.5 预测中的应用. [J]现代计算机, 2019