2017 届研究生硕士学位论文

分类号：_____ 学校代码：10269

密　级：_____ 学　　号：51143901091

華東師範大學

East China Normal University

硕士学位论文

MASTER'S　DISSERTATION

论文题目：基于随机森林的上海市 $PM_{2.5}$ 质量浓度预测研究

院　　系：地理科学学院

专　　业：地图学与地理信息系统

研究方向：空间数据挖掘

指导教师：过仲阳　教授

学位申请人：王雨晨

2017 年 5 月 12 日

Dissertation for master's degree in 2017

Student ID: 51143901091
University code:10269

East China Normal University

Tile:A Prediction Model of $PM_{2.5}$ Concentrations in Shanghai based on Random Forest

Department:School of Geographic Sciences

Major:Cartography and Geographic Information System

Research direction:Spatial Data Mining

Supervisor:*Prof.* Zhongyang Guo

Graduate:Yuchen Wang

May, 2017

华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《基于随机森林的上海市 $PM_{2.5}$ 质量浓度预测研究》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名：王雨晨　　日期：2017 年 05 月 16 日

华东师范大学学位论文著作权使用声明

《基于随机森林的上海市 $PM_{2.5}$ 质量浓度预测研究》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的研究成果归本人所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和相关机构如国家图书馆、中信所和“知网”送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于（请勾选）

（　）.经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（√ ）.不保密，适用上述授权。

导师签名　过仲阳　　本人签名　王雨晨

2017 年 05 月 16 日

* “涉密”学位论文应是已经华东师范大学学位评定委员会办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权）。

王雨晨硕士学位论文答辩委员会成员名单

姓名	职称	单位	备注
吴健平	教授	华东师范大学	主席
王军	教授	华东师范大学	
李响	教授	华东师范大学	
王占宏	教授级高工	上海众恒信息产业股份有限公司	
陆衍	教授级高工	上海市地质调查研究院	

摘 要

随着我国经济的快速发展，工业化和城市化进程加快，以 $PM_{2.5}$ 为主的大气污染问题日益突出。雾霾等大气污染问题不仅严重影响了人们的日常生活和身体健康，也对社会的可持续发展造成了巨大冲击。因此，实现对大气中污染物浓度的精准预测具有重要的现实意义和社会价值。

本文从以下几方面对上海市 $PM_{2.5}$ 质量浓度进行了研究：

首先，对于原始数据中存在的缺失值，本文综合考虑了其他因素对于缺失值的影响，使用 KNN 算法对其进行了填充，填充后的数据与原数据的波动规律基本吻合。

其次，本文从月际和周际的时间尺度分析了上海市 $PM_{2.5}$ 的分布变化规律，分别总结出上海市每月平均 $PM_{2.5}$ 浓度变化、每月空气质量等级占比、一周内每一天的 $PM_{2.5}$ 浓度变化。

然后，分析了 $PM_{2.5}$ 与其他污染物、$PM_{2.5}$ 与气象因素之间的相关关系，并计算其 Pearson 相关系数矩阵，确定了 $PM_{2.5}$ 与其他因子间相关性的大小与方向。同时，建立以赤池信息准侧为判停标准的逐步回归仿真方程，并对回归假设做出诊断。实验表明，气象因子和离子浓度因子的加入使逐步回归仿真方程的拟合优度从 66%提升至 85%。

最后，利用随机森林算法建立用于 $PM_{2.5}$ 质量浓度预测的逐小时模型和极值模型，分别对未来 1~6 小时每小时的 $PM_{2.5}$ 质量浓度以及 6~12 小时、12~24 小时、24~48 小时的 $PM_{2.5}$ 质量浓度最大值和最小值进行预测。实验表明，随机森林算法的预测精度在 90%以上，相较于基准模型，精度最大提升 30%。通过基于 OOB 误差估计的变量筛选方法选择出最优预测变量子集，可以使模型的拟合优度平均提升 1.05%。

关键词： $PM_{2.5}$，逐步回归，OOB 误差估计，随机森林

ABSTRACT

With the rapid development of China's economy and the acceleration of industrialization and urbanization, air pollution problems which mainly about $PM_{2.5}$ are becoming more and more serious. Haze and other air pollution problems not only seriously affect people's daily lives and physical health, but also has a great impact on sustainable development of society. Therefore, it has great practical significance and social value to accurately predict the concentration of atmospheric pollutants

The concentration of $PM_{2.5}$ in Shanghai is studied from the following aspects in this paper.

Firstly, this paper comprehensively considers the influence of other factors on the deficiencies in the original data, and uses the KNN algorithm to fill them, fluctuation after filling the data and the original data are consistent.

Secondly, this paper analyzes the distribution of $PM_{2.5}$ in Shanghai from the monthly and weekly time scales, and respectively summed up the monthly average concentration change of $PM_{2.5}$, the monthly air quality ratio, the $PM_{2.5}$ concentration changes per day in a week.

Thirdly, the correlation between $PM_{2.5}$ and other pollutants and between $PM_{2.5}$ and meteorological factors was analyzed. Then the Pearson correlation coefficient matrix was calculated, and the magnitude and direction of the correlation between $PM_{2.5}$ and other factors was identified. At the same time, the stepwise regression simulation equation is established to determine the stopping criterion of the Akaike information.The experimental results show that the goodness of fit of the stepwise regression equation is increased from 66% to 85%.

Finally, the hourly model and the extreme value model for $PM_{2.5}$ mass concentration prediction were established by using the random forest algorithm. Respectively predict the $PM_{2.5}$ concentration per hour in the next 1 to 6 hours, and predict the maximum and minimum $PM_{2.5}$ concentration in the next 6 to 12 hours, 12

to 24 hours, 24 to 48 hours. The experimental results show that the prediction accuracy of the random forest algorithm is more than 90%, compared with the reference model, the accuracy is improved by 30%. By selecting the optimal subset of variables based on the OOB error estimation method, the goodness of fit of the model can be increased by 1.05%.

Keywords: *[$PM_{2.5}$] [stepwise regression] [OOB error estimation] [random forest]*

插图清单

附表清单

目 录

第一章 绪论

1.1 研究背景和意义

随着我国大气污染模式的转变，雾霾天气频繁出现，空气质量等级六级以上的严重污染天气时有发生，$PM_{2.5}$已经成为我国大气污染的首要污染物之一。上海气象局数据显示，近年来长三角地区悬浮颗粒物污染不论是浓度还是影响范围都有逐年上升的趋势。

粒径，也就是空气动力学当量直径，是大气环境中悬浮颗粒物种类划分的重要依据之一。悬浮颗粒物按其粒径大小可分为 TSP、PM_{10}、$PM_{2.5}$等。其中，空气动力学当量直径小于或等于 100μm 的称为 TSP，又称总悬浮颗粒物；空气动力学当量直径小于或等于 10μm 的称为 PM_{10}，又称可吸入颗粒物；空气动力学当量直径小于或等于 2.5μm 的称为 $PM_{2.5}$，又称细颗粒物。可以看出，颗粒物粒径的划分是存在包含关系的。

在多种污染悬浮颗粒物中，颗粒物的粒径越小，其对人体健康的危害就越大。这是因为细小的颗粒物能够渗透进人体的上呼吸系统和下呼吸系统，使得颗粒物上附着的有害物质可以轻易地侵入人体内。大量研究证明，长期暴露在 $PM_{2.5}$污染的环境中不利于人体健康，极端情况下甚至会引发人群呼吸系统、心血管系统、神经系统、免疫系统等多个系统的疾病[33]。此外，大气中高浓度的悬浮颗粒物污染也将引发许多潜在的环境问题。例如，大气能见度降低[21]、植物物种的生长发育迟缓或死亡[4]、气候变化[11]、文物古迹损坏[24]等。

大气中悬浮颗粒物的来源主要包括自然源和人为源。其中，自然因素会造成 $PM_{2.5}$浓度升高的有：土壤风沙尘、海盐、火山喷发等，人为因素会造成 $PM_{2.5}$浓度升高的有：机动车尾气排放、火力发电、工业生产排放、钢铁煤源、餐饮及露天焚烧等。此外，自然源和人为源的前驱物排放转化，也可能导致 $PM_{2.5}$浓度上升。例如，一次污染物 SO_2转化成二次污染物硫酸盐[18]、一次污染物 NO_x转化成二次污染物硝酸盐[27]等。

如今，大部分机构都将 $PM_{2.5}$质量浓度监测数据纳入到了空气质量指标当中，如：污染物标准指数（Pollutant Standards Index，PSI）、空气质量指数（Air Quality

Index，AQI）等。这些指数的建立是为了给公众一个空气污染程度的参考标准，并且，指数的每个级别预示着空气污染对健康的影响大小。2012 年 12 月，美国环境保护局（EPA）决定修改 $PM_{2.5}$ 浓度在 AQI 中的地位，以此来完善其空气质量标准体系。具体来说，就是将空气质量“良好”等级的上限从 15μg/m^3 降低至 12μg/m^3。虽然新标准的差别只有 3μg/m^3，但这表明了 EPA 认为之前做出的 $PM_{2.5}$ 对公众健康影响严重程度的判断可能估计不足。另一方面，3μg/m^3 的差距也表明，$PM_{2.5}$ 质量浓度预测模型需要拥有更加精确的预测能力，以满足日益严格的新标准。目前，EPA 的 $PM_{2.5}$ 污染物空气质量三级标准为：小时质量浓度在 0-12μg/m^3 之间时，空气质量定为“良好”；小时质量浓度在 12.1-55.4μg/m^3 之间时，空气质量定为“敏感”；小时质量浓度在 55.5μg/m^3 以上时，空气质量定为“有害”。

$PM_{2.5}$ 是一种由空气中固体和液体的悬浮颗粒物混合而成且分布广泛的大气污染物。相较于传统污染物，$PM_{2.5}$ 的污染范围超越了一般的地理界线。所以，有效治理 $PM_{2.5}$ 污染是一个全球化的问题，急需一种跨学科的方案予以解决。在治理 $PM_{2.5}$ 污染的过程中，对于未来一段时间内污染物浓度的预测是对大气污染物进行有效控制和管理的重要前提之一，及时、准确地预测未来大气中的污染物浓度，也将有助于控制大气污染以及规划大气质量管理过程。因此，城市范围的大气污染物预测是十分必要的。

1.2 国内外研究现状

1.2.1 $PM_{2.5}$ 预测研究进展

空气污染是影响人类健康的一大全球性问题[29]，已经引起了政府和学术界的广泛关注。随着人们环保意识和健康意识的提高，$PM_{2.5}$ 浓度成为人们日常关注的重要信息之一。$PM_{2.5}$ 的准确预测将有助于有关部门的环境管理和提高公众防范意识，这也成为国内外学者的研究热点。

目前已经有多种算法模型被国内外学者应用于预测大气中的颗粒物浓度，并且被证明能够取得比较好的结果。这些模型的主要类型可以分为：时间序列模型、化学传输模型、线性或非线性回归、数据挖掘模型等。

其中，用于预测空气质量的时间序列模型包括：移动平均模型（Moving Average, MA）、自回归模型（Autoregressive, AR）、自回归移动平均模型（Autoregressive Moving Average, ARMA）、自回归积分移动平均模型（Autoregressive Integrated Moving Average, ARIMA）等。Jian 等利用 ARIMA 模型成功预测了街道范围 $PM_{1.0}$ 浓度[12]。余辉等使用 ARMAX 模型对 $PM_{2.5}$ 小时浓度进行预测[42]，并提出了一种基于模型精度的实时更新方法。

研究表明，化学传输模型在预测大气中颗粒物浓度时，其精确性不如 ARMA 或 ARIMA[23]，并且计算的复杂度较高。

得益于回归模型的可解释性（这种可解释性体现在人们可以通过回归方程直观地知道污染物与模型其他变量之间的系数关系），构建多元线性回归方程目前任然是一种被普遍使用的颗粒物预测建模思路。孙云海等使用多元线性回归拟合了前日颗粒物浓度在当日气象条件下与当日颗粒物浓度的关系[39]。在训练模型时，分别对春、夏、秋、冬四季建立回归方程，其预测数值的准确率在 58.1%-70.2% 之间，预测等级的准确率在 52.1%-77.5%之间。付倩娆通过对大气污染颗粒物和气象数据间拟合多元线性回归方程[32]，预测未来一天、三天及七天的日均 $PM_{2.5}$，命中率在 81.67%-85%之间。崔寒等对比了 BP 神经网络与逐步回归在大雾预报上的性能，并建立了能见度模型和大雾判别模型[30]。

数据挖掘方面，由于人工神经网络模型结合有效的训练算法可以检测到预测变量与响应变量间复杂的潜在非线性关系，该模型成为当前的主流。Gholamreza 等使用人工神经网络和马尔科夫链，将 PM_{10}、NO、NO_2、NO_x、CO、SO_2 作为输入，对每小时的 $PM_{2.5}$ 浓度进行仿真[10]。Maher 等对比了多元线性回归和前馈反向神经网络在预测不同季节室内通风环境中空气质量时的性能[16]。研究表明，前馈反向神经网络的预测精度要高于多元线性回归，在秋季、冬季和春季，室内 $PM_{2.5}$ 浓度预测精度分别为 75%、78%、79%，室内 $PM_{2.5-10}$ 浓度预测精度分别为 65%、73%、78%。Bun 等提出深度递归神经网络模型用于预测 $PM_{2.5}$ 浓度[6]。模型使用的是基于时间序列预测自动编码的动态预训练方法，此外，传感器的选择是在不降低预测精度的基础上利用网络稀疏获得的。Shadi 等评估了自适应模糊

神经网络推理系统（ANFIS）、整体经验模态分解和广义回归神经网络混合模型（EEMD-GRNN）、主成分回归（PCR）、多元线性回归（MLR）用于 $PM_{2.5}$ 预测时的性能[22]。模型中使用到数据包括空气质量数据（$PM_{2.5}$、PM_{10}、SO_2、NO_2、CO、O_3）和气象数据（气温、气压、降水量、相对湿度、风速），实验结果表明，EEMD-GRNN 模型 0.79 的 R^2 表现最优。从上也可以看出，在预测污染物浓度时，气象因素是除颗粒物因素外最常被考虑纳入模型中的变量之一。

1.2.2 随机森林算法研究进展

随机森林算法是 Breiman 于 2001 年提出的，是一种基于分类回归树的组合模型。每棵分类回归树使用自助样本生成，并且每一次分裂的候选变量集是所有变量的随机子集。因此，随机森林算法结合了 CART、随机属性选择以及装袋（bagging）的算法思想。每棵分类回归树都完全生长，从而得到了低偏移的树，同时，随机属性选择和装袋的方法使得随机森林中的每个个体其相关性都较低。随机森林算法建立的模型具有较低的偏移和方差[19]，该算法适用于解决分类和回归问题[37]。相对于其他算法中常遇到的过度拟合和多重共线性问题，随机森林对其均不敏感。

随机森林算法只需要进行少量的参数调整就可以实现较好的性能。在随机森立的所有参数中，对模型性能影响最大的是随机属性选择的个数。李毓等通过 OOB 数据对该参数的最优值进行了估计[38]。同时，随机森林特有的变量筛选机制也促进了多种变量选择算法的提出。姚登举等提出的 RFFS 算法[40]，利用随机森林对变量重要性的排序，采用后向搜索的方法筛选变量集合。尹华等提出 IBRFVS 变量选择算法[41]，解决了在高维类不平衡数据上的变量选择问题。

随机森林算法近年在国内外来发展迅速，在生物医学、遥感、经济等各领域得到广泛应用。

在生物医学方面，Gaspar 等使用随机森林对激酶配体、核激素受体及其他酶的分子描述符进行自动筛选[9]。Liu 等通过建立随机森林模型对微生物细胞表面疏水性进行预测[15]。

在遥感方面，Vogels 等利用随机森林方法从黑白航片中识别出耕地[26]，识别

精度在 90%以上。Rodriguez 等使用随机森林算法对 Landsat-5 遥感影像的 14 中不同地类进行分类[20]，达到 92%的分类精度，Kappa 指数为 0.92。Onisimo 等使用 WorldView-2 卫星影像和随机森林算法估算了湿地植被高密度生物量[17]，该方法相比多元线性回归误差降低了 3%。Li 等利用随机森林算法通过机载激光雷达和多光谱影像数据对城市场景分类[14]，实现了 95%的总体准确性，但在密度较大的场景中，准确性在 70%左右。

在经济方面，Bart 等实现了通过随机森林预测客户的留存率和盈利率[3]，Xie 等在解决类不平衡问题的基础上使用随机森林预测了银行客户的流失率[28]。利用随机森林的变量评估机制，林成德等确定了一个企业信用指标体系[35]。

在其他方面，Anush 等通过随机森林方法实现了自适应的潜在指纹分割[2]。张华伟等将随机森林用于文本分类[43]。甄亿位等利用随机森林算法对中长期降水量进行了预测[44]。

由于随机森林算法准确度高、防止过拟合能力强等特点，并且在许多研究实验中表明随机森林算法优于神经网络、支持向量机等其他算法，因此，该算法目前已经有了非常广泛的应用，本文也将使用随机森林算法对 $PM_{2.5}$ 质量浓度的预测进行研究。

1.3 研究内容

本文的研究目标是通过统计分析和数据挖掘等方法，对 1~6 小时的 $PM_{2.5}$ 质量浓度逐小时进行预测，对 6~12 小时、12~24 小时、24~48 小时进行最大值和最小值的预测。

首先，对研究区域内 $PM_{2.5}$ 质量浓度的总体变化趋势和变化规律进行分析，在各时间尺度上把握 $PM_{2.5}$ 质量浓度值的分布特征。

其次，对于污染物因子和气象因子两种主要因子与 $PM_{2.5}$ 之间的相关性进行分析，同时确定各相关性的大小和方向。

再次，建立 $PM_{2.5}$ 与其他污染物、离子、气象条件等因子的逐步回归仿真方程，定量描述 $PM_{2.5}$ 与各因子之间的关系。

最后，利用随机森林算法建立 $PM_{2.5}$ 质量浓度的 1~6 小时逐小时预测模型和

6~12 小时、12~24 小时、24~48 小时的极值预测模型，同时，尝试通过变量筛选的方法进一步提高随机森林的预测准确性，并给出验证结果。

1.4 创新点

（1）本文使用随机森林算法对上海地区大气中的 $PM_{2.5}$ 颗粒物浓度进行了定量预测，并在预测时综合考虑了离子、气象、周期、预报等多维多角度的因素；

（2）使用 OOB 误差估计的方法对输入随机森林预测模型的各变量进行了筛选，从理论上为变量选择方法提供了依据；

（3）对于 $PM_{2.5}$ 质量浓度预测的时间尺度缩短至小时，同时，预测目标包含了未来一定时间范围内 $PM_{2.5}$ 浓度的具体数值以及变化区间。

1.5 论文结构

本文共分为 6 章，具体结构如下：

第一章为绪论，论述了 $PM_{2.5}$ 质量浓度预测的选题背景和意义，并介绍了国内外在 $PM_{2.5}$ 污染物预测和随机森林算法研究上的最新进展。同时，提出了本文的研究内容以及论文的基本框架。

第二章介绍了随机森林的研究方法，包括其原理和性质，从理论上证明了随机森林模型的泛化误差收敛性。解释了如何利用随机森林的 OOB 误差估计对模型的输入因子进行筛选，使得模型复杂度降低的同时提高其准确性。另外，给出了随机森林算法的具体实现步骤，针对离散型和连续型两种不同的响应变量，对应给出了各自具体的 CART 算法。

第三章介绍了研究区域和数据来源，本文数据均来自上海市徐汇区的实测数据。对原始数据中存在缺失值的情况，使用 KNN 算法进行了补充，并展示补充效果。在原有输入因子的基础上加入了周期因子和预报因子。最后，具体分析了 $PM_{2.5}$ 质量浓度在各时间尺度上的变化规律。

第四章首先介绍了相关分析的原理，并对 $PM_{2.5}$ 与污染物因子、$PM_{2.5}$ 与气象因子之间的相关性进行分析。其次，介绍了逐步回归的原理和步骤，建立了 $PM_{2.5}$ 质量浓度的逐步回归仿真方程，并对回归方程进行诊断，验证其是否满足假设条

件。

第五章建立了用于预测未来 1~6 小时 $PM_{2.5}$ 质量浓度的逐小时预测模型以及用于预测未来 6~12 小时、12~24 小时、24~48 小时 $PM_{2.5}$ 质量浓度最大值和最小值的极值预测模型。通过 OOB 误差估计进行变量筛选，得到各输入因子的重要性排序，进一步提高模型准确率。最后，给出了各模型在检验数据上的验证结果。

第六章为结论和展望，总结了本文的主要研究成果，并明确了未来的研究方向。

第二章 随机森林研究方法

2.1 随机森林原理与性质

2.1.1 随机森林原理

组合分类方法是一种有助于提高模型分类准确率的技术[1]。其基本思想是：一个组合分类器（ensemble）是由多个个体分类器组合而成，每个个体分类器都有各自的分类结果，组合分类器的分类结果由个体分类器联合决定。对于分类问题，待预测的响应变量是类别变量，组合分类器的结果由多有个体分类器投票决定；对于回归问题，待预测的响应变量是数值变量，组合分类器的结果取个体分类器结果的平均值。

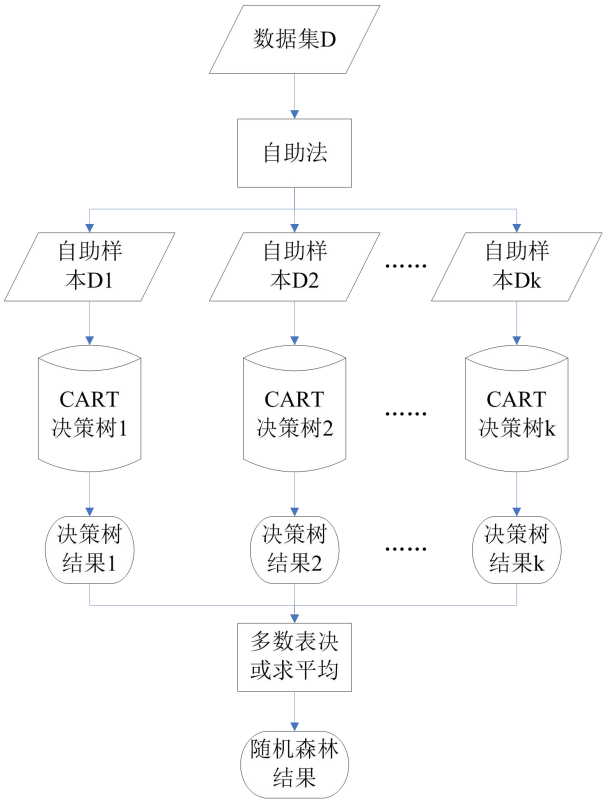


图 2-1 随机森林原理

随机森林算法是 Breiman 在 2001 年提出的，该算法属于组合分类方法的一种[2][13][34]。随机森林中的每个个体分类器都是基于 CART 算法建立的决策树，因此随机森林算法可以解决分类和回归问题，每棵决策树的训练数据通过自助法

（bootstrap）也就是有放回的等概率随机抽样从原数据集中抽取，每个个体分类器拥有各自不同的训练数据，所有个体分类器分类结果中的多数类或者平均值成为随机森林的最终结果。随机森林算法原理详见图 2-1。

随机森林在生成过程中有以下特点：

（1）单棵决策树的训练样本是通过自助法得到的，也称为自助样本。自助法，就是从原数据集中通过有放回的随机抽样的方式获取训练样本的方法，每个样本被抽中的概率相等。由于是有放回，所以数据集中的有些记录可能被多次抽取到训练集中，而未被抽取的记录可以放到检验集中用于估计模型的准确率。

自助法有时也称作.632 自助法，这是因为，对于拥有 d 条记录的数据集 D，有放回地从 D 抽取 d 次，每次抽取时，D 中每条记录被抽中的概率为 $1/d$，因此每条记录未被抽中的概率为 $(1-1/d)$。由于是抽取 d 次，因此，某一条记录最终一次也没被抽中的概率为 $(1-1/d)^d$。当数据集 D 中记录很多，d 的值很大时，$(1-1/d)^d$ 近似等于 $e^{-1}=0.368$。所以，由于未被抽中而放入检验集中的记录数约占数据集 D 记录总数 d 的 36.8%，而被抽中放入训练集的记录数约占数据集 D 记录总数 d 的 63.2%。需要注意的是，这里的 63.2%是指训练集中不同的记录是 d 的 63.2%，而训练集的数据量仍为 d。

由于自助法的特性，会产生 36.8%的检验数据，所以随机森林在生成过程中就可以对模型的准确率进行估计，实验证明，这种估计属于无偏估计。

（2）单棵决策树内部节点分裂时使用的候选属性集并非全属性，而是从所有属性中随机抽取获得。CART 算法在节点分裂过程中，对于使用何种属性作为分裂属性，使用的评价标准是基尼指数。基尼指数是基于每一个属性进行二元划分后的结果子集进行计算的，在内部节点分裂时选择能够产生最小基尼指数的分裂属性。随机森林中的决策树在计算基尼指数时，每一个内部节点的候选属性都是从所有属性中随机抽取的。也就是对于数据集 D 中的 F 个属性，每次只抽取 f 个，并且 f 的值远小于 F。在生成每棵决策树的过程中，f 的值是固定的。

（3）随机森林中的每棵决策树在生长完全后都不进行剪枝处理。传统的 CART 算法在生成决策树后一般需要进行剪枝，其使用的剪枝策略是代价复杂度，

剪枝的目的消除决策树对训练数据的过度拟合。代价复杂度是决策树中叶节点个数与决策树错误率的函数，导致较大代价复杂度的子树会被剪枝，该方法使得模型的结构风险最小化，在保证准确率的基础上维持了模型的简洁。但随机森林中的决策树在生成过程中使其完全生长，并不剪枝，这是因为随机森林算法的两次随机过程（随机抽取训练数据、随机抽取分裂属性）能够很好地避免过度拟合。不剪枝的另一个好处就是消除了决策树的偏移。

（4）随机森林的最终结果是由所有决策树投票表决得到的，每棵决策树拥有相同的投票权重，使用投票作为确定最终结果的策略使得随机森林算法的结果更加稳定。对于离散型的响应变量，使用多数表决的方法，选择所有结果中得票比例最高的作为随机森林分类的最终结果；对于连续型的响应变量，汇总所有决策树的结果，统计其平均值作为随机森林回归的最终结果。

2.1.2 随机森林性质

泛化能力是衡量模型性能的重要指标。泛化能力指的是在未经训练的数据上模型所表现出的性能，模型的泛化能力越强，说明其在未知数据上的预测能力越强。事实上，可以得到一个在训练数据集100%准确率的模型，但一般情况下，这个模型用在新的数据上时准确率会表现得较低，也就是泛化能力较弱。当数据量较大时，模型的泛化能力可以用模型在检验数据上的误差水平表示；在检验数据并不充分的情况下，可以理论推导出随机森林模型的泛化误差概率上界。

假设给定一个随机森林的组合分类器表示为$\{h_1(X),h_2(X),\ldots,h_k(X)\}$，$h(X)$是单个分类器对于输入向量$X$所产生的输出结果。对于通过自助法随机抽取的训练集向量X、Y，Y是训练集中X的对应分类结果，定义间隔函数（margin function）见公式（2-1）：

$$mg(X,Y)=av_k I(h_k(X)=Y)-\max_{j\neq Y} av_k I(h_k(X)=j) \qquad (2\text{-}1)$$

式（2-1）中，$I(\ )$为指示函数。间隔函数衡量了平均正确分类数超过平均错误分类数的程度。间隔函数越大，说明模型分类的置信水平越高。

因此，随机森林的泛化误差可以定义为公式（2-2）：

$$PE^{*} = P_{X,Y}(mg(X,Y) < 0) \qquad (2\text{-}2)$$

对于随机森林模型，$h_k(X) = h(X, \Theta_k)$，Θ 是单棵决策树独立同分布的参数向量。当随机森林的规模很大时，其性能遵循大数定律规律，即随着树的棵数的增加，几乎可以肯定，所有序列$\Theta_1, \ldots, PE^{*}$收敛于公式（2-3）：

$$P_{X,Y}(P_{\Theta}(h(X,\Theta) = Y) - \max_{j \neq Y} P_{\Theta}(h(X,\Theta) = j) < 0) \qquad (2\text{-}3)$$

这就说明，随着随机森林中树的棵数的增加，模型不会陷入过度拟合[31]，其泛化误差有一个理论上的极限值。

另外，随机森林算法还具有以下优点[2]：

（1）随机森林算法的性能可以和 Adaboost 算法媲美，有时甚至表现得更好；

（2）随机森林对噪声点和离群值更加鲁棒；

（3）由于内部节点分裂时只考虑很少的属性，随机森林相较于装袋（bagging）或提升（boosting）计算更快；

（4）随机森林可以在不需要额外数据的情况下给出错误率、相关性、变量重要性的估计；

（5）随机森林算法易于实现且有利于并行化计算。

2.2 基于 OOB 误差估计的变量选择方法

随机森林可以返回多种衡量变量重要性的度量，其中最可靠的是基于树节点分裂属性被随机替换后的分类准确度降低值[19]。为了筛选变量，需要迭代拟合随机森林，在每次迭代过程中生成一个新的森林，这些森林是由剔除最不重要变量后的变量集合生成的，最终选择的是产生最小 OOB 误差的变量集合。

OOB 误差通过 OOB 估计（Out-of-Bag Estimation）获得[25]。对于随机森林中的每个个体分类器来说，每次选择用于学习的训练数据都是通过自助法从原数据集中有放回地等概率抽取的自助样本，原数据集中约 37%的数据在自主样本之外，这部分数据被称为袋外（Out-of-Bag）数据。利用这些数据可以估计每个个体分类器的各种统计量，如泛化误差、强度、相关性等，随机森林的 OOB 误差是每个个体分类器的 OOB 误差估计值平均后得到的[8]。

基于 OOB 误差估计的变量选择，其基本原理是：在每次迭代的过程中，通过预测变量 X_j 的随机置换，该预测变量与响应变量 Y 的关联被打破[7]。当被置换后的预测变量 X_j 连同其他未被置换的预测变量被用于预测袋外数据的响应变量时，预测准确度会发生变化，若准确度发生较大幅度的下降，说明置换前的预测变量 X_j 与响应变量有关，该变量的重要性较大。

因此，可以利用预测变量 X_j 随机置换前后准确度变化的差值，所有决策树平均后作为衡量变量重要性的度量，计算过程如下：

首先，计算预测变量 X_j 在第 t 棵树中的重要性，计算见公式（2-4）：

$$VI^{(t)}(X_j)=\frac{\sum_{i\in\overline{B}^{(t)}} I\left(y_i=y_i^{(t)}\right)}{\left|\overline{B}^{(t)}\right|}-\frac{\sum_{i\in\overline{B}^{(t)}} I\left(y_i=y_{i,\pi_j}^{(t)}\right)}{\left|\overline{B}^{(t)}\right|} \tag{2-4}$$

式（2-4）中，$\overline{B}^{(t)}$ 是第 t 棵树的 OOB 数据，$\hat{y}_i^{(t)}=f^{(t)}(x_i)$ 是预测变量 X_j 被替换前第 i 条记录的预测结果，$\hat{y}_{i,\pi_i}{}^{(t)}=f^{(t)}(x_{i,\pi_i})$ 是预测变量 X_j 被替换后第 i 条记录的预测结果。由于节点的候选分裂属性是随机抽取的，因此会出现第 t 棵树中没有预测变量 X_j 的情况，此时定义 $VI^{(t)}(X_j)=0$。

变量 X_j 的重要性计算结果是取所有树中该变量重要性的平均值，计算见公式（2-5）：

$$VI\left(X_j\right)=\frac{\sum_{t=1}^{ntree} VI^{(t)}\left(X_j\right)}{ntree} \tag{2-5}$$

最终选择变量数目最少且导致最小误差率的变量集合。

2.3 随机森林算法步骤

2.3.1 随机森林单棵决策树生成算法

随机森林中的决策树是通过 CART 算法生成的，该算法在每个节点只考虑二元划分，因此最终产生的决策树是一棵二叉树。但在生成随机森林的具体过程中，训练集和候选分裂属性集的获取与传统的 CART 算法略有不同。

根据响应变量的不同，决策树的类型可分为分类树和回归树[36]。针对离散型的响应变量，决策树也称为分类树，分裂规则使用的是基尼指数；针对连续型的

响应变量，决策树可称为回归树，使用的分裂规则是平方误差。

基尼指数是衡量分区内数据不纯度的指标，计算公式见（2-6）：

$$Gini(D)=1-\sum_{i=1}^{m}p_i^{\,2} \tag{2-6}$$

式（2-6）中，D 是数据分区，m 是数据分区中类的个数，$p_i=\left|C_{i,D}\right|/\left|D\right|$ 是数据分区 D 中记录属于 C_i 类的概率，$\left|D\right|$ 是数据分区 D 中记录的总条数，$\left|C_{i,D}\right|$ 是数据分区 D 中属于 C_i 类的记录的条数。

对于属性 A，当 A 中有 v 个不同的值时，在不考虑全集和空集的情况下，有 $\left(2^v-2\right)/2$ 种不同的方法可以将数据集 D 划分成两个分区 D_1 和 D_2。对于特定的二元划分，计算基于属性 A 的二元分裂子集的基尼指数加权和，也就是该划分的基尼指数，计算见公式（2-7）：

$$Gini_A(D)=\frac{\left|D_1\right|}{\left|D\right|}Gini(D_1)+\frac{\left|D_2\right|}{\left|D\right|}Gini(D_2) \tag{2-7}$$

对于每个属性，分别计算其每一种可能的划分结果所得到的基尼指数，选择划分后分区内纯度最大，也就是分裂子集各自基尼指数加权后总和最小的子集划分方式作为该属性的分裂策略。

选择基于属性 A 的划分，也将导致数据集整体不纯度有所降低，其降低值的大小计算见公式（2-8）：

$$\Delta Gini(A)=Gini(D)-Gini_A(D) \tag{2-8}$$

在数据集的基尼指数固定的情况下，按属性划分后的基尼指数越小，总体基尼指数的下降程度就越大，分裂后的集合中纯度就越高，因此，为了最大化基尼指数的下降程度，提升数据分区内的纯度，选择基尼指数最小的属性作为该节点的分裂属性。

分裂属性和分裂子集一起构成了 CART 分类树算法的分裂准则。

基于基尼指数的随机森林 CART 分类树算法见图 2-2。

算法：随机森林 CART 分类树算法

输入：

• 数据集 D

• 候选属性集合 L_A

• 抽取属性个数 f

输出：一棵随机森林的 CART 分类树

方法：

（1）自助法从数据集 D 中抽取训练数据 D_k；

（2）创建节点 N；

（3）**if** D_k 中的记录都同属于一个类 C **then**

（4） 返回节点 N，标记 N 的类为 C；

（5）**for** 内部节点

（6） 从 L_A 中抽取 f 个属性作为候选分裂属性集合 L_{Ai}；

（7） 对于 L_{Ai} 中每个属性的每种可能的划分计算其基尼指数，确定二元划分；

（8） 将 D_k 划分为两部分 D_{k1} 和 D_{k2}；

（9） **if** D_{kj} 满足停止条件 **then**

（10） 添加叶节点到 N，其类别为 D_{kj} 中的多数类；

（11） **else** 添加内部节点到 N；

（12）**end for**

图 2-2 CART 分类树算法

回归树中衡量一个数据分区纯度的指标是平方误差，假设将数据集 D 划分为 m 个子集 $D_1, D_2, \ldots, D_m$，子集 D_m 中的最优输出值见公式（2-9）：

$$\hat{c}_m = ave\left(y_i \middle| x_i \in D_m\right) \tag{2-9}$$

式（2-9）中，$ave(\ )$ 计算了子集 D_m 内所有输入向量 x_i 对应输出值 y_i 的平均值。

对于属性 B 进行划分，当 B 中有 v 个不同的连续值时，将 B 中的数值按升序或降序排序后，每次将该属性可能的分裂点 s 选在两个相邻值之间，也就是中点

处，这样对于属性 B 就有 $v-1$ 种不同的方法可以将数据集 D 划分成两个子集 D_1 和 D_2，子集的定义见公式（2-10）：

$$D_1(B,s)=\{x|x^{(B)}\le s\} \text{ 和 } D_2(B,s)=\{x|x^{(B)}>s\} \tag{2-10}$$

子集 D_1 和 D_2 划分后，子集内的平方误差见公式（2-11）：

$$SE_1=\sum_{x_i\in D_1(B,s)}(y_i-c_1)^2 \text{ 和 } SE_2=\sum_{x_i\in D_2(B,s)}(y_i-c_2)^2 \tag{2-11}$$

最优分裂属性 B 和最优分裂点 s 应该满足公式（2-12）：

$$\min_{B,s}\left(\min_{c_1}(SE_1)+\min_{c_2}(SE_2)\right) \tag{2-12}$$

在确定了节点处的分裂属性 B 和分裂点 s 后，划分后的子集 D_1 和子集 D_2 的相应输出值计算见公式（2-13）：

$$\hat{c}_1=ave(y_i|x_i\in D_1(B,s)) \text{ 和 } \hat{c}_2=ave(y_i|x_i\in D_2(B,s)) \tag{2-13}$$

对于每个属性的每种可能的划分计算其平方误差，选择使得划分后集合的纯度最大，也就是拥有最小平方误差的划分点作为该属性的分裂点，对应属性作为节点的分裂属性。

分裂属性和分裂点一起构成了 CART 回归树算法的分裂准则。

基于平方误差的随机森林 CART 回归树算法见图 2-3。

算法：随机森林 CART 回归树算法

输入：

• 数据集 D

• 候选属性集合 L_A

• 抽取属性个数 f

输出：一棵随机森林的 CART 回归树

方法：

（1）自助法从数据集 D 中抽取训练数据 D_k；

（2）创建节点 N；

（3）**if** L_A 为空 **then**

（4） 返回节点 N，标记 N 的值为 D_k 中所有输出值的均值；

（5）**for** 内部节点

（6） 从 L_A 中抽取 f 个属性作为候选分裂属性集合 L_{Ai}；

（7） 对于 L_{Ai} 中每个属性的每种可能的分裂点算其平方误差，确定二元划分；

（8） 将 D_k 划分为两部分 D_{k1} 和 D_{k2}；

（9） **if** D_{kj} 满足停止条件 **then**

（10） 添加叶节点到 N，其值为 D_{kj} 中所有输出值的均值；

（11） **else** 添加内部节点到 N；

（12）**end for**

图 2-3 CART 回归树算法

2.3.2 随机森林组合算法

随机森林是多个决策树的组合，每棵决策树的输出结果不尽相同，使用不同的组合策略得到的结果也是不同的，随机森林使用的组合策略是，让模型中每棵决策树输出的结果都具有相同的权重。也就是，对于结果为离散型的输出，随机森林取占多数的分类结果；对于结果为连续型的输出，随机森林取所有输出结果的平均值。

随机森林算法之所以能够显著提高个体决策树的准确率，是因为复合模型降低了个体模型的方差。

随机森林的组合算法汇总见图 2-4。

算法：随机森林组合算法

输入：

- 数据集 D
- 随机森林中的模型数 k

输出：随机森林模型

方法：

（1）**for** i = 1 to k **do**

（2）自助法从数据集 D 中抽取训练数据 D_k；

（3）使用 D_k 和随机森林 CART 算法生成模型 M_k；

（4）**end for**

（5）离散数据多数表决，连续数据取均值；

图 2-4 随机森林组合算法

第三章 研究区域与数据

3.1 研究区域概况

上海市是我国四大直辖市之一，同时也是我国的经济、贸易、金融、科技、工业、交通、航运和会展中心。

徐汇区位于上海市中心城区的西南部，是上海市西南部的城市副中心，地理位置优越，北与黄浦区、静安区、长宁区接壤；东南与浦东新区隔江相望；西南与闵行区分界，详见图 3-1。全境总面积 54.76 平方千米，区内总人口 108.91 万人，人口密度位列全市各区县第 7 位（2016 年上海市统计年鉴）。

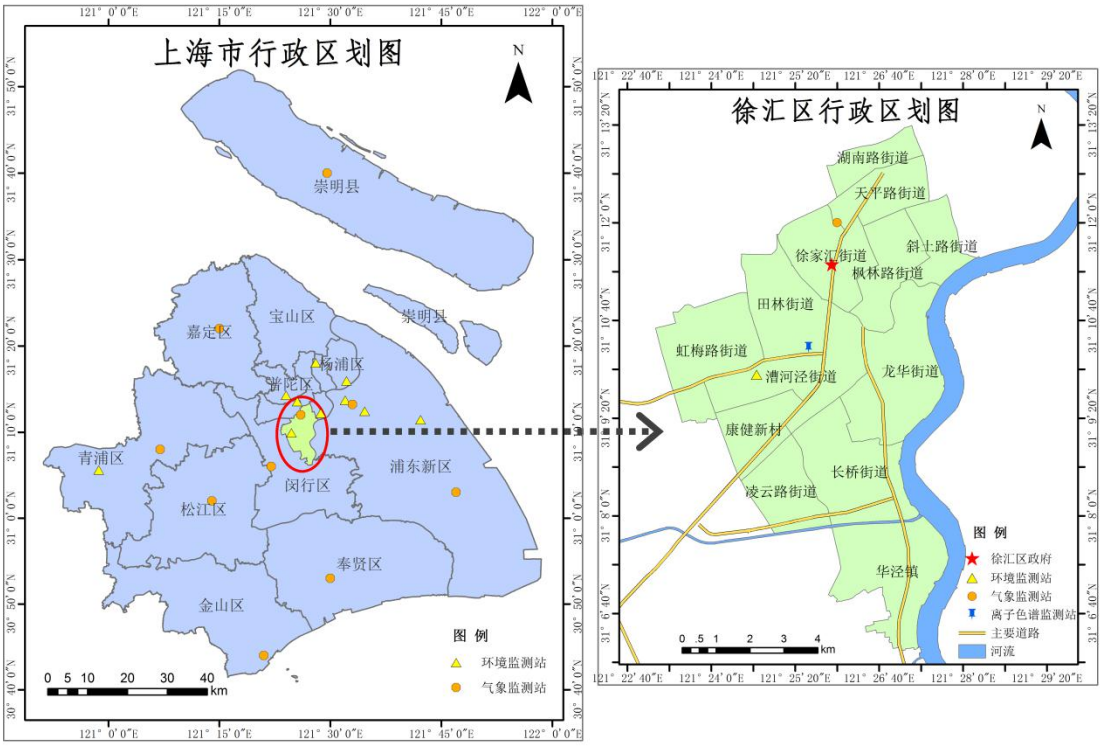


图 3-1 研究区域

上海市共有气象数据国家控制点 10 处，分别位于：嘉定、浦东、徐家汇、崇明、惠南、金山、奉贤、小洋山、闵行、青浦、松江，从图 3-1 中可以看出，气象监测站点分布较为均衡，基本覆盖了上海市全境，其中，位于徐汇区的有一处。另外，上海市环境监测站点也有 10 处，分别位于：静安、虹口、浦东新区、青浦淀山湖、黄浦区十五厂、普陀、徐汇上师大、杨浦四漂、浦东川沙、浦东张

江。与气象监测站点不同的是，环境监测站点分布较为集中，主要在中心城区，其中，位于徐汇区的有一处。另外，离子色谱监测站点也位于徐汇区，因此，本文将徐汇区作为研究区域。

3.2 数据来源

本文所使用到的数据均来源于上海市环境科学研究院专业系统平台，选取的监测时段为 2016 年 3 月至 2016 年 10 月，数据的采样频率为每隔一小时获得一组实测数据并上传服务器。本文使用到的原始数据共包含 3 个部分，分别是：

（1）环境监测数据，主要为各污染物的质量浓度，共 4986 条，单位为μg/m^3。本文主要使用徐汇上师大监测站数据，数据解释及编码见表 3-1。

表 3-1 污染物质量浓度输入

因子	解释	编码
$PM_{2.5}$	颗粒物，粒径小于 2.5μm	101
PM_{10}	颗粒物，粒径小于 10μm	102
O_3-1	臭氧，1 小时平均	103
O_3-8	臭氧，8 小时滑动平均	104
CO	一氧化碳	105
SO_2	二氧化硫	106
NO_2	二氧化氮	107

（2）MARGA 在线离子色谱数据，共 5119 条，单位为μg/m^3。该数据监测站点位于上海市徐汇区，数据解释及编码见表 3-2。

表 3-2 离子质量浓度输入

因子	解释	编码
NH_4^+	铵根离子	201
NO_3^-	硝酸根离子	202
SO_4^{2-}	硫酸根离子	203
K^+	钾离子	204
Ca^{2+}	钙离子	205
Na^+	钠离子	206
Mg^{2+}	镁离子	207
Cl^-	氯离子	208

（3）气象实况数据，主要是气象部门共享的自动站监测数据，共 4825 条，本文主要使用徐家汇监测站数据，数据解释及编码见表 3-3。

表 3-3 气象因子输入

因子	解释	编码
PRS	气压	301
TEM	气温	302
RHU	相对湿度	303
PRE_1H	降水量，1 小时平均	304
WIN_D_AVG_2MI	风向，2 分钟平均	305
WIN_S_AVG_2MI	风速，2 分钟平均	306
WIN_D_AVG_10MI	风向，10 分钟平均	307
WIN_S_AVG_10MI	风速，10 分钟平均	308
VIS_HOR_1MI	能见度，1 分钟平均	309
VIS_HOR_10MI	能见度，10 分钟平均	310

3.3 数据预处理

为了方便之后的数据处理，先以时间为关键字，将颗粒物、离子、气象三张表的数据整合成一张表，该表反映的是当前时间下的 AQI 指标污染物的质量浓度情况、离子质量浓度情况和各类气象条件情况。由于传输处理等方面原因，时间字段可能会有几分钟的偏差，因此，时间列只精确到小时。

3.3.1 缺失值处理

由于存在机器故障、质控检修、传输不稳定等客观原因，数据在在线接收时难免会存在一定数量的缺失值，缺失值的存在将导致一些分析方法无法应用，因此，为了提高数据质量，需要对缺失值进行处理。

处理缺失值的方法一般包括：

（1）将含有缺失值的记录删除。该方法最为简单，同时可以保证数据的真实性，当缺失记录在数据集中所占的比例非常小时这种处理方法较为合理。该方法的另一个策略是，不把所有包含缺失值的记录都删除，只是删除某些缺失值个数较多的记录。例如，如果某条记录的缺失属性个数超过总属性数的 20%，可以考虑删除将该条记录。

（2）人工补全缺失值。某些情况下，可以通过人工查找日志等方式填补缺失值。但在数据量较大的情况下，该方法费时费力，一般不采用。

（3）使用全局常量填补缺失值。将所有缺失值填补为一个全局常量，如

“MISSING”，在分类模型中可以当成一个属性值使用，尽管该方法一定程度上可以反映数据集的规律，但并不十分可靠。

（4）使用缺失属性的中心度量填补缺失值。一般使用反映中心趋势的值填补，对称分布的数据使用均值填补，倾斜分布的数据使用中位数填补。例如，使用 CO 质量浓度的均值填补缺失的 CO 数据。

（5）使用同类样本的中心度量填补缺失值。与直接使用缺失属性即含有缺失值的列的中心度量不同的是，该方法考虑了同类数据之间的相关性，也就是通过一个第三方属性对参与计算中心度量的记录进行了约束。例如，按照空气质量等级对数据集分类，使用空气质量等级相同情况下的 CO 质量浓度的均值或中位数填补其缺失值。

（6）推测缺失部分最可能的值。一般使用数据挖掘的方法，如决策树、贝叶斯、神经网络等，通过属性间的关系，推测归纳填补缺失值。该方法通过使用大部分已有的数据信息预测小部分缺失值，更大几率保留了各属性之间的内在相互联系。

方法（1）和方法（2）属于无偏处理，方法（3）~（6）会使数据有偏，因为填补的数据可能不正确。

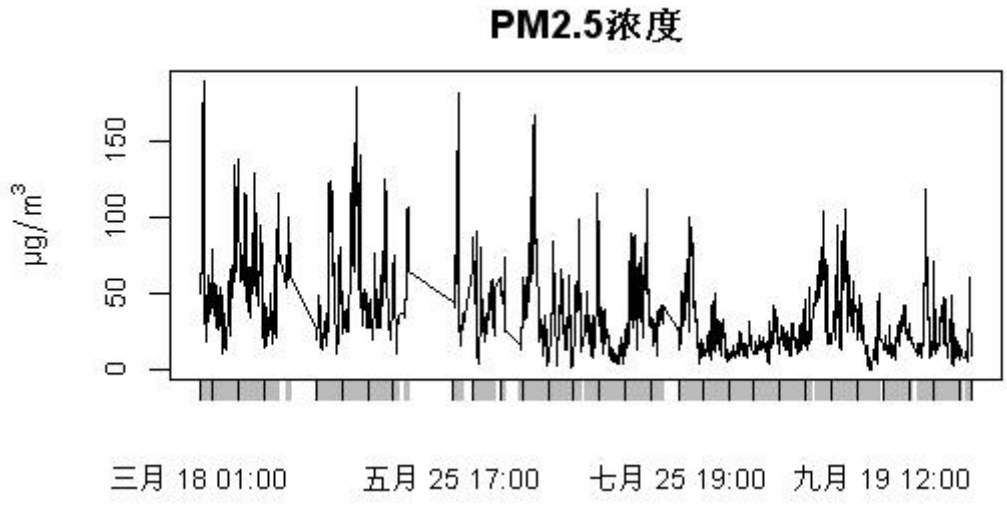


图 3-2 $PM_{2.5}$ 质量浓度变化

本文预测模型的响应变量即预测目标是 $PM_{2.5}$ 的质量浓度值，采集到的原始样本中其随时间的变化情况见图 3-2，图中横坐标灰色线条部分标记了数据采集

的时间，空白部分为缺失数据，可以看出，$PM_{2.5}$质量浓度的缺失主要集中在某几个时间段，并非随机缺失。因此，在缺少足够依据的情况下盲目填补将造成数据倾斜。为了使$PM_{2.5}$的质量浓度数据无偏且准确，使用方法（1）处理该部分缺失值，即剔除所有$PM_{2.5}$质量浓度缺失的记录。同时，由于 MARGA 设备原因，在某些时间段，所有离子色谱数据均无法接收，因此，删除离子色谱数据全为空的记录。处理后，剩余数据 3911 条，各因子缺失值情况见表 3-4。

表 3-4 各因子缺失值数量

因子编码	缺失数量	因子编码	缺失数量	因子编码	缺失数量
101	0	201	0	301	1
102	862	202	0	302	1
103	2	203	0	303	1
104	0	204	0	304	2
105	0	205	0	305	1458
106	0	206	0	306	1
107	8	207	0	307	1200
		208	0	308	1
				309	1
				310	1

从表 3-4 中可以看出，经过处理后的数据，缺失值主要集中在 102、305、307 这三个属性上，离子质量浓度（因子编码 201~208）不再含有缺失值，说明离子质量浓度因子的数值缺失很可能是由于机器设备检修或故障等原因造成的整体缺失，而并非随机产生的缺失值。处理后的数据在全部 25 个属性中，缺失值主要集中在其中 3 个属性上，缺失属性占的比例较小。为了尽可能地使填补的缺失值准确，考虑使用方法（6）再次对缺失值进行处理。

本文使用 KNN（k-最近邻分类，k-nearest-neighbor-classifier）方法填补缺失值。该方法试图寻找与含有缺失值的记录最相似的 k 条记录，并用这 k 条记录的中心度量来填补缺失值。在数据集中，每一条记录由 n 个属性描述，因此，每一条记录相当于 n 维空间中的一个点，这个点可以通过特定的距离公式找到与自己距离最接近的 k 个点，这 k 个点就是该点的“最近邻”。

本文衡量记录之间相似性使用的是欧几里得距离，这个距离可以定义任意两条记录之间属性值差值的平方和，两条记录$X=(x_1,x_2,\ldots,x_n)$和$Y=(y_1,y_2,\ldots,y_n)$

的欧几里得距离计算见公式（3-1）：

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}\delta_i(x_i,y_i)} \tag{3-1}$$

其中，$\delta_i(\ )$表示对应的两个属性i之间的距离，计算见公式（3-2）：

$$\delta_i(v_1,v_2)=\begin{cases}1,\text{当}i\text{是标称属性且}v_1=v_2\text{时}\\0,\text{当}i\text{是标称属性且}v_1\neq v_2\text{时}\\(v_1-v_2)^2,\text{当}i\text{是数值属性时}\end{cases} \tag{3-2}$$

为了避免值域范围较大的属性对值域范围较小的属性权重的影响，例如，$PM_{2.5}$质量浓度分别为188μg/m^{3}和189μg/m^{3}的记录明显比CO质量浓度为1.0μg/m^{3}和0.5μg/m^{3}的记录距离要接近得多，但前者的欧几里得距离大于后者，因此，在计算距离之前，需要对数值属性进行标准化，计算见公式（3-3）：

$$y_i=\frac{x_i-\bar{x}}{\sigma_x} \tag{3-3}$$

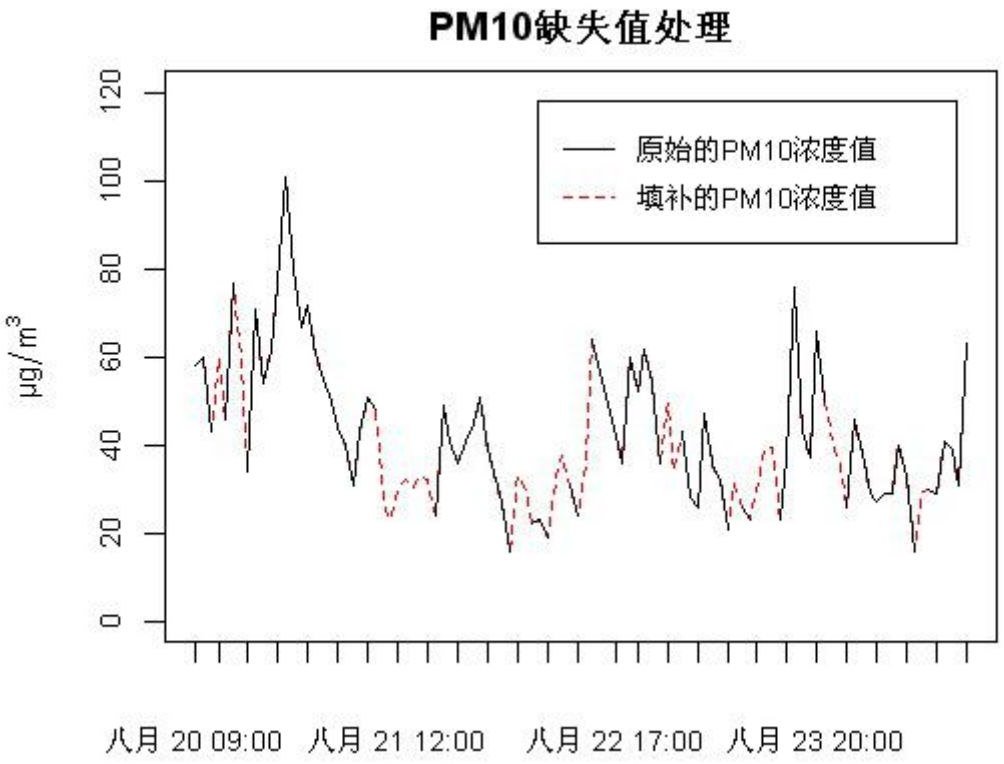


图3-3 PM_{10}缺失值处理对比

本文选择k的个数为10，并通过10条最近邻记录的中位数来填补缺失值。处理完成后，数据集中不再含有缺失值，图3-3反映了部分PM_{10}质量浓度的填

补效果。图 3-3 中黑色实线为收集到的 PM_{10} 质量浓度值，红色虚线为填补的缺失值，从图 3-3 中可以看出，填补后的 PM_{10} 质量浓度曲线较大程度上保持了原有的波动规律。

3.3.2 输入因子的改进

为了得到更好的预测效果，在引入化学因子（包括污染物和离子质量浓度）和气象因子的同时，再在数据集中加入周期因子和预报因子。

周期因子主要包括 $PM_{2.5}$ 质量浓度的月平均数据和周平均数据，其反映了 $PM_{2.5}$ 的周期变化规律。在取得均值后，还需要进行标准化处理，并假设未来 n 年污染物质量浓度的周期变化基本符合该趋势。最终，月度周期质量浓度取值如下：三月为 1.31（$PM_{2.5}$ 质量浓度为 58μg/m^3）、四月为 1.18（$PM_{2.5}$ 质量浓度为 56μg/m^3）、五月为 0.71（$PM_{2.5}$ 质量浓度为 49μg/m^3）、六月为 0.18（$PM_{2.5}$ 质量浓度为 41μg/m^3）、七月为-0.61（$PM_{2.5}$ 质量浓度为 29μg/m^3）、八月为-1.34（$PM_{2.5}$ 质量浓度为 18μg/m^3）、九月为-0.41（$PM_{2.5}$ 质量浓度为 32μg/m^3）、十月为-1.01（$PM_{2.5}$ 质量浓度为 23μg/m^3）；周度周期质量浓度取值如下：周一为-1.02（$PM_{2.5}$ 质量浓度为 32μg/m^3）、周二为-0.62（$PM_{2.5}$ 质量浓度为 34μg/m^3）、周三为-0.62（$PM_{2.5}$ 质量浓度为 34μg/m^3）、周四为 0.96（$PM_{2.5}$ 质量浓度为 42μg/m^3）、周五为 1.75（$PM_{2.5}$ 质量浓度为 46μg/m^3）、周六为-0.42（$PM_{2.5}$ 质量浓度为 35μg/m^3）、周日为-0.03（$PM_{2.5}$ 质量浓度为 37μg/m^3）。数据解释及编码见表 3-5。

表 3-5 周期因子输入

因子	解释	编码
MONTH	$PM_{2.5}$ 月平均质量浓度	401
WEEK	$PM_{2.5}$ 周平均质量浓度	402

预报因子主要指的是未来一段时间的气象预报，包括预测时段的气压、气温、相对湿度、降水量、风向和风速预报，但不包含能见度，一方面是因为一般不提供能见度方面的预报数据，另一方面，本文中的能见度为实测数据，而能见度和污染物的质量浓度有明显的相关关系，作为预报数据加入后会影响预测结果的合理性，因此不予考虑。

本文将要预测的是 $PM_{2.5}$ 污染物 1~6 小时的逐小时质量浓度以及 6~12 小时、

12~24 小时、24~48 小时的质量浓度极值，因此，针对不同的预测目标，所选择的预报因子也不尽相同。其中，针对 1~6 小时的逐小时预测，预报因子共 6 组，分别为未来 1~6 小时每小时的气压、气温、相对湿度、降水量、风向、风速的预报数值；针对极值预测，预报因子共 6 组，分别为未来 6~12 小时、12~24 小时、24~48 小时的各时间区间下气象因子预报数值的最大值和最小值。

考虑到本文数据在时间序列上有空白区间，因此，将预报数值为空的记录删除，最终得到数据集记录 3820 条。

3.4 上海市 $PM_{2.5}$ 变化特征

3.4.1 月际变化分析

2016 年 3 月至 10 月上海市 $PM_{2.5}$ 月平均质量浓度的分布呈现出一定的季节性规律，如图 3-4 所示，3 月至 10 月 $PM_{2.5}$ 月平均浓度在时间尺度上呈现整体下降趋势，从 6 月份开始，空气质量由良转为优（空气质量指数等级及对应颜色为：优-绿色、良-黄色、轻度污染-橙色、中度污染-红色、重度污染-紫色、严重污染-褐红色），8 月份空气污染程度处于全年最低水平。原因是，夏季高温多雨的气候特点，可以有效抑制雾霾天气的发生，同时这个时间段也是雷雨大风等强对流天气的高发时段，有助于雾霾的消散。从图中可以看出，$PM_{2.5}$ 各月的平均浓度之间差距较大，最大差别在 3 倍以上。

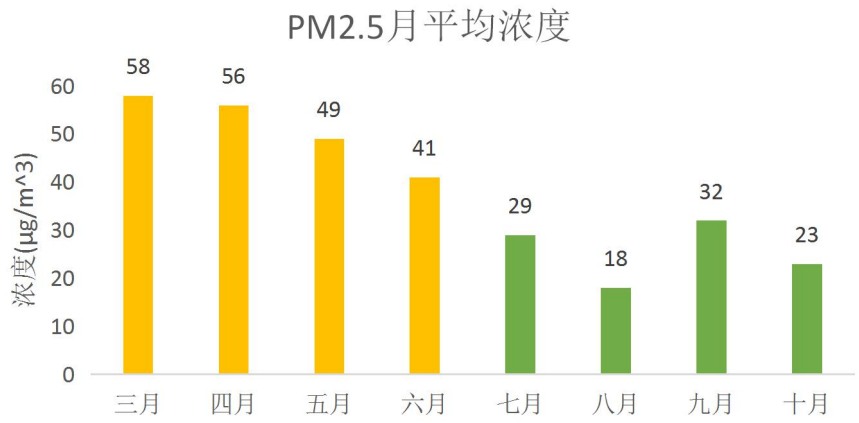


图 3-4 $PM_{2.5}$ 月平均浓度变化

图 3-5 反映了各月各空气质量等级出现时间的占比，从图 3-5 中可以看出，

中度以上污染天气主要集中在 3 月和 4 月，原因是上海冬季温和少雨的特点，强对流天气较少，不利于雾霾消散，气候因素对空气质量有一定影响。但 7 月之后再也没有出现过中度以上污染的情况，8 月几乎全月处于空气质量为优的环境下。总体上看，上海市空气质量为优或良的时间占据绝大多数，整体空气质量表现较为良好。

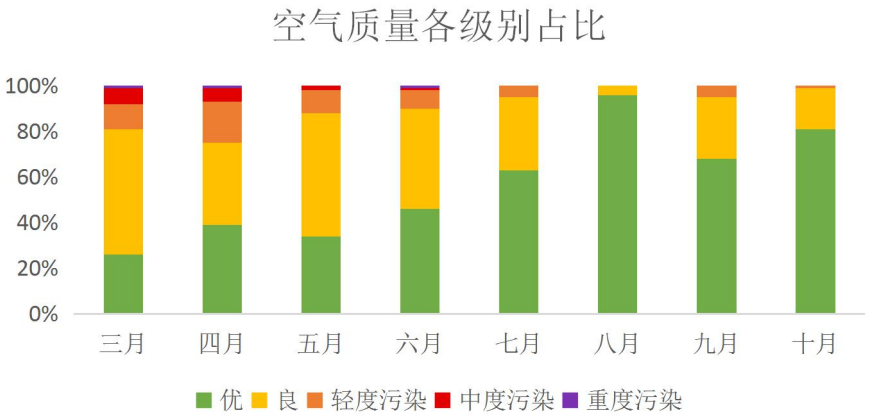


图 3-5 空气质量等级月变化

3.4.2 周际变化分析

2016 年上海市 $PM_{2.5}$ 周平均质量浓度分布变化如图 3-6，周四、周五的 $PM_{2.5}$ 质量浓度明显高于其他日期，从周一到周日，$PM_{2.5}$ 质量浓度变化主要表现为先升高后降低。原因是，周四、周五中心城区内车流量较大，尤其是在早高峰和晚高峰时段，而在周末，一方面驾车出行的总体数量减少，另一方面轿车出行多前往郊区公园等周边地区，中心城区 $PM_{2.5}$ 质量浓度因此有所下降。

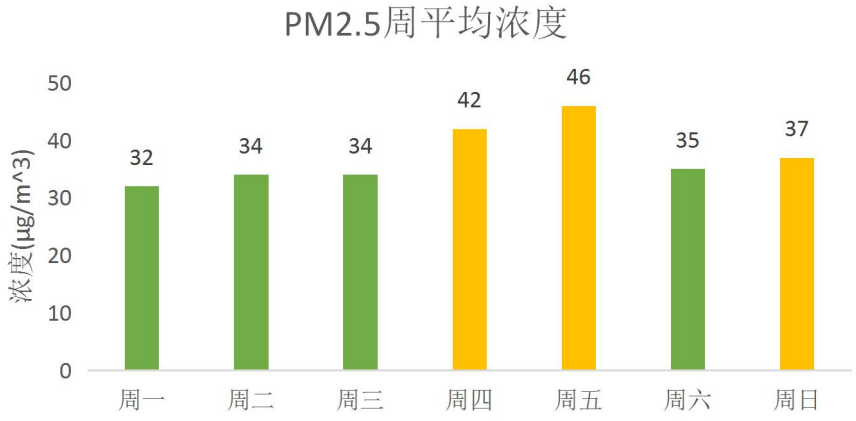


图 3-6 $PM_{2.5}$ 周平均浓度变化

第四章 相关分析与仿真

4.1 相关分析

在建立 $PM_{2.5}$ 质量浓度预测模型时，直接对数据进行数值运算不免存在一定的主观性和盲目性，所以在此之前还需要对属性之间的联系进行定量分析。在本节，将分别定量地分析 $PM_{2.5}$ 与 AQI 其他指标污染物以及 $PM_{2.5}$ 与各气象因子之间的相关性。

4.1.1 相关分析原理

相关性主要描述的是两个属性之间的一种潜在关系，这种关系衡量了其中一个属性对另一个属性的蕴含程度。对于标称属性，通常使用 χ^2（卡方）检验衡量其相关性，而对于数值属性，常用到的方法是相关系数和协方差。具体到相关系数，按照适用的数据类型又细分为 Pearson 相关系数、Spearman 相关系数等。其中，Pearson 相关系数衡量了两个连续型数值属性之间定量的线性相关程度，Spearman 相关系数主要描述了分级或有序属性之间的相关程度。本文选择 Pearson 相关系数作为评价标准对 $PM_{2.5}$ 与其他因子之间的相关性进行分析，计算公式见（4-1）：

$$r_{A,B}=\frac{\sum_{i=1}^{n}\left(a_i-\overline{A}\right)\left(b_i-\overline{B}\right)}{n\sigma_A\sigma_B}=\frac{\sum_{i=1}^{n}(ab)-n\overline{A}\,\overline{B}}{n\sigma_A\sigma_B} \tag{4-1}$$

其中，n 表示数据的长度即数据量，a_i 与 b_i 表示属性 A 与属性 B 的第 i 个值，$\overline{A}$ 与 $\overline{B}$ 表示属性 A 与属性 B 的均值，σ_A 与 σ_B 表示属性 A 与属性 B 的标准差。

相关系数的绝对值在[0,1]之间，绝对值为 0 表示两个属性完全不相关，绝对值为 1 表示两个属性完全相关，系数值越大说明相关性越强。相关系数的符号指明的是两个属性之间相关关系的方向：“+”表示正相关，即一个属性值随着另一个属性值的增加而增加，“-”表示负相关，即一个属性值随着另一个属性值的增加而减少。

需要注意的是，相关关系和因果关系是不同的，A 与 B 相关性较强并不等同于 A 的发生会造成 B 的发生或者 B 的发生会造成 A 的发生。

4.1.2 $PM_{2.5}$ 与其他污染物之间的相关分析

污染颗粒物 $PM_{2.5}$ 与其他污染物之间的相关分析主要是指作为空气质量指数参评污染物的 $PM_{2.5}$ 与 PM_{10}、CO、NO_2、SO_2、O_3 之间的相关关系。首先，对 6 中污染物作散点图，如图 4-1 所示，任意两个污染物的散点图可以在行列交叉处找到，主对角线画出了各污染物的核密度曲线和轴须图，散点图中的绿色实线表示了两种污染物质量浓度的线性拟合曲线，红色虚线表示了两种污染物质量浓度值之间平滑后的拟合曲线。

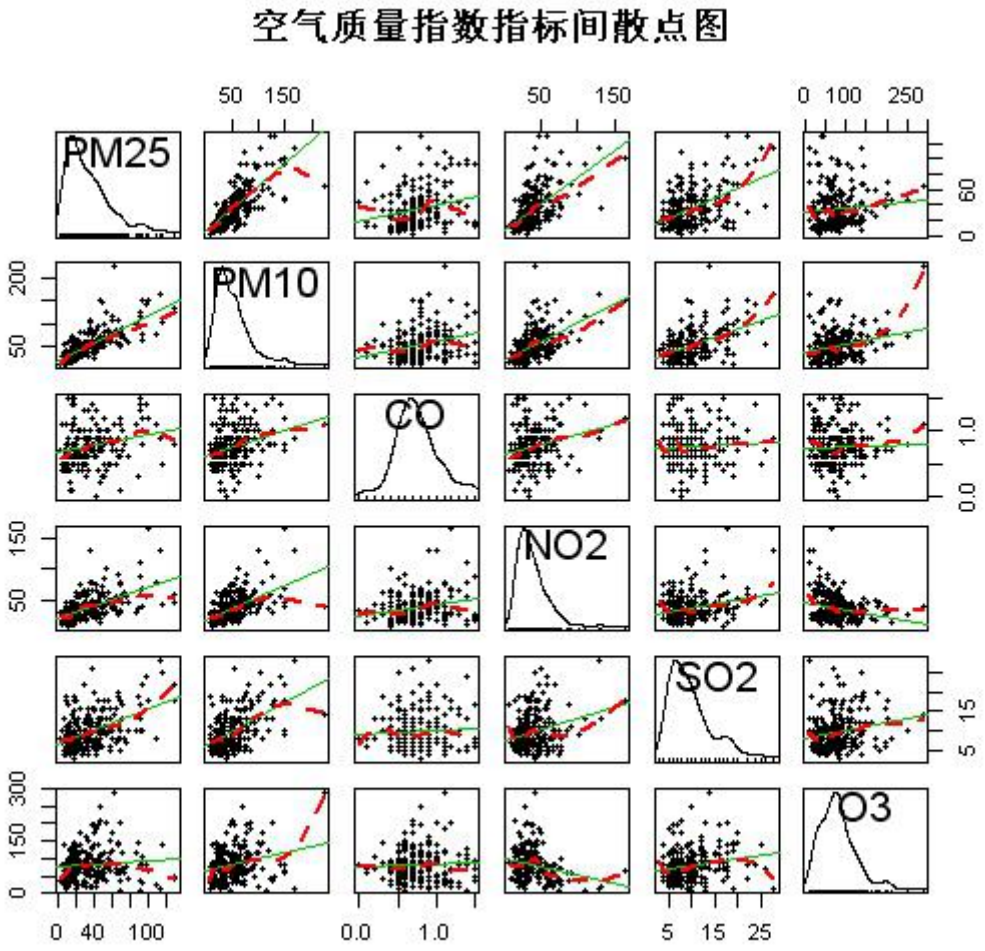


图 4-1 空气质量指数指标间散点图

从图中可以得出以下结论：

（1）从相关性的强弱上看，$PM_{2.5}$与PM_{10}之间的相关性较强，两者间有比较明显的线性相关趋势，$PM_{2.5}$与NO_2、SO_2之间的相关性较弱。此外，$PM_{2.5}$与CO、O_3之间基本没有线性相关关系。

（2）从相关性的方向上看，$PM_{2.5}$与PM_{10}、CO、NO_2、SO_2、O_3之间均呈现出正相关的趋势，但$PM_{2.5}$与O_3之间的正相关性并不明显；PM_{10}与其他污染物之间也存在类似趋势。

（3）从污染物浓度的核密度曲线看，$PM_{2.5}$、PM_{10}、NO_2、SO_2、O_3的质量浓度主要集中在较低水平，发生较高浓度污染的概率不大，CO 的质量浓度较大概率出现在 0.5μg/m^3-1μg/m^3 的中间值区段。

为了进一步定量地分析各污染物之间的相关关系，计算获得各污染物之间的 Pearson 相关系数矩阵，见表 4-1。

表 4-1 各污染物相关系数矩阵

	$PM_{2.5}$	PM_{10}	CO	NO_2	SO_2	O_3
$PM_{2.5}$	1.00	0.74	0.39	0.61	0.50	0.09
PM_{10}	0.74	1.00	0.28	0.49	0.45	0.16
CO	0.39	0.28	1.00	0.34	0.15	-0.05
NO_2	0.61	0.49	0.34	1.00	0.35	-0.32
SO_2	0.50	0.45	0.15	0.35	1.00	0.13
O_3	0.09	0.16	-0.05	-0.32	0.13	1.00

同时，根据表所示相关系数矩阵绘制污染物指标间的相关系数图，见图 4-2。图 4-2 中下三角部分，红底正斜线方块表示两个指标间呈现正相关，蓝底反斜线方块表示两个指标间呈现负相关，方块的颜色越深表明两属性之间的相关性越强。图 4-2 中上三角部分，相关系数大小由饼图的填充面积表示：正相关的指标，饼图顺时针填充；负相关的指标，饼图逆时针填充。

从图 4-2 和表 4-1 中我们发现，$PM_{2.5}$与PM_{10}、CO、NO_2、SO_2、O_3之间相关系数的符号均为正，说明$PM_{2.5}$的质量浓度会随着其他污染物浓度的增加而增加。此外，$PM_{2.5}$与PM_{10}的相关性最强，其 Pearson 相关系数为 0.74，其次是$PM_{2.5}$与NO_2，其 Pearson 相关系数为 0.61，$PM_{2.5}$与 CO 以及$PM_{2.5}$与O_3之间的 Pearson 相关系数分别为 0.39 和 0.09，基本没有线性相关性。这与图 4-1 的分析结果一致。

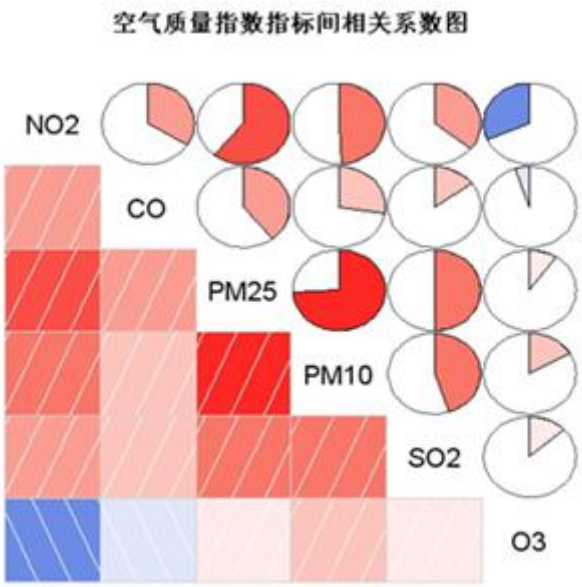


图 4-2 空气质量指数指标间相关系数图

4.1.3 $PM_{2.5}$ 与气象因子之间的相关分析

本文不仅分析了 $PM_{2.5}$ 与其他污染物之间的相关性，还将气象因子纳入研究范围。主要包括 $PM_{2.5}$ 质量浓度与气压、气温、相对湿度、降水量、风速之间的关系。绘制因子间散点图，见图 4-3。从图 4-3 中可以看出，$PM_{2.5}$ 与气象因子之间的相关性并不十分明显，除与气压呈现正相关关系外，$PM_{2.5}$ 与其他气象因子多表现为负相关关系，其中可以看出的是，随着气温的升高，$PM_{2.5}$ 质量浓度有下降趋势。与传统观念不同的是，并不能从散点图中看出，随着风速的增加 $PM_{2.5}$ 质量浓度会有所下降，相反，$PM_{2.5}$ 质量浓度与风速之间基本不存在线性相关性。从核密度曲线看，气压、气温的核密度曲线基本呈高斯分布，相对湿度在 60% 以上的概率较大，降水量多集中在 0.5mm 以下，风速有较大概率低于 1m/s。

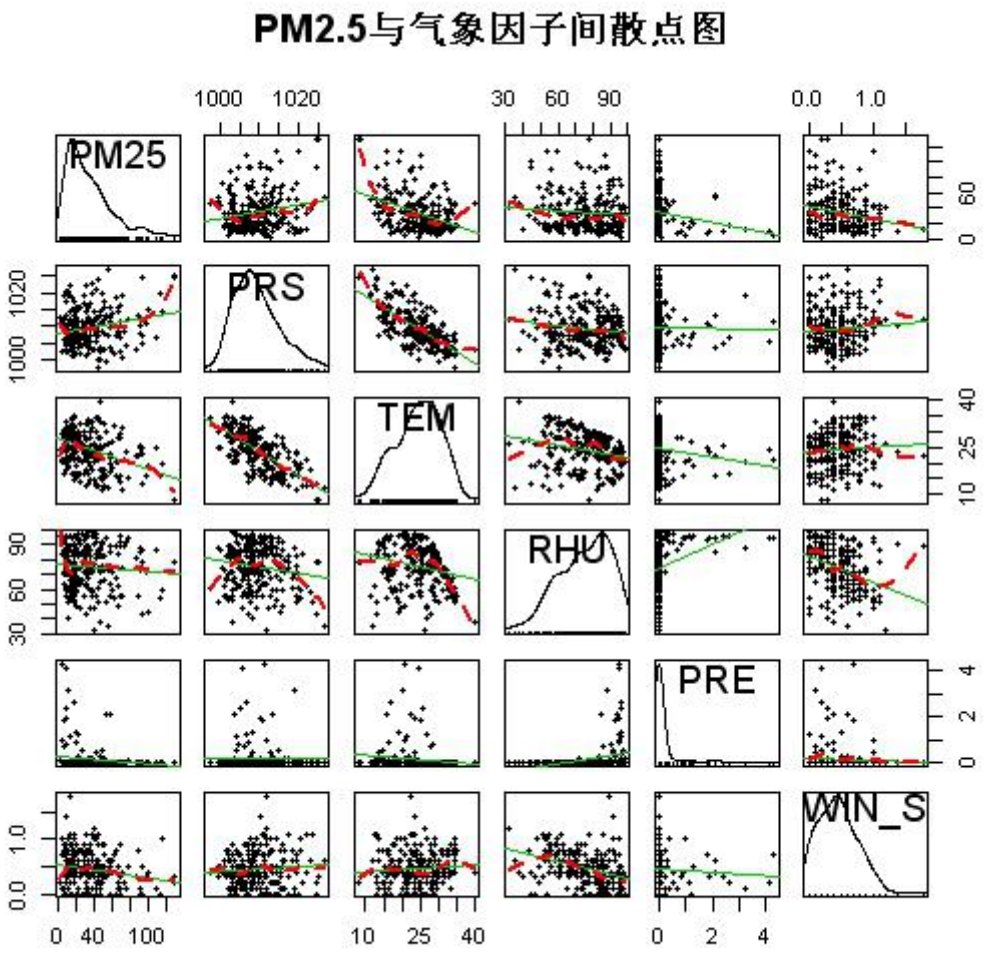


图 4-3 $PM_{2.5}$ 与气象因子间散点图

进一步定量地计算 $PM_{2.5}$ 与气象因子之间的 Pearson 相关系数矩阵，计算结果见表 4-2。同时，根据表中所示相关系数值绘制 $PM_{2.5}$ 与气象因子间的相关系数图，见图 4-4，图 4-4 中各元素的意义与图 4-2 相同。

表 4-2 PM2.5 与气象因子之间相关系数矩阵

	$PM_{2.5}$	PRS	TEM	RHU	PRE	WIN_S
$PM_{2.5}$	1.00	0.14	-0.28	-0.04	-0.04	-0.16
PRS	0.14	1.00	-0.73	-0.28	-0.06	0.19
TEM	-0.28	-0.73	1.00	-0.15	-0.04	0.04
RHU	-0.04	-0.28	-0.15	1.00	0.19	-0.51
PRE	-0.04	-0.06	-0.04	0.19	1.00	-0.06
WIN_S	-0.16	0.19	0.04	-0.51	-0.06	1.00

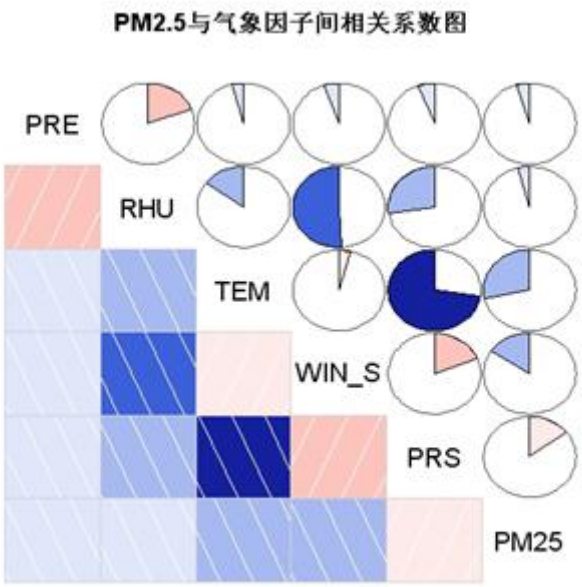


图 4-4 PM 与气象因子间相关系数图

从相关图 4-4 中可以看出，$PM_{2.5}$ 与气象因子之间的相关性并不明显，较强的相关性仅出现在气温与气压之间，气温和气压存在较强的负相关关系，当气温升高时，气压会随之降低。从具体数据看，$PM_{2.5}$ 与气象因子之间的 Pearson 相关系数绝对值都没有超过 0.3，说明 $PM_{2.5}$ 与气象因子间的整体相关性较弱。从相关系数的符号上看，$PM_{2.5}$ 与气压之间的相关系数符号为正，说明随着气压的升高，$PM_{2.5}$ 质量浓度会有所上升；$PM_{2.5}$ 与气温、相对湿度、降水量、风速之间相关系数的符号方向为负，说明说明随着这些气象因子数值的升高，$PM_{2.5}$ 质量浓度会有所下降。$PM_{2.5}$ 与风速之间的 Pearson 相关系数是-0.16，虽然呈现负相关，但整体意义不大，与高风速天气环境下会降低大气中污染物浓度的传统映像有一定的差距。数据与图分析的结果基本一致。

4.2 $PM_{2.5}$ 质量浓度逐步回归仿真

多元线性回归模型描述了多个预测变量与一个响应变量（也就是多个自变量和一个因变量）之间的线性关系，通过建立 PM_{10}、CO、NO_2、SO_2、O_3 与 $PM_{2.5}$ 的多元回归仿真模型，可以知道在特定的污染物浓度环境下，$PM_{2.5}$ 的质量浓度是多少。并通过向预测变量中加入离子浓度因子和气象因子，一方面为提高仿真模型的准确度，另一方面也希望发现其中一些潜在的规律。

4.2.1 逐步回归原理

在建立回归模型解决实际问题时，受限于计算成本和解释难度，人们总是希望多元回归模型的变量都是有效的。这就需要对变量按照一定的标准进行筛选，从而到达“最理想”的模型。所谓“最理想”的标准是，模型的准确性尽可能得高，同时，模型的复杂度尽可能得低。这就需要使每一个加入到回归模型中的自变量对于因变量的影响都具有较高的显著性，而对于因变量影响不显著的自变量将被剔除在回归模型之外。

于是，引出多元回归模型自变量自动搜索的逐步回归方法，该方法的基本思想是，在每一步向模型中添加或者删除一个变量，直到模型满足某个判停标准为止。按照增删变量的方向不同，逐步回归分为向前逐步回归（forward stepwise）、向后逐步回归（backward stepwise）、向前向后逐步回归（stepwise stepwise）。向前逐步回归是指，在拟合回归模型的过程中每一步向模型添加一个变量，直到添加模型外的变量后不再能提高模型的拟合优度为止；向后逐步回归是指，每一步从模型的已有变量中删除一个，直到删除变量的行为会降低模型的拟合优度为止；向前向后逐步回归结合了向前逐步回归和向后逐步回归的特点，每次添加或删除变量时会重新计算变量对模型的贡献，因此，有的变量可能会被添加、删除多次，直到模型稳定为止。

本文选择的逐步回归判停标准是 AIC（Akaike Information Criterion，赤池信息准则）。AIC 估计量综合考虑了模型的复杂度和输出结果的拟合优度，提供了一个对于回归模型在拟合过程中信息损失大小的相对估计。AIC 的值越小越好，它说明该模型既能很好地解释已有数据，又含有较少的参数。AIC 值的计算见公式（4-2）：

$$AIC = -2\ln\left(\hat{L}\right) + 2k \tag{4-2}$$

其中，$\hat{L}$ 是模型的极大似然函数，k 是模型中独立变量的个数。

本文选择使用向前向后的逐步回归方法，其一般步骤是：

(1)准备两个集合 A 和 B，集合 A 负责存放模型中存在的变量，也可以说是

待删除的变量，集合 A 的长度为 i，集合 B 负责存放从模型中删除的变量，也就是待增加的变量，集合 B 的长度为 j；

(2)计算 A 集合中的每一个变量被删除后模型的 AIC 值为 a_{1i}，计算 B 集合中的每一个变量增加到模型中时模型的 AIC 值为 a_{2j}，计算模型没有变量的增删操作时模型的 AIC 值为 a_3；

(3)当 $\min(a_{1i})<\min(a_{2j})$且 $\min(a_{1i}, a_{2j})<a_3$ 时，将集合 A 中 $\min(a_{1i})$对应的变量添加到集合 B 中；

(4)当 $\min(a_{2j})<\min(a_{1i})$且 $\min(a_{1i}, a_{2j})<a_3$ 时，将集合 B 中 $\min(a_{2j})$对应的变量添加到集合 A 中；

(5)当 $a_3<\min(a_{1i}, a_{2j})$时，模型稳定，不再做变量的增删操作；

(6)重复第(2)~(5)步，直到模型稳定为止；

(7)使用 A 集合中的变量拟合生成回归模型，并计算模型中各参数的值。

4.2.2 $PM_{2.5}$ 质量浓度逐步回归模型

首先，建立 $PM_{2.5}$ 与 PM_{10}、CO、NO_2、SO_2、O_3 的多元回归模型，模型各参数见表 4-3。

表 4-3 污染物多元回归模型参数

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-19.28	2.03	-9.48	< 2E-16
PM_{10}	0.25	0.02	16.51	< 2E-16
CO	10.56	1.87	5.65	2.28E-08
NO_2	0.45	0.03	14.53	< 2E-16
SO_2	0.83	0.12	7.11	2.69E-12
O_3	0.08	0.01	6.64	5.82E-11
Residual standard error: 14.79			p-value: < 2.2E-16	
Multiple R-squared: 0.6607			Adjusted R-squared: 0.6586	

表 4-3 的变量参数中，“Estimate”列给出了回归模型中每个系数的估计值。“Std. Error”列是各变量回归系数估计值对应的标准误差，反映了对各系数变化程度的估计。可以看出，除了截距项和 CO 变量的回归系数变化程度相对较大外，PM_{10}、NO_2、SO_2、O_3 对应回归系数的变化程度均较小。“t value”和“Pr(>|t|)”反映了变量的重要程度，也就是变量的回归系数为 0 的概率。“t value”的定义

是每个变量的回归系数估计值与其对应的标准误差的比值。从“Pr(>|t|)”列可以看出，每个变量其原假设成立的概率均较小，所有变量的 Pr(>|t|)值都在 0.0001 以下，也就是说，有 99.99%的置信度可以认定变量的回归系数不为 0，每一个预测变量对于响应变量的影响都是显著的。

表 4-3 的模型参数中，14.79 的残差标准误可以认为是模型对 $PM_{2.5}$ 质量浓度仿真的平均误差。“p-value”的原假设是响应变量与每一个预测变量都没有关系，也就是每一个预测变量的回归系数都为 0 的概率，从< 2.2E-16 的 p 值可以知道原假设不成立，模型要首先通过这个检验，否则单独对每个变量做 t 检验就没有意义。

多元 R^2 和调整 R^2 反映的是模型的拟合优度，也就是模型预测结果与实际数据之间的相关系数，R^2 的值在 0~1 之间，R^2 越大，说明模型对于真实值的拟合程度越好。该模型解释了 $PM_{2.5}$ 质量浓度 66%左右的方差，结果并不理想。

从表 4-3 中可以得到多元回归模型（4-3）：

$$Y_{PM25} = -19.28 + 0.25X_{PM10} + 10.56X_{CO} + 0.45X_{NO2} + 0.83X_{SO2} + 0.08X_{O3} \quad (4\text{-}3)$$

从式 4-3 中可以看出，每个污染物浓度的回归系数都为正，也就是 $PM_{2.5}$ 与其他污染物之间都是正相关，当固定其他污染物浓度不变时，每一个污染物质量浓度升高时，$PM_{2.5}$ 的质量浓度也将相应地升高。从每一个预测变量的系数看，PM_{10} 的质量浓度每升高或降低 1μg/m³，$PM_{2.5}$ 的质量浓度将预期升高或降低 0.25μg/m³；CO 的质量浓度每升高或降低 1μg/m³，$PM_{2.5}$ 的质量浓度将预期升高或降低 10.56μg/m³；NO_2 的质量浓度每升高或降低 1μg/m³，$PM_{2.5}$ 的质量浓度将预期升高或降低 0.45μg/m³；SO_2 的质量浓度每升高或降低 $_1$μg/m³，$PM_{2.5}$ 的质量浓度将预期升高或降低 0.83μg/m³；O_3 的质量浓度每升高或降低 1μg/m³，$PM_{2.5}$ 的质量浓度将预期升高或降低 0.08μg/m³。从数据上看，相较于其他污染物，CO 的质量浓度变化将有可能造成 $PM_{2.5}$ 的质量浓度更快地变化。因此，应该尽可能降低空气中 CO 的质量浓度。

可以看出，只有 5 个污染物浓度因子的多元回归模型的拟合效果并不理想，于是，将污染物因子 102~107、离子浓度因子 201~208、气象因子 301~308 共 22

个预测变量进行逐步回归构建模型，逐步回归步骤见表 4-4。

在模型的初始状态，所有 22 个变量均在集合 A 中，也就是用所有变量进行回归，此时模型的 AIC 值为 3627；

第 2 步，计算集合 A 中每一个变量被删除后模型的 AIC 值，将 AIC 值最小的 206 加入到集合 B，再通过集合 A 中剩余变量回归，此时模型的 AIC 值为 3625；

第 3 步，计算集合 A 中每一个变量被删除后模型的 AIC 值，以及集合 B 中每一个变量被加入到模型后的 AIC 值，将 AIC 值最小的 303 加入到集合 B，再通过集合 A 中剩余变量回归，此时模型的 AIC 值为 3623；

第 4 步，计算集合 A 中每一个变量被删除后模型的 AIC 值，以及集合 B 中每一个变量被加入到模型后的 AIC 值，将 AIC 值最小的 308 加入到集合 B，再通过集合 A 中剩余变量回归，此时模型的 AIC 值为 3622；

第 5 步，计算集合 A 中每一个变量被删除后模型的 AIC 值，以及集合 B 中每一个变量被加入到模型后的 AIC 值，将 AIC 值最小的 305 加入到集合 B，再通过集合 A 中剩余变量回归，此时模型的 AIC 值为 3620；

第 6 步，计算集合 A 中每一个变量被删除后模型的 AIC 值，以及集合 B 中每一个变量被加入到模型后的 AIC 值，将 AIC 值最小的 307 加入到集合 B，再通过集合 A 中剩余变量回归，此时模型的 AIC 值为 3619；

第 7 步，计算集合 A 中每一个变量被删除后模型的 AIC 值，以及集合 B 中每一个变量被加入到模型后的 AIC 值，将 AIC 值最小的 304 加入到集合 B，再通过集合 A 中剩余变量回归，此时模型的 AIC 值为 3618；

第 8 步，计算集合 A 中每一个变量被删除后模型的 AIC 值，以及集合 B 中每一个变量被加入到模型后的 AIC 值，将 AIC 值最小的 208 加入到集合 B，再通过集合 A 中剩余变量回归，此时模型的 AIC 值为 3617；

第 9 步，计算集合 A 中每一个变量被删除后模型的 AIC 值，以及集合 B 中每一个变量被加入到模型后的 AIC 值，将 AIC 值最小的 207 加入到集合 B，再通过集合 A 中剩余变量回归，此时模型的 AIC 值为 3616；

最后，不再对集合 A 和集合 B 内变量做增删操作时模型的 AIC 值最小，逐

步回归结束。

表 4-4 逐步回归步骤

Step	A & B	AIC
1	A={102,103,104,105,106,107,201,202,203,204,205,206, 207,208,301,302,303,304,305,306,307,308} B={}	3627
2	A={102,103,104,105,106,107,201,202,203,204,205,207, 208,301,302,303,304,305,306,307,308} B={206}	3625
3	A={102,103,104,105,106,107,201,202,203,204,205,207, 208,301,302,304,305,306,307,308} B={206,303}	3623
4	A={102,103,104,105,106,107,201,202,203,204,205,207, 208,301,302,304,305,306,307} B={206,303,308}	3622
5	A={102,103,104,105,106,107,201,202,203,204,205,207, 208,301,302,304,306,307} B={206,303,305,308}	3620
6	A={102,103,104,105,106,107,201,202,203,204,205,207, 208,301,302,304,306} B={206,303,305,307,308}	3619
7	A={102,103,104,105,106,107,201,202,203,204,205,207, 208,301,302,306} B={206,303,304,305,307,308}	3618
8	A={102,103,104,105,106,107,201,202,203,204,205,207, 301,302,306} B={206,208,303,304,305,307,308}	3617
9	A={102,103,104,105,106,107,201,202,203,204,205,301, 302,306} B={206,207,208,303,304,305,307,308}	3616

建立 $PM_{2.5}$ 与集合 A 中变量的多元回归模型，模型各参数见表 4-5。加入新的预测变量后，除截距项回归系数估计值的标准误差较大外，其他变量回归系数估计值的标准误差均较小。203（硫酸根离子）和 306（风速）分别有 11%和 12%的置信度认为这两个变量的回归系数为 0，说明其对响应变量的贡献较小。其他变量回归系数为 0 的概率较小，对响应变量的影响更显著。

就新的回归模型而言，整体 9.58 的残差标准误相较于污染物模型降低了 5.21，R^2 值也从 66%提高到了 85%以上，新模型可以解释 $PM_{2.5}$ 质量浓度 85%的方差，说明新模型的拟合能力较强，模型仿真结果与实际值更加吻合。模型 p 值<

2.2E-16，说明所有变量的回归系数全为 0 的概率极低，也拒绝了响应变量与所有预测变量均无关的假设。

表 4-5 逐步回归模型参数

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	340.81	103.25	3.30	0.001
102	0.20	0.01	19.30	< 2E-16
103	0.09	0.01	7.42	2.97E-13
104	-0.06	0.01	-4.86	1.46E-06
105	4.39	1.61	2.73	0.006
106	0.71	0.09	8.13	1.63E-15
107	0.21	0.02	8.94	< 2E-16
201	2.47	0.52	4.77	2.19E-06
202	0.37	0.19	1.97	0.049
203	0.39	0.24	1.60	0.11
204	-9.46	2.43	-3.89	1.09E-04
205	-6.87	1.53	-4.45	9.66E-06
301	-0.34	0.10	-3.32	9.42E-04
302	-0.63	0.10	-6.29	5.39E-10
306	1.48	0.95	1.56	0.12
Residual standard error: 9.58			p-value: < 2.2E-16	
Multiple R-squared: 0.8593			Adjusted R-squared: 0.8568	

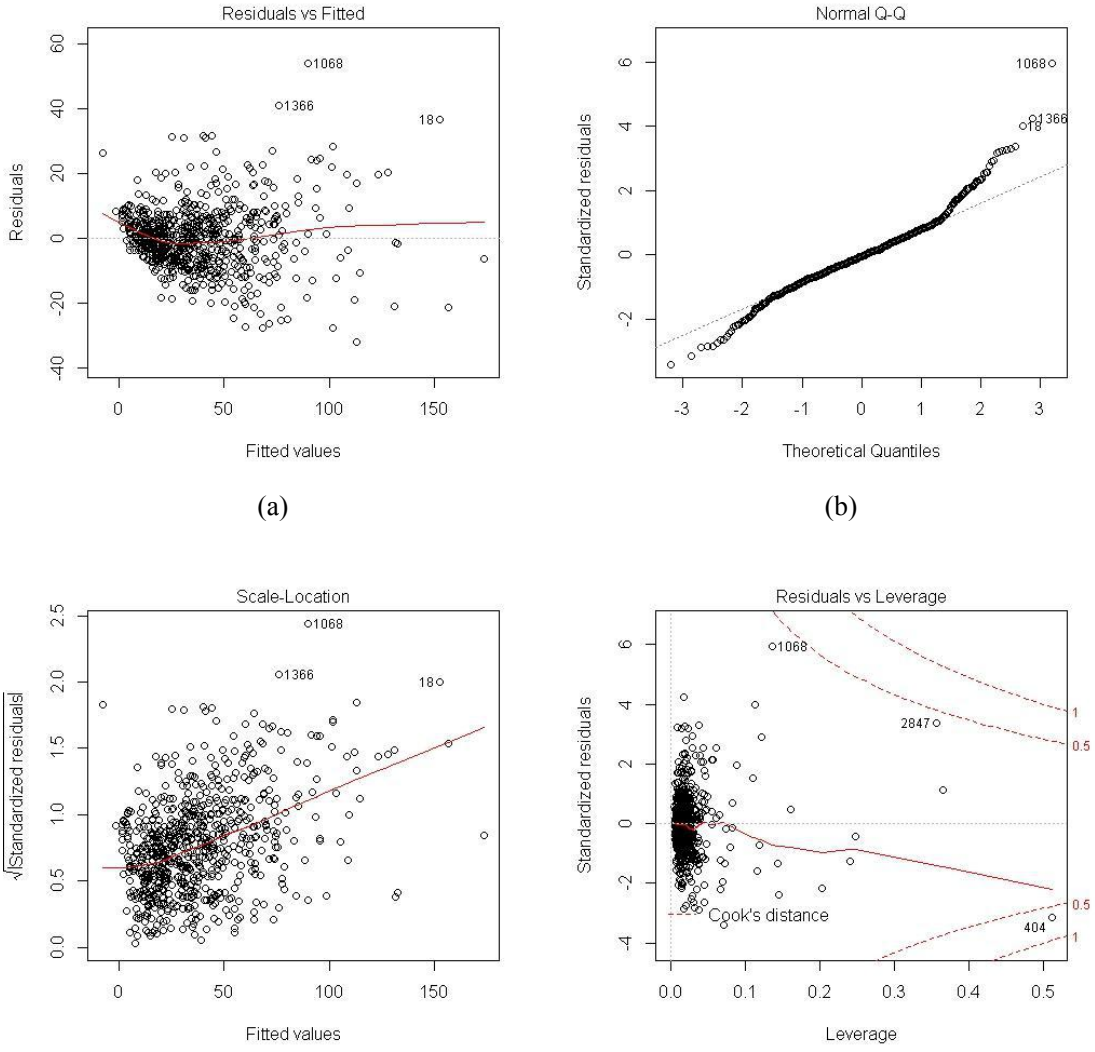
从表 4-5 中得到多元回归模型（4-4）：

$$\begin{aligned} Y_{101} &= 340.81 + 0.2X_{102} + 0.09X_{103} - 0.06X_{104} + 4.39X_{105} \\ &+ 0.71X_{106} + 0.21X_{107} + 2.47X_{201} + 0.37X_{202} + 0.39X_{203} \\ &- 9.46X_{204} - 6.87X_{205} - 0.34X_{301} - 0.63X_{302} + 1.48X_{306} \end{aligned} \quad (4\text{-}4)$$

式 4-4 中回归系数表明，污染物因子与离子浓度因子对 $PM_{2.5}$ 质量浓度的影响以正相关为主，气象因子对 $PM_{2.5}$ 质量浓度的影响以负相关为主。比较特殊的是，204（钾离子）质量浓度每升高 1 个单位，$PM_{2.5}$ 质量浓度将降低 9.46 个单位；205（钙离子）质量浓度每升高 1 个单位，$PM_{2.5}$ 质量浓度将降低 6.87 个单位。306（风速）每升高 1 个单位，$PM_{2.5}$ 质量浓度也将升高 1.48 个单位，说明上海市 $PM_{2.5}$ 污染程度相较于周边地区处于较低水平，空气流动有可能将大气中的污染物从浓度较高的区域传送到浓度较低的区域，所以并不能笼统地概括为风速越高就越有助于当前区域污染物的消散。

4.2.3 回归诊断

回归模型建立后，还需要对其是否满足统计假设进行检验，包括线性、正态性、同方差性以及强影响点的检验。统计假设的成立与否，将决定模型的显著性检验结果以及置信区间是否精确。图 4-5 给出了对多因子多元逐步回归模型诊断的结果。



(a) (b)

(c) (d)

图 4-5 逐步回归诊断图

图 4-5(a)用于检验线性假设。线性假设是指所选择的预测变量和响应变量之间是的关系必须是线性的。若满足线性假设，模型的残差值和拟合值就应该没有关联，图中点应该在水平线上下随机分布。从图 4-5(a)可以看出，模型基本满足线性假设。

图 4-5(b)用于检验正态性假设。正态性假设是指当预测变量的值固定时，响

应变量值的分布是正态分布，而残差值也应该满足正态分布。若满足正态性假设，在分位数-分位数图中，所有点应该分布在 45 度直线上。从图 4-5(b)可以看出，回归模型基本满足正态性假设。

图 4-5(c)用于检验同方差性假设。同方差性假设是指响应变量的方差不随预测变量值的变化而发生改变。若满足同方差性假设，图中点应该在水平线周围随机分布。从图 4-5(c)中可以看出，误差有增大的趋势，产生异方差的原因可能是响应变量与预测变量之间还存在着某些非线性关系。

图 4-5(d)可以识别数据中的强影响点。强影响点是指会显著影响模型斜率和截距的数据点。检测是否为强影响点一般使用 Cook 距离统计量作为判定标准，若 Cook 距离大于 1，则认为该点为强影响点。从图 4-5(d)中可以看出，第 1068 和 2847 条记录的 Cook 距离接近 0.5，第 404 条记录的 Cook 距离在 0.5 至 1 之间，没有出现 Cook 距离大于 1 的强影响点。

第五章 基于随机森林的 $PM_{2.5}$ 小时浓度预测

本章将建立基于随机森林的 $PM_{2.5}$ 质量浓度预测模型，对未来 1~6 小时的 $PM_{2.5}$ 质量浓度进行逐小时的预测，以及对未来 6~12 小时、12~24 小时、24~48 小时 $PM_{2.5}$ 质量浓度的最高值和最低值进行预测。最后，对随机森林模型的预测结果进行检验。

5.1 $PM_{2.5}$ 污染物 1~6 小时逐小时质量浓度值预测

5.1.1 数据准备

生成随机森林模型所需要的预测变量共分为 4 部分，包含了实测因子、周期因子、预报因子和预测因子，其结构见表 5-1。其中，因子 A 是当前在线收集到的各项实测数据，包括当前 $PM_{2.5}$、PM_{10}、O_3、CO、SO_2、NO_2 等污染物的质量浓度，当前 NH_4^+、NO_3^-、SO_4^{2-}、K^+、Ca^{2+}、Na^+、Mg^{2+}、Cl^-等离子的质量浓度，当前气压、气温、相对湿度、降水量、风向、风速、能见度等气象情况；因子 B 是反映 $PM_{2.5}$ 质量浓度周期变化规律的数据，包括 $PM_{2.5}$ 的月平均质量浓度以及周平均质量浓度；因子 C 是所需预测时间段的气象预报，包括气压、气温、相对湿度、降水量、风向、风速，但不包含能见度；因子 D 是各模型逐小时预测的结果，该因子的加入是考虑了 $PM_{2.5}$ 质量浓度数据自身在时间序列上的变化规律。因子 A、因子 B、因子 C 在模型生成前已经准备完成，而因子 D 需要在预测过程中不断加入到训练数据集中。另外，模型的响应变量为待预测时段的 $PM_{2.5}$ 质量浓度实际值。最后，把原数据集按 9:1 的比例分为训练数据集和检验数据集，本文抽取原数据集中约 90%的数据进行模型训练，训练数据共 3446 条，剩余的 374 条数据用来检验模型性能。

表 5-1 逐小时模型的预测变量

因子 A	因子 B	因子 C	因子 D
当前的实测数据： 污染物质量浓度（101~107）； 离子质量浓度（201~208）； 气象数据（301~310）；	周期数据： $PM_{2.5}$ 月平均质量浓度（401）； $PM_{2.5}$ 周平均质量浓度（402）；	C1：1 小时后气象预报数据（包括气压、气温、相对湿度、降水量、风向、风速）； C2：2 小时后气象预报数据； C3：3 小时后气象预报数据； C4：4 小时后气象预报数据； C5：5 小时后气象预报数据； C6：6 小时后气象预报数据；	D1：1 小时后 $PM_{2.5}$ 质量浓度预测值； D2：2 小时后 $PM_{2.5}$ 质量浓度预测值； D3：3 小时后 $PM_{2.5}$ 质量浓度预测值； D4：4 小时后 $PM_{2.5}$ 质量浓度预测值； D5：5 小时后 $PM_{2.5}$ 质量浓度预测值； D6：6 小时后 $PM_{2.5}$ 质量浓度预测值；

5.1.2 预测步骤

在对 1~6 小时逐小时 $PM_{2.5}$ 的质量浓度进行预测时，使用 6 个随机森林模型分别对每小时的 $PM_{2.5}$ 质量浓度进行预测，每个模型所需的训练数据集是在动态变化的，其具体步骤见图 5-1：

（1）构建用于预测 1 小时后 $PM_{2.5}$ 质量浓度的随机森林模型 1，此时将当前的各项实测数据即因子 A、周期数据即因子 B、1 小时后的气象预报数据即因子 C1 作为模型的输入因子，模型训练完成后，预测未来第 1 个小时的 $PM_{2.5}$ 质量浓度，预测值为 D1；

（2）构建用于预测 2 小时后 $PM_{2.5}$ 质量浓度的随机森林模型 2，此时的输入因子在因子 A 和因子 B 的基础上，将气象预报数据换成 2 小时后的预报数据即因子 C2，并将模型 1 的预测值即因子 D1 加入到训练数据中，模型训练完成后，预测未来第 2 个小时的 $PM_{2.5}$ 质量浓度，预测值为 D2；

（3）构建用于预测 3 小时后 $PM_{2.5}$ 质量浓度的随机森林模型 3，此时的输入因子包括因子 A 和因子 B，气象预报数据换成 3 小时后的数据即因子 C3，并将模型 1、模型 2 的预测值即因子 D1、D2 加入到训练数据中，模型训练完成后，预测未来第 3 个小时的 $PM_{2.5}$ 质量浓度，预测值为 D3；

（4）构建用于预测 4 小时后 $PM_{2.5}$ 质量浓度的随机森林模型 4，此时的输入因子包括因子 A 和因子 B，气象预报数据换成 4 小时后的数据即因子 C4，并将模型 1、模型 2、模型 3 的预测值即因子 D1、D2、D3 加入到训练数据中，模型训练完成后，预测未来第 4 个小时的 $PM_{2.5}$ 质量浓度，预测值为 D4；

（5）构建用于预测 5 小时后 $PM_{2.5}$ 质量浓度的随机森林模型 5，此时的输入因子包括因子 A 和因子 B，气象预报数据换成 5 小时后的数据即因子 C5，并将模型 1、模型 2、模型 3、模型 4 的预测值即因子 D1、D2、D3、D4 加入到训练数据中，模型训练完成后，预测未来第 5 个小时的 $PM_{2.5}$ 质量浓度，预测值为 D5；

（6）构建用于预测 6 小时后 $PM_{2.5}$ 质量浓度的随机森林模型 6，此时的输入因子包括因子 A 和因子 B，气象预报数据换成 6 小时后的数据即因子 C6，并将模型 1、模型 2、模型 3、模型 4、模型 5 的预测值即因子 D1、D2、D3、D4、D5 加入到训练数据中，生成模型后，预测未来第 6 个小时的 $PM_{2.5}$ 质量浓度，预测值为 D6。

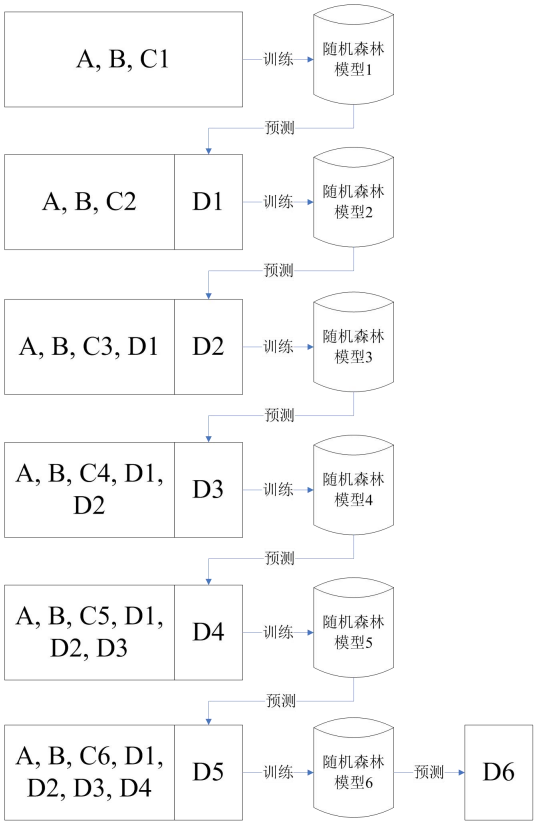


图 5-1 1-6 逐小时模型预测流程

5.1.3 构建模型

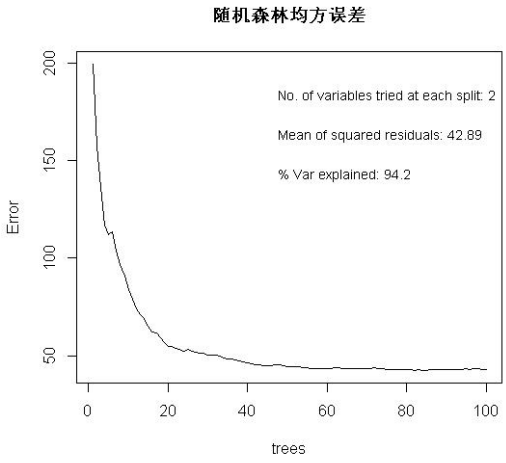
随机森林模型的构建一般包括以下几步：

（1）确定随机森林模型的各个参数，其中的主要参数包括随机森林中树的棵数以及每棵树的内部节点分裂时所需抽取候选变量的数量；

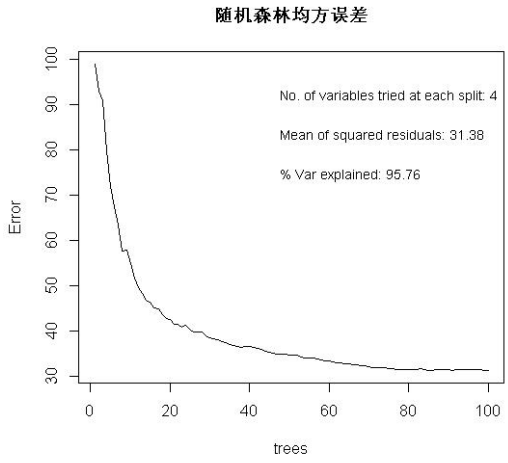
（2）对输入的变量进行筛选，本文选用 OOB 误差估计后的准确率降低值作为筛选标准，筛选后的变量将构成新的训练集参与模型建立；

（3）对模型的性能进行评估，主要是验证模型在检验集，也就是未经训练的数据上预测的准确性，本文选择评估模型性能的指标是平均绝对误差、均方误差、标准化后的平均绝对误差、标准化后的均方误差以及拟合优度。

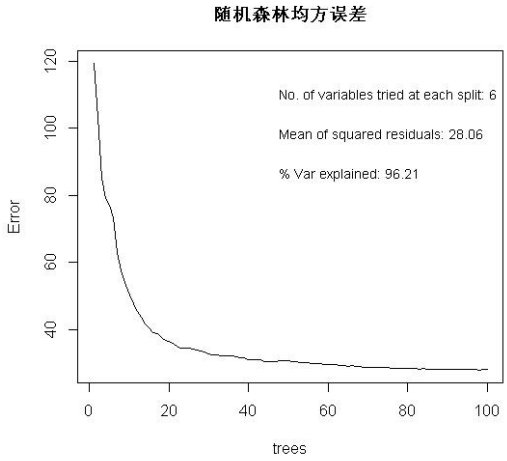
随机森林作为一种组合算法，在生成森林的过程中，单棵树的训练数据集通过自助法从原训练数据集中抽取，在单棵树的分裂过程中，每个内部分裂节点的分裂属性也是分别从原输入因子中抽取，随机森林的最终预测结果综合多棵树的结果决定。因此，对模型性能影响最大的两个参数分别是分裂点抽取属性的个数以及树的棵数。本文通过控制变量的方法，观察当节点分裂属性抽取个数在 2~12 个，森林规模从 0 到 100 时，模型交叉验证结果的均方误差和方差解释程度的变化情况，参数调整结果见图 5-2，图 5-2 的横坐标为随机森林模型中树的棵数，纵坐标为交叉验证后得到模型的均方误差。图 5-2 中展示的是随机森林模型 1 的参数调整情况，也就是用于预测 1 小时后 $PM_{2.5}$ 质量浓度的模型。



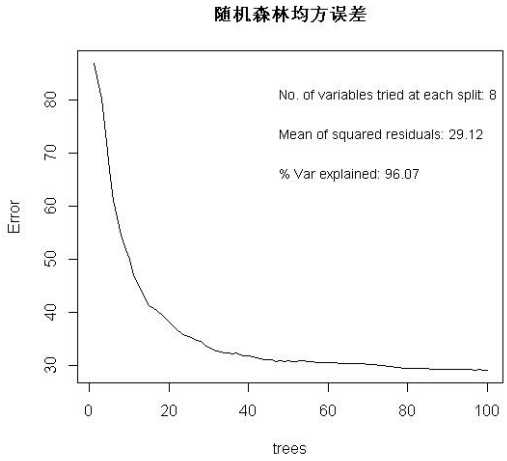
(a)



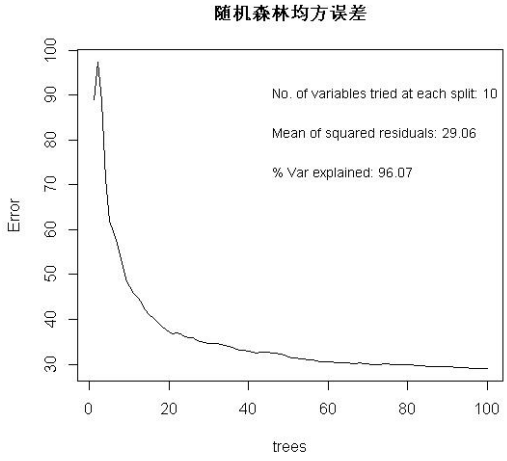
(b)



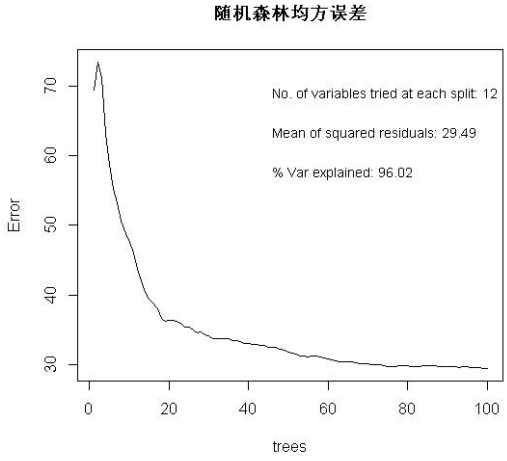
(c)



(d)



(e)



(f)

图 5-2 随机森林模型参数调整

图 5-2(a)表明，当内部节点分裂候选属性数为 2 个时，随着森林中树的棵数从 0 增加到 100，模型的交叉验证均方误差降低至 42.89，能够解释响应变量 94.2%的方差，模型在树的棵数达到 60 时基本趋于稳定；

图 5-2(b)表明，当内部节点分裂候选属性数为 4 个时，随着森林中树的棵数从 0 增加到 100，模型的交叉验证均方误差降低在 31.38，能够解释响应变量 95.76%的方差，模型在树的棵数达到 80 时基本趋于稳定；

图 5-2(c)表明，当内部节点分裂候选属性数为 6 个时，随着森林中树的棵数从 0 增加到 100，模型的交叉验证均方误差降低在 28.06，能够解释响应变量 96.21%的方差，模型在树的棵数达到 100 时基本趋于稳定；

图 5-2(d)表明，当内部节点分裂候选属性数为 8 个时，随着森林中树的棵数从 0 增加到 100，模型的交叉验证均方误差降低在 29.12，能够解释响应变量 96.07%的方差，模型在树的棵数达到 100 时基本趋于稳定；

图 5-2(e)表明，当内部节点分裂候选属性数为 10 个时，随着森林中树的棵数从 0 增加到 100，模型的交叉验证均方误差降低在 29.06，能够解释响应变量 96.07%的方差，模型在树的棵数达到 100 时基本趋于稳定；

图 5-2(f)表明，当内部节点分裂候选属性数为 12 个时，随着森林中树的棵数从 0 增加到 100，模型的交叉验证均方误差降低在 29.49，能够解释响应变量 96.02%的方差，模型在树的棵数达到 100 时基本趋于稳定。

一般认为，单次分裂所抽取的属性数取 $sqrt(p)$ 个左右较为合适，其中，p 是训练数据中变量的个数。因此，结合验证的结果，选择随机森林模型中单棵树内部节点分裂时随机抽取的属性个数为 6，模型规模即随机森林中树的棵数为 100。

在生成随机森林的过程中，每棵树的训练样本通过自助法即有放回的随机抽样获得，这种抽样的结果是，原数据集中的有些数据会被重复抽取进训练集，而有些数据则不会被抽取。利用自助法的这种特性，可以将未被抽取的数据当成检验数据放入检验集，用于验证模型的准确率。随机森林的另一个特点是，在生成单棵树的过程中，内部节点在每次分裂时都会从候选属性集中抽取一定数量的分裂属性，利用这个特点，如果在每次迭代的过程中，随机置换掉其中一个属性，此时若模型的准确性出现很大幅度的降低，则说明该属性对于预测结果的重要性非常高，以此来对输入因子的重要性进行评估。

本文对于衡量输入因子重要性所使用到的是 R 语言提供的 varSelRF 包，该包提供的方法要求响应变量必须是类别类型，因此，需要将 $PM_{2.5}$ 质量浓度的数值属性泛化成类别属性，本文选择的泛化标准是将 $PM_{2.5}$ 质量浓度转化成对应的空气质量等级，其对应关系见表 5-2。

表 5-2 $PM_{2.5}$ 浓度对应空气质量

$PM_{2.5}$($\mu g/m^3$)	IAQI	空气质量级别	空气质量类别
0~35	0~50	一级	优
36~75	51~100	二级	良
76~115	101~150	三级	轻度污染
116~150	151~200	四级	中度污染
151~250	200~300	五级	重度污染
> 250	> 300	六级	严重污染

首先，将 $PM_{2.5}$ 质量浓度的具体数值转化为对应的空气质量分指数（IAQI）。空气质量指数（AQI）是由当前时刻的首要污染物 IAQI 决定的，也就是各污染物中最高的 IAQI 作为当前时刻 AQI，因此在确定当前空气质量级别时需要考虑同一时刻所有污染物的空气质量分指数。但对于本模型，预测的变量仅仅是 $PM_{2.5}$，所以直接将其 IAQI 对应到相应的空气质量指数的级别。因此，将质量浓度在 0~35$\mu g/m^3$ 之间的响应变量泛化成一级，将质量浓度在 36~75$\mu g/m^3$ 之间的响应变量泛化成二级，将质量浓度在 76~115$\mu g/m^3$ 之间的响应变量泛化成三级，将质量浓度在 116~150$\mu g/m^3$ 之间的响应变量泛化成四级，将质量浓度在 151~250$\mu g/m^3$ 之间的响应变量泛化成五级，将质量浓度大于 250$\mu g/m^3$ 的响应变量泛化成六级。如此，响应变量的数值属性就转化成了类别属性。

每次构建随机森林模型时，都需要对输入因子的重要性进行评估，并选择重要程度最高、误差最小且因子数量最少的组合方案。图 5-3 给出了每次因子重要评估的结果，图 5-3 中横坐标是当因子被替换时模型准确率的降低程度，并按照因子的所属类别分别进行排序，包括污染物因子、离子浓度因子、周期因子、气象因子、气象预报因子和预测因子。

图 5-3(a)为预测 1 小时后 $PM_{2.5}$ 质量浓度模型其输入因子的重要性程度排序，从中可以看出，对预测结果影响最大的是当前的 $PM_{2.5}$ 质量浓度，当该因子被替换后，模型的准确率将降低 27%左右，之后分别是当前的铵根离子浓度和当前能见度，被替换后模型准确率将分别降低 8%和 5%左右，从整体看，污染物因子、离子浓度因子、气象因子的重要性程度较高，周期因子、气象预报因子的重要性程度相对较低；

图 5-3(b)为预测 2 小时后 $PM_{2.5}$ 质量浓度模型其输入因子的重要性程度排序，

从中可以看出，对预测结果影响最大的是 1 小时后 $PM_{2.5}$ 质量浓度的预测值，当该因子被替换后，模型的准确率将降低 36%左右，相较于图 5-3(a)，当前 $PM_{2.5}$ 质量浓度的重要性减弱，周期因子中，$PM_{2.5}$ 月平均质量浓度的重要性升高，整体上看，重要性较高的因子均集中在质量浓度数据上；

图 5-3(c)为预测 3 小时后 $PM_{2.5}$ 质量浓度模型其输入因子的重要性程度排序，从中可以看出，对预测结果影响最大的是 2 小时后 $PM_{2.5}$ 质量浓度的预测值，当该因子被替换后，模型的准确率将降低 41%左右，重要性最突出的分别是两个预测因子，铵根离子的重要性明显降低，气象因子中，当前气温成为最主要的因素；

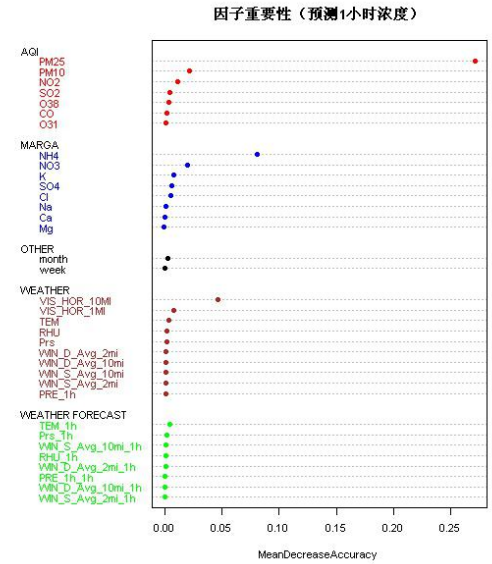
图 5-3(d)为预测 4 小时后 $PM_{2.5}$ 质量浓度模型其输入因子的重要性程度排序，从中可以看出，对预测结果影响最大的是 3 小时后的 $PM_{2.5}$ 质量浓度预测值，当该因子被替换后，模型的准确率将降低 40%左右，重要性最突出因子集中在最近 2 小时的预测值上，气象因子与预报因子重要性之间的差距在减弱；

图 5-3(e)为预测 5 小时后 $PM_{2.5}$ 质量浓度模型其输入因子的重要性程度排序，从中可以看出，对预测结果影响最大的 4 小时后 $PM_{2.5}$ 质量浓度的预测值，当该因子被替换后，模型的准确率将降低 40%左右，污染物因子和气象因子的重要性程度有所提升；

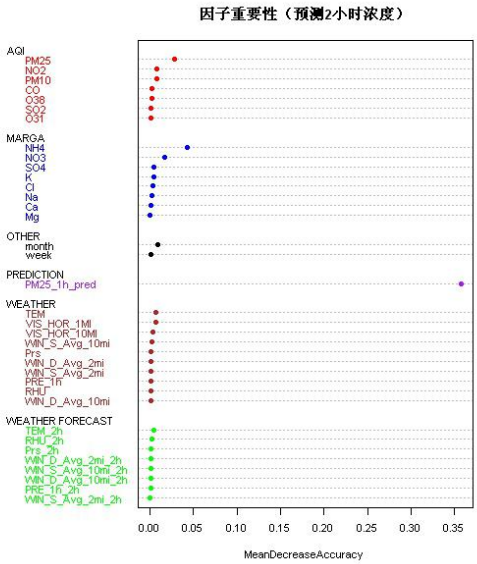
图 5-3(f)为预测 6 小时后 $PM_{2.5}$ 质量浓度模型其输入因子的重要性程度排序，从中可以看出，对预测结果影响最大的 5 小时后 $PM_{2.5}$ 质量浓度的预测值，当该因子被替换后，模型的准确率将降低 38%左右，除预测因子外，其他各类因子重要性之间的差距不再明显。

从 6 个模型的因子重要性排序也可以看出，最重要的因子始终是预报时刻前一小时的 $PM_{2.5}$ 质量浓度数据，但随着时间的向前推移，该类因子的重要性急剧减弱，其他类型因子在预测时也起到相当重要的作用。

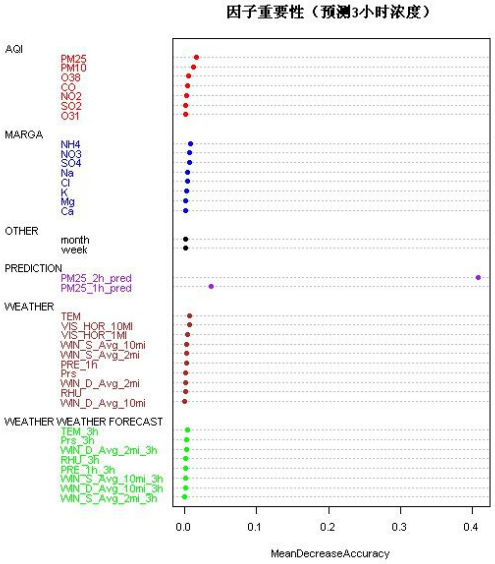
输入因子筛选完成后，将构建预测模型并输出预测结果。


因子重要性（预测1小时浓度）
AQI
PM25
PM10
NO2
SO2
O38
CO
O31
MARGA
NH4
NO3
K
SO4
Cl
Na
Ca
Mg
OTHER
month
week
WEATHER
VIS_HOR_10MI
VIS_HOR_1MI
TEM
RHU
Prs
WIN_D_Avg_2mi
WIN_D_Avg_10mi
WIN_S_Avg_10mi
WIN_S_Avg_2mi
PRE_1h
WEATHER FORECAST
TEM_1h
Prs_1h
WIN_S_Avg_10mi_1h
RHU_1h
WIN_D_Avg_2mi_1h
PRE_1h_1h
WIN_D_Avg_10mi_1h
WIN_S_Avg_2mi_1h
0.00 0.05 0.10 0.15 0.20 0.25
MeanDecreaseAccuracy

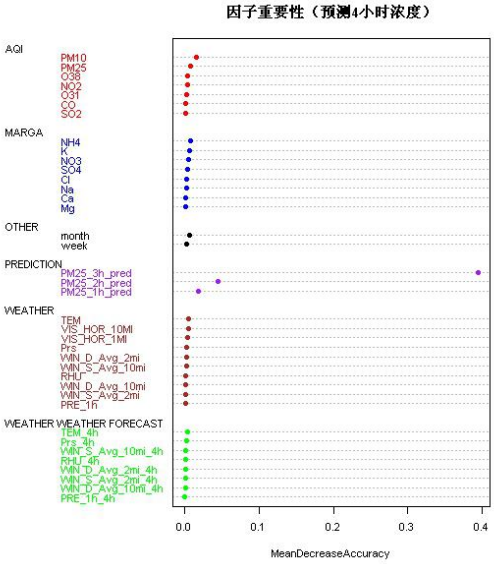
(a)


因子重要性（预测2小时浓度）
AQI
PM25
NO2
PM10
CO
O38
SO2
O31
MARGA
NH4
NO3
SO4
K
Cl
Na
Ca
Mg
OTHER
month
week
PREDICTION
PM25_1h_pred
WEATHER
TEM
VIS_HOR_1MI
VIS_HOR_10MI
WIN_S_Avg_10mi
Prs
WIN_D_Avg_2mi
WIN_S_Avg_2mi
PRE_1h
RHU
WIN_D_Avg_10mi
WEATHER FORECAST
TEM_2h
RHU_2h
Prs_2h
WIN_D_Avg_2mi_2h
WIN_S_Avg_10mi_2h
WIN_D_Avg_10mi_2h
PRE_1h_2h
WIN_S_Avg_2mi_2h
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35
MeanDecreaseAccuracy

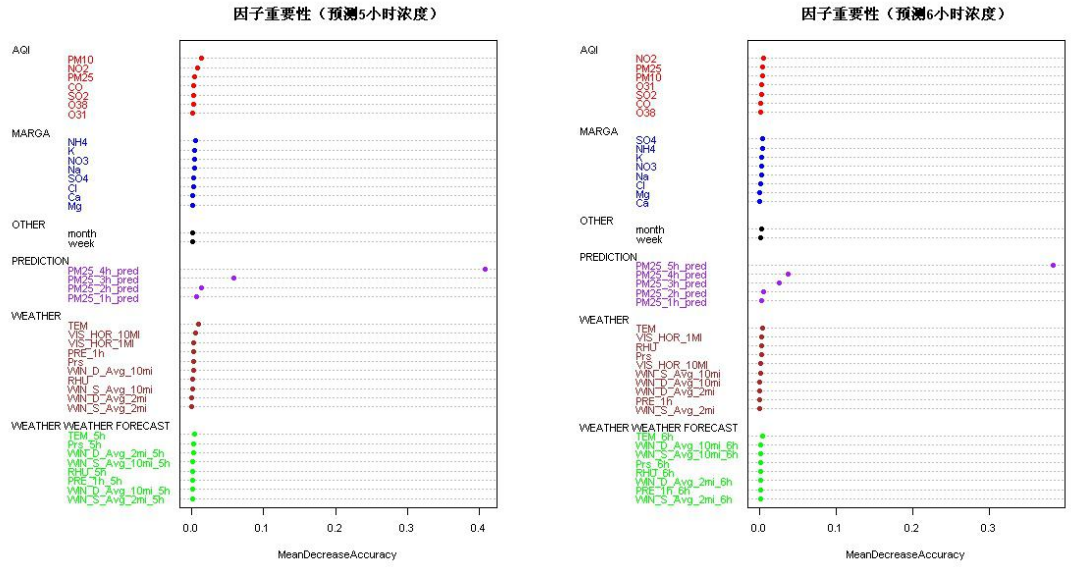
(b)


因子重要性（预测3小时浓度）
AQI
PM25
PM10
O38
CO
NO2
SO2
O31
MARGA
NH4
NO3
SO4
Na
Cl
K
Mg
Ca
OTHER
month
week
PREDICTION
PM25_2h_pred
PM25_1h_pred
WEATHER
TEM
VIS_HOR_10MI
VIS_HOR_1MI
WIN_S_Avg_10mi
WIN_S_Avg_2mi
PRE_1h
Prs
WIN_D_Avg_2mi
RHU
WIN_D_Avg_10mi
WEATHER WEATHER FORECAST
TEM_3h
Prs_3h
WIN_D_Avg_2mi_3h
RHU_3h
PRE_1h_3h
WIN_S_Avg_10mi_3h
WIN_D_Avg_10mi_3h
WIN_S_Avg_2mi_3h
0.0 0.1 0.2 0.3 0.4
MeanDecreaseAccuracy

(c)


因子重要性（预测4小时浓度）
AQI
PM10
PM25
O38
NO2
O31
CO
SO2
MARGA
NH4
K
NO3
SO4
Cl
Na
Ca
Mg
OTHER
month
week
PREDICTION
PM25_3h_pred
PM25_2h_pred
PM25_1h_pred
WEATHER
TEM
VIS_HOR_10MI
VIS_HOR_1MI
Prs
WIN_D_Avg_2mi
WIN_S_Avg_10mi
RHU
WIN_D_Avg_10mi
WIN_S_Avg_2mi
PRE_1h
WEATHER WEATHER FORECAST
TEM_4h
Prs_4h
WIN_S_Avg_10mi_4h
RHU_4h
WIN_D_Avg_2mi_4h
WIN_S_Avg_2mi_4h
WIN_D_Avg_10mi_4h
PRE_1h_4h
0.0 0.1 0.2 0.3 0.4
MeanDecreaseAccuracy

(d)



(e) (f)

图 5-3 模型输入因子重要性

5.1.4 结果分析

使用训练数据构建完模型后，还需要对模型的预测性能进行不同维度的评估，如果所构建的模型性能较差则模型没有意义。验证模型性能主要是通过将检验集输入模型，比较预测值与实际值之间的差异，差异越小则说明模型性能越好、准确性越高。由于检验集并没有参与模型训练，所以使用检验集验证模型性能时不会造成评估过于乐观，能够反映模型的泛化能力。

模型性能的评估有多个方面，比如预测准确性、可解释性、运算效率等，本文主要考虑模型的准确性。使用到的准确性度量有：

拟合优度（R^2），衡量了预测值对实际值的拟合情况，R^2 的值在 0~1 之间，越接近 1（说明模型几乎解释了 100%的方差）越好，计算公式见（5-1）-（5-3）。

$$SS = \sum_{i-1}^{n} (y_i - \bar{y})^2 \tag{5-1}$$

$$RSS = \sum_{i-1}^{n} (y_i - \hat{y}_i)^2 \tag{5-2}$$

$$R^2 = \frac{SS - RSS}{SS} \tag{5-3}$$

式中（5-1）-（5-3），SS 是离差平方和（Sum of Squares of Deviations），指

的是实际值与平均值的差平方和，y_i 是第 i 条记录的实际值，$\bar{y}$ 是实际值的均值，RSS 是残差平方和（Residual Sum of Squares），指的是实际值与预测值的差平方和，$\hat{y}_i$ 是第 i 条记录的预测值。以下符号解释相同。

平均绝对误差（MAE），计算公式见（5-4）：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5-4}$$

规范化后的平均绝对误差（NMAE），计算公式见（5-5）：

$$NMAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \Big/ \frac{1}{n}\sum_{i=1}^{n}|y_i - \bar{y}| \tag{5-5}$$

均方误差（MSE），计算公式见（5-6）：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5-6}$$

规范化后的均方误差（NMSE），计算公式见（5-7）：

$$NMSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{5-7}$$

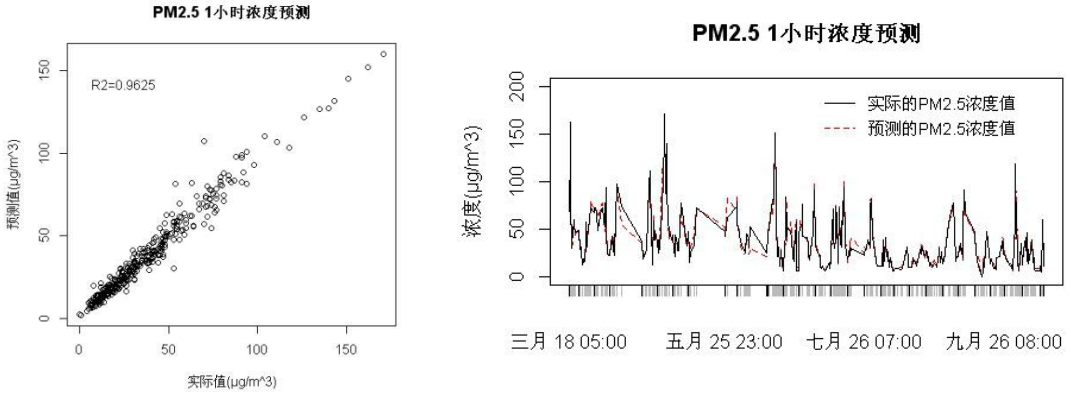
平均绝对误差衡量的是预测结果与实际值之间的绝对差异，但仅仅使用绝对误差判断模型的好坏会产生一定的偏差，这种偏差是由实际值大小区间不同造成的。均方误差作为常用度量，其不足之处是计算后与响应变量的单位变得不一致。规范化后的平均绝对误差与规范化后的均方误差是将预测模型与一个基准模型进行比较，其值越小，说明使用预测模型比基准模型更好。一般设置基准模型为均值模型。若规范化后的平均绝对误差或规范化后的均方误差值等于 1，则说明使用模型进行预测和仅仅使用响应变量的平均值作为预测结果效果是一样的，若值大于 1，则模型的建立完全没有意义。因此，鉴于各衡量标准均有自己的优缺点，本文使用多标准来评估模型的性能。

首先，绘制预测结果与实际值的散点图和折线图，见图 5-4，可以直观地评估模型的预测效果。图 5-4 中预测值是由因子筛选后的随机森林模型得到的。

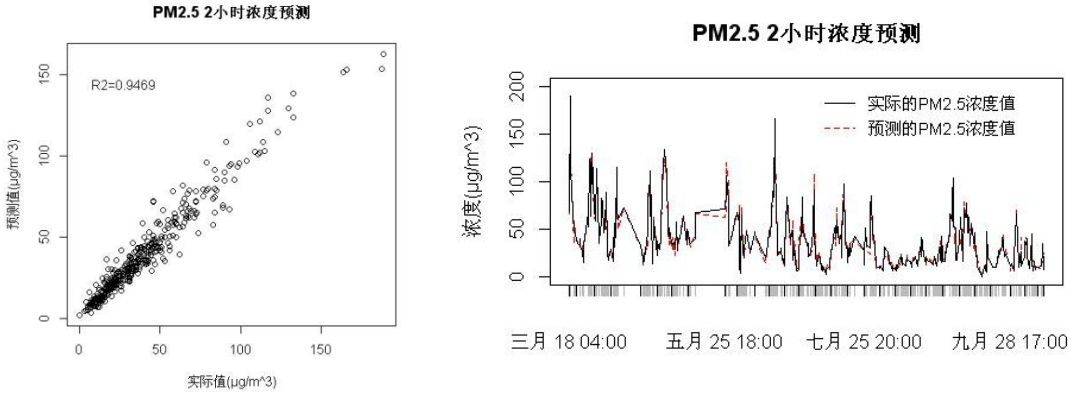
左侧散点图的横坐标是 $PM_{2.5}$ 质量浓度在预测时段的实际值，纵坐标为预测值，若预测结果准确，图 5-4 中各点应该分布在 45 度的斜线上。右侧折线图反

映了当前时段 n 小时后 $PM_{2.5}$ 的质量浓度情况与预测情况，例如，图 5-4(a)反应的是横坐标显示时段 1 小时后的 $PM_{2.5}$ 的质量浓度实际值与预测值，图 5-4(f)反应的是横坐标显示时段 6 小时后的 $PM_{2.5}$ 的质量浓度实际值与预测值。其中，实际值由黑色实线标出，预测值由红色虚线标出。

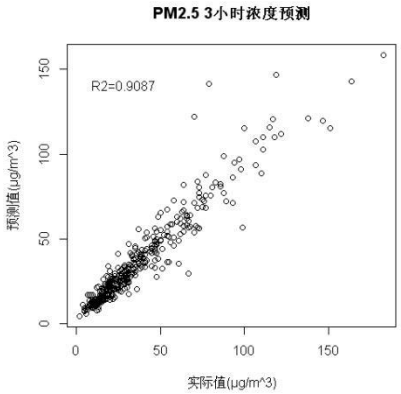
从图 5-4 中对于未来 1~6 小时 $PM_{2.5}$ 浓度的预测情况变化中可以看出，随着预测时间的向后推移，预测值与实际值之间的吻合程度呈现逐渐降低的趋势，但下降幅度在可以接受的范围内，总体吻合程度依然较高。

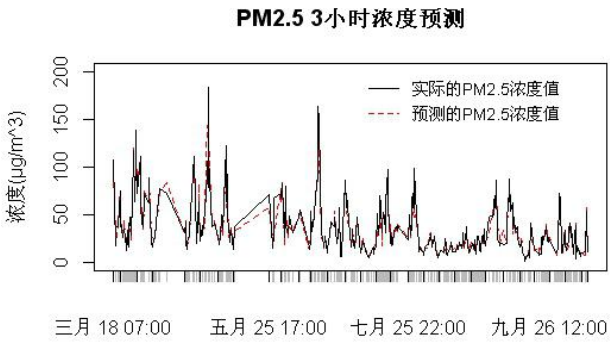


(a)

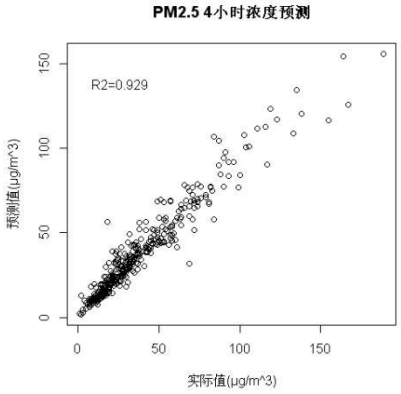


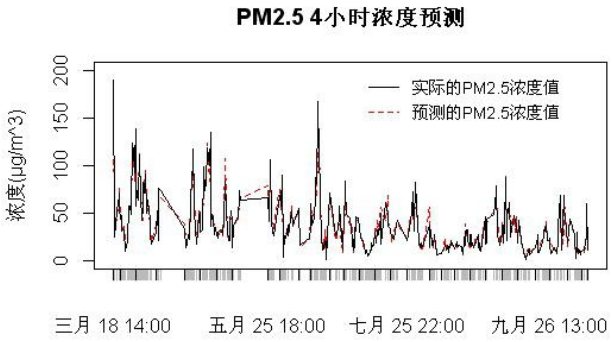
(b)


PM2.5 3小时浓度预测
R2=0.9087
预测值(μg/m^3)
实际值(μg/m^3)
0
50
100
150

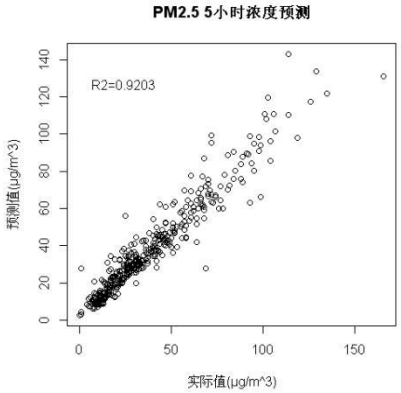

PM2.5 3小时浓度预测
实际的PM2.5浓度值
预测的PM2.5浓度值
浓度(μg/m^3)
三月 18 07:00
五月 25 17:00
七月 25 22:00
九月 26 12:00

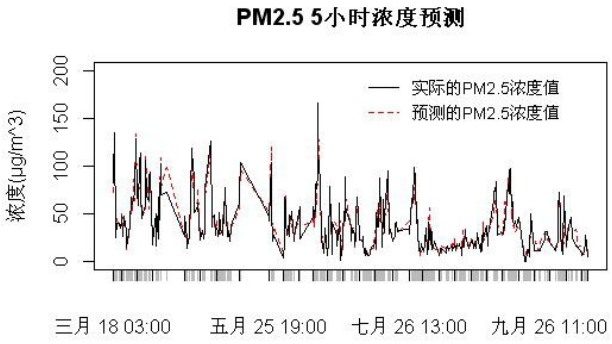
(c)


PM2.5 4小时浓度预测
R2=0.929
预测值(μg/m^3)
实际值(μg/m^3)
0
50
100
150

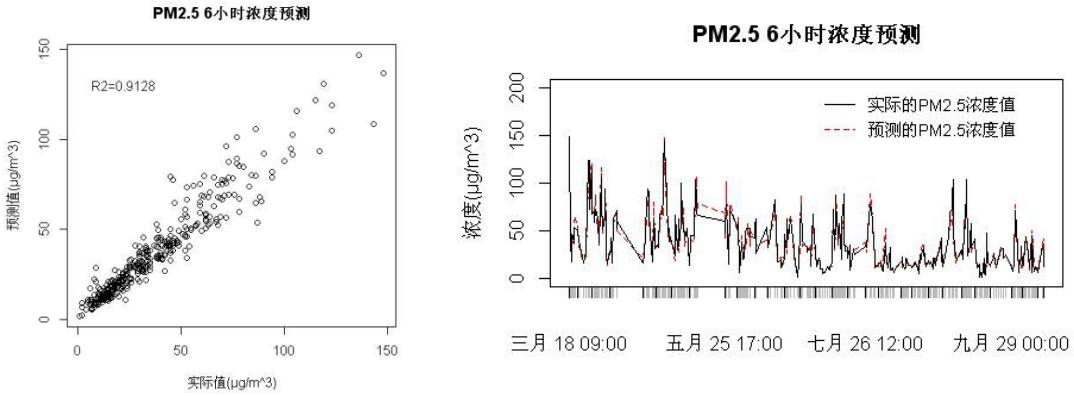

PM2.5 4小时浓度预测
实际的PM2.5浓度值
预测的PM2.5浓度值
浓度(μg/m^3)
三月 18 14:00
五月 25 18:00
七月 25 22:00
九月 26 13:00

(d)


PM2.5 5小时浓度预测
R2=0.9203
预测值(μg/m^3)
实际值(μg/m^3)
0
50
100
150


PM2.5 5小时浓度预测
实际的PM2.5浓度值
预测的PM2.5浓度值
浓度(μg/m^3)
三月 18 03:00
五月 25 19:00
七月 26 13:00
九月 26 11:00

(e)



(f)

图 5-4 1-6 逐小时预测结果与实际值

定量计算随机森林模型对未来 1~6 小时逐小时 $PM_{2.5}$ 质量浓度预测时的拟合优度、平均绝对误差、规范化后的平均绝对误差、均方误差、规范化后的均方误差等度量，计算结果见表 5-3。

本文选择使用逐步回归模型作为基准模型，用以对比随机森林模型的优劣，该模型各步骤的训练数据和检验数据与随机森林模型保持一致，基准模型的评价指标只记录拟合优度一项。

表 5-3 各模型 1-6 逐小时预测精度

	1h 精度	2h 精度	3h 精度	4h 精度	5h 精度	6h 精度
	基准模型					
R^2	0.9578	0.8850	0.8199	0.7486	0.7249	0.6279
	随机森林模型（因子筛选前）					
R^2	0.9614	0.9439	0.9025	0.9048	0.9043	0.9001
MAE	3.64	4.68	5.44	5.55	5.71	5.51
NMAE	0.1732	0.2164	0.2568	0.2580	0.2680	0.2681
MSE	29.72	47.40	78.31	79.86	70.59	68.74
NMSE	0.0386	0.0561	0.0975	0.0952	0.0957	0.0991
	随机森林模型（因子筛选后）					
R^2	0.9625	0.9469	0.9087	0.9290	0.9203	0.9128
MAE	3.61	4.40	5.16	4.90	5.10	5.36
NMAE	0.1717	0.2080	0.2436	0.2277	0.2392	0.2609
MSE	28.86	44.87	73.38	59.62	58.80	60.51
NMSE	0.0374	0.0531	0.0913	0.0710	0.0797	0.0872

根据表中数据可以得到以下结论：

（1）随着时间的推移，基准模型的准确性下降幅度较大，而随机森林模型的准确性基本能够维持在一个较高的水平，模型的稳定性较强。从数据中可以看

出，基准模型在 1~6 小时的预测时，拟合优度逐小时的降低幅度分别为 7.28%、6.51%、7.13%、2.37%、9.7%，而随机森林拟合优度逐小时的变化幅度分别为 1.56%、3.82%、-2.03%、0.87%、0.75%，随机森林拟合优度的波动幅度要远小于基准模型。从第 1 小时到第 6 小时，基准模型拟合优度下降了 32.99%，而随机森林模型仅降低了 4.97%。

（2）在每个时段，随机森林拟合优度相较于基准模型均有提高。除预测 1 小时浓度时随机森林 96.25%的拟合优度与基准模型的 95.78%相差不大外，从 2 小时开始，随机森林比基准模型在拟合优度上分别提高了 6.19%、8.88%、18.04%、19.54%、28.49%，并且可以看出，随机森林的提高幅度在不断增大。

（3）在进行对输入因子筛选的步骤后，随机森林的性能相较于筛选前会有小幅提升。因子筛选后的模型，其拟合优度平均提升 1.05%，绝对误差平均降低 0.33。

（4）未进行因子筛选的随机森林模型以及基准模型，其 1~6 小时的预测性能均呈现持续降低的趋势，而因子筛选后的随机森林模型，其性能会出现回升现象。例如，因子筛选后的随机森林模型 4 小时的拟合优度比 3 小时的提高了 2.03%。

（5）使用随机森林对 $PM_{2.5}$ 质量浓度进行逐小时预测，其拟合优度总体在 90%以上，规范化后的均方误差在 10%以下，说明该模型的效果较为理想。

5.2 $PM_{2.5}$ 污染物 6~12、12~24、24~48 小时浓度极值预测

5.2.1 数据准备

由于预测目标的不同，对于 1~6 小时的逐小时预测，一次只需要预测一个 $PM_{2.5}$ 质量浓度值，而对于 6~48 小时各时段的极值预测，每次需要预测出 $PM_{2.5}$ 质量浓度的变化范围即一个最低值和一个最高值，因此，需要准备的数据也不尽相同。

极值预测模型所需要的数据包含 5 个部分，结构见表 5-4。其中，因子 A 和因子 B 与逐小时预测模型所需的数据相同，均为当前的各项实测数据，包括当前污染物的质量浓度、离子的质量浓度、当前气压、气温、相对湿度、降水量、

风向、风速、能见度等气象情况，以及反映 $PM_{2.5}$ 质量浓度周期变化规律的数据，包括 $PM_{2.5}$ 的月平均质量浓度以及周平均质量浓度。因子 C 是待预测时段内的气象预报极值，包括气压、气温、相对湿度、降水量、风向、风速的最大值和最小值。因子 D 为逐小时模型所预测的未来 1~6 小时 $PM_{2.5}$ 质量浓度的预测值。因子 E 是极值模型的预测结果。

极值模型的训练数据和检验数据所抽取的时间节点均与逐小时预测模型保持相同。

表 5-4 极值模型的预测变量

因子 A	当前的实测数据： 污染物质量浓度（101~107）；离子质量浓度（201~208）；气象数据（301~310）；
因子 B	周期数据： $PM_{2.5}$ 月平均质量浓度（401）；$PM_{2.5}$ 周平均质量浓度（402）；
因子 C	C1：未来 6~12 小时气象预报数据极值（包括气压、气温、相对湿度、降水量、风向、风速） C2：未来 12~24 小时气象预报数据极值 C3：未来 24~48 小时气象预报数据极值
因子 D	1~6 小时的 $PM_{2.5}$ 质量浓度逐小时预测数据
因子 E	E1：未来 6~12 小时 $PM_{2.5}$ 质量浓度最低值预测 E2：未来 6~12 小时 $PM_{2.5}$ 质量浓度最高值预测 E3：未来 12~24 小时 $PM_{2.5}$ 质量浓度最低值预测 E4：未来 12~24 小时 $PM_{2.5}$ 质量浓度最高值预测 E5：未来 24~48 小时 $PM_{2.5}$ 质量浓度最低值预测 E6：未来 24~48 小时 $PM_{2.5}$ 质量浓度最高值预测

5.2.2 预测步骤

对于未来 6~12、12~24、24~48 小时的 $PM_{2.5}$ 质量浓度进行预测时，需要使用 6 个随机森林模型分别对各时段 $PM_{2.5}$ 质量浓度的最高值和最低值进行预测，同一时段的两个模型所需的训练数据集是相同的，而不同时段的训练数据是在动态变化的，其具体步骤见图 5-5：

（1）构建用于预测未来 6~12 小时后 $PM_{2.5}$ 质量浓度最低值的随机森林模型 7 和用于预测未来 6~12 小时后 $PM_{2.5}$ 质量浓度最高值的随机森林模型 8。此时将当前的各项实测数据即因子 A、周期数据即因子 B、未来 6~12 小时的气象预报

数据极值即因子 C1、1~6 小时的 $PM_{2.5}$ 质量浓度逐小时预测数据即因子 D 作为模型的输入因子。模型训练完成后，分别预测未来 6~12 小时 $PM_{2.5}$ 质量浓度最低值 E1 以及未来 6~12 小时 $PM_{2.5}$ 质量浓度最高值 E2。

（2）构建用于预测未来 12~24 小时后 $PM_{2.5}$ 质量浓度最低值的随机森林模型 9 和用于预测未来 12~24 小时后 PM2.5 质量浓度最高值的随机森林模型 10。此时的输入因子包括因子 A、因子 B、因子 D，气象预报数据换成未来 12~24 小时的气象预报数据极值即因子 C2，并将模型 7、模型 8 的预测值即因子 E1、E2 加入到训练数据中。模型训练完成后，分别预测未来 12~24 小时 $PM_{2.5}$ 质量浓度最低值 E3 以及未来 12~24 小时 $PM_{2.5}$ 质量浓度最高值 E4。

（3）构建用于预测未来 24~48 小时后 $PM_{2.5}$ 质量浓度最低值的随机森林模型 11 和用于预测未来 24~48 小时后 $PM_{2.5}$ 质量浓度最高值的随机森林模型 12。此时的输入因子包括因子 A、因子 B、因子 D，气象预报数据换成未来 24~48 小时的气象预报数据极值即因子 C3，并将模型 7、模型 8、模型 9、模型 10 的预测值即因子 E1、E2、E3、E4 加入到训练数据中。模型训练完成后，分别预测未来 24~48 小时 $PM_{2.5}$ 质量浓度最低值 E5 以及未来 12~24 小时 $PM_{2.5}$ 质量浓度最高值 E6。

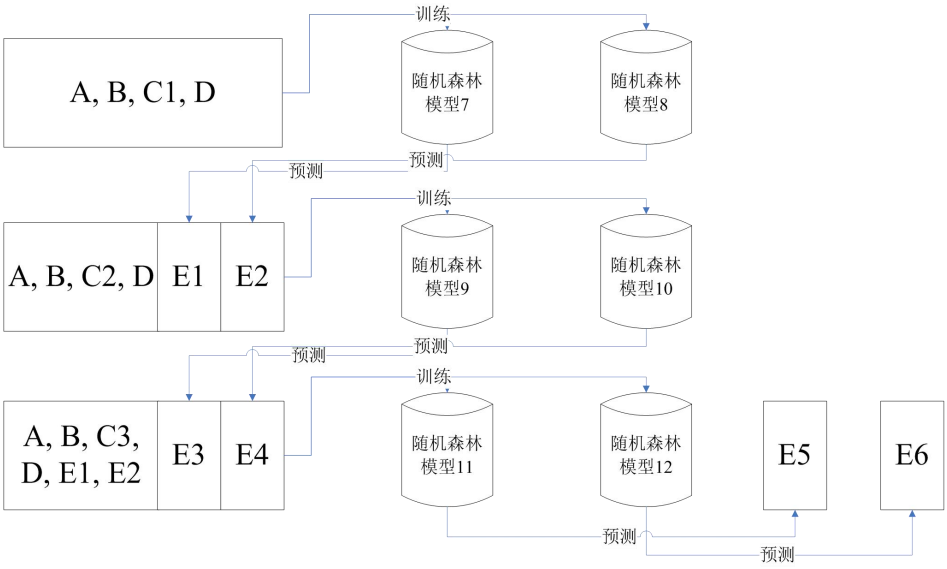


图 5-5 极值模型预测流程

5.2.3 构建模型

极值模型与逐小时模型在生成随机森林时的步骤基本相同。

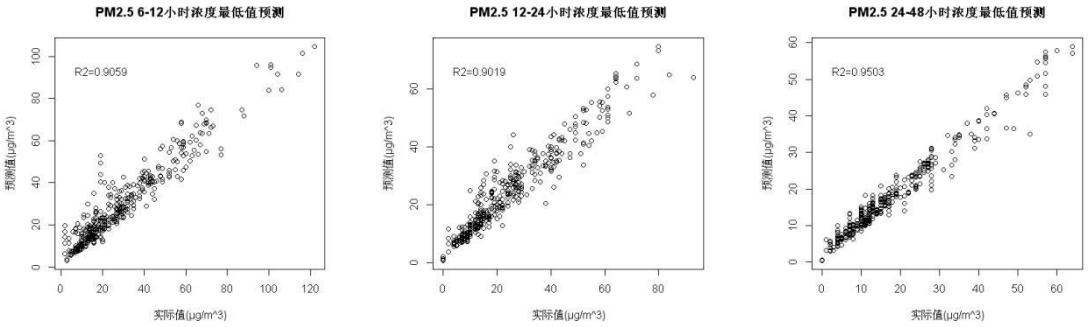
首先，采用控制参数变量以及 10 折交叉验证结果的方法，调整不同的森林规模即随机森林模型中树的棵数以及单棵树的内部节点每次分裂时从候选变量中抽取的变量数，同时观察 10 折交叉验证的结果，选择验证误差最低的一组参数。最终确定森林规模为每个随机森林模型由 100 棵树构成，由于整个候选变量集合中的因子数量在 42~46 个之间，结合交叉验证结果以及最优抽取数推荐，即单次分裂所抽取的属性数取 $sqrt(p)$ 个左右，确定随机森林模型中单棵树内部节点分裂时随机抽取的属性个数为 7。

其次，利用随机森林的特性，结合因子被替换后的 OOB 误差变化，确定各输入因子的重要性。具体标准是，因子被替换后模型准确率下降越多，说明该因子越重要。选择重要性较高且满足误差标准的因子组合。

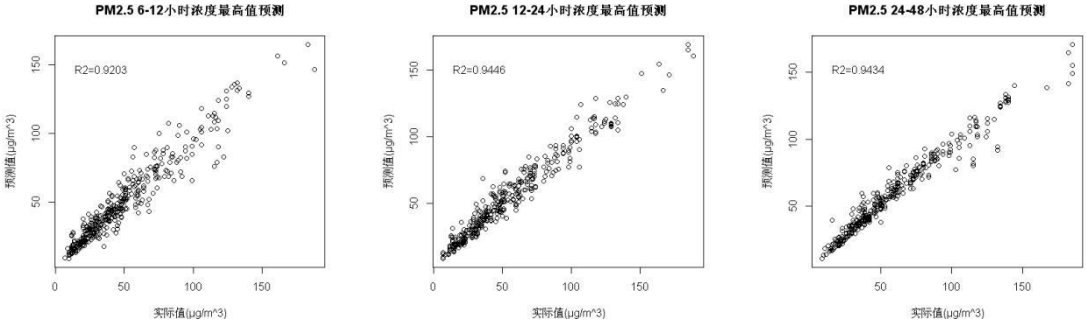
最后，输出模型预测结果，并与实际值对比，分析模型性能。

5.2.4 结果分析

根据构建的未来 6~12 小时、12~24 小时、24~48 小时 $PM_{2.5}$ 质量浓度预测模型，分别输出各时段 $PM_{2.5}$ 质量浓度最高值和最低值的预测值，并绘制时段内实际值与模型预测值的散点图和折线图。散点图见图 5-6，横坐标是实际值，纵坐标是预测值，图 5-6（a）（b）（c）是对各时段最低值的预测，图 5-6（d）（e）（f）是对各时段最高值的预测，折线图见图 5-7。

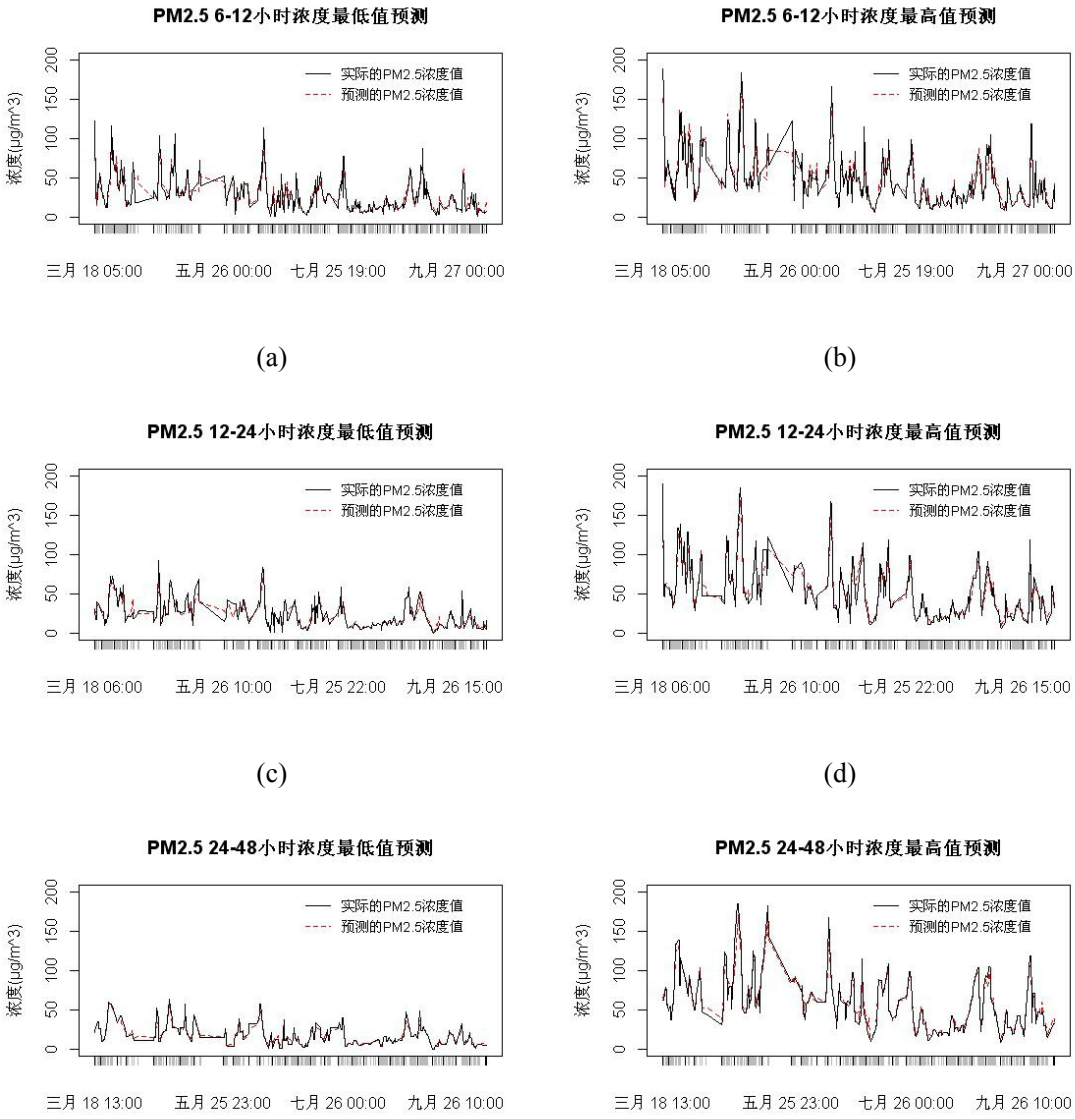


(a) (b) (c)



(d) (e) (f)

图 5-6 极值模型预测结果散点图



(a) (b)

(c) (d)

(e) (f)

图 5-7 极值模型预测结果折线图

从图 5-6 和图 5-7 中可以看出，不论是预测 $PM_{2.5}$ 质量浓度的最低值或是最

高值，随着时间的向后推移加上预测时段范围的扩大，模型性能的趋势是拟合程度越来越高，未来 6~48 小时各时间尺度下的总体拟合效果较好。

为了验证极值模型在未来 6~48 小时各时段的性能，分别计算各模型预测值的拟合优度（R^2）、平均绝对误差（MAE）、标准化后的平均绝对误差（NMAE）、均方误差（MSE）、标准化后的均方误差（NMSE），计算结果见表 5-5。

表 5-5 极值模型预测精度

	R^2	MAE	NMAE	MSE	NMSE
6~12h 最低值精度	0.9059	4.49	0.2675	45.02	0.0941
6~12h 最高值精度	0.9203	6.25	0.2540	82.62	0.0797
12~24h 最低值精度	0.9019	3.69	0.2669	29.68	0.0981
12~24h 最高值精度	0.9446	6.05	0.2170	69.92	0.0554
24~48h 最低值精度	0.9503	2.03	0.2031	8.65	0.0497
24~48h 最高值精度	0.9434	5.47	0.1935	72.42	0.0566

根据表 5-5 中数据可以得到以下结论：

（1）随着时间的推移，最低值和最高值的拟合精度都呈现出逐步上升的趋势。在逐小时的随机森林预测模型中，模型性能是随着时间的推移而呈现一个降低趋势的，而对于极值模型，从 6~12 小时到 24~48 小时，各时段最低值的拟合优度从 90.59%上升到 95.03%，最高值的拟合优度从 92.03%上升到 94.34%，最低值标准化后的均方误差从 9.41%降低至 4.87%，最高值标准化后的均方误差从 7.97%降低至 5.66%，这说明模型的性能越来越好、拟合精度越来越高。

（2）在评价模型性能时，并不能简单地比较绝对误差。可以看出，在预测未来 6~12 小时的 $PM_{2.5}$ 质量浓度时，最低值的预测值绝对误差是 4.49，均方误差是 45.02，最高值的预测值绝对误差是 6.25，均方误差是 82.62。如果单从绝对误差和均方误差上看，预测此时段最低值的精度要高于最高值。但从拟合优度、标准化后的平均绝对误差、标准化后的均方误差上看，最低值的预测值拟合优度是 90.59%，规范化后的平均绝对误差是 26.75%，规范化后的均方误差是 9.41%，最高值的预测值拟合优度是 92.03%，规范化后的平均绝对误差是 25.4%，规范化后的均方误差是 7.79%，预测此时段最低值的精度要低于最高值。造成这种现象的原因是，在同一时段，最高值的实际值要高于最低值的实际值，在相同拟合程度的情况下，预测最低值时模型的绝对误差和均方误差低于预测最高值时模型

的绝对误差和均方误差是一定的。12~24 小时、24~48 小时的预测结果也基本表现出相同的规律，因此，本文选择多维度的评价标准有其合理性。

（3）总体上看随机森林模型在各时段预测 $PM_{2.5}$ 质量浓度最高值和最低值时的表现，拟合优度均在 90%以上，规范化后的均方误差均在 10%以下，说明该模型的效果较为理想。

第六章 结论与展望

6.1 结论

$PM_{2.5}$是目前最受关注的城市空气污染源之一，对其在更小时间尺度下的精准预测具有重要的科学意义和现实意义。

本文在相关和回归分析的基础上，采用随机森林算法对2016年上海市徐汇区3月到10月的$PM_{2.5}$质量浓度建立预测模型，对未来1~6小时的$PM_{2.5}$逐小时质量浓度以及6~12小时、12~24小时、24~48小时$PM_{2.5}$质量浓度的最大值和最小值进行预测。

本文的主要结论如下：

（1）$PM_{2.5}$与不同因子之间有明显的相关倾向。

本文采用Pearson相关系数，衡量了$PM_{2.5}$与不同因子之间的相关性。其中，$PM_{2.5}$与其他污染物因子之间整体呈现正相关，平均相关系数为0.47；$PM_{2.5}$与气象因子之间整体呈现负相关，平均相关系数为-0.08。虽然$PM_{2.5}$与气象因子的相关性并不强，但加入气象等多因子后的逐步回归仿真方程，其拟合优度从66%提高到85%，说明改进输入因子对提高模型准确性是必要的。

（2）模型的输入因子存在最优子集。

本文通过OOB误差估计的方法，衡量了不同模型中输入因子的重要性程度，并根据OOB误差筛选出重要程度高且因子数量少的输入因子组合。新的因子组合使模型的拟合优度均得到提升，提升幅度在0.11%~2.42%之间，平均提升1.05%。该方法为因子筛选提供了理论依据，避免了因子选择过程中的盲目性和主观性。

（3）随机森林算法可以满足$PM_{2.5}$质量浓度小时预测的要求。

本文采用随机森林算法对未来48小时各时间粒度下的$PM_{2.5}$质量浓度进行预测，1~6小时的逐小时预测精度在90.87%~96.25%之间，6~12小时、12~24小时、24~48小时最大值和最小值预测精度在90.19%~95.03%之间。对比逐步回归方法建立的基准预测模型，预测精度最大提升30%左右。

6.2 展望

本文在对 $PM_{2.5}$ 质量浓度进行建模预测方面还存在一些不足，需要进一步讨论和分析，有待继续研究和完善的工作有以下几方面：

（1）对长时间跨度下的 $PM_{2.5}$ 质量浓度进行分析。

本文获取的数据主要集中在 2016 年 3 月至 10 月，数据时间跨度覆盖较短，使得难以对 $PM_{2.5}$ 质量浓度较高的秋冬季节进行完整的分析。在接下来的研究中，一方面，对全年时段的 $PM_{2.5}$ 质量浓度进行分析，总结其在不同季节、冷暖月等时间范围上呈现的规律，并将其加入到预测模型的周期因子中；另一方面，比较不同年份 $PM_{2.5}$ 质量浓度的变化情况，探索其在更大时间尺度上的演变规律。

（2）随机森林算法建立的预测模型更有利于解释非线性规律。

随机森林作为一个非线性的分类模型，理论上，其分类边界对特征空间是非线性切分的，因此，数据在线性或时间序列上的方差需要进一步解释。在接下来的研究中，考虑使用线性模型或时间序列模型对随机森林模型预测的残差进行进一步的拟合，以此校正随机森林模型的预测结果。

（3）$PM_{2.5}$ 质量浓度的预测流程可以系统化。

本文的数据加载、预处理、分析、建模等流程相互独立，在接下来的工作中，可以将数据、建模、展示等模块之间的壁垒打通，使该预测功能做到“一键式”展示，更有利于预测模型的应用。同时，考虑更丰富的可视化展现。

参考文献

[1] Anantha M, Louis R, Andy L. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction[J]. Ecosystems, 2006, 9: 181-199.

[2] Anush S, Aayush J, Tarun V, et al. Adaptive latent fingerprint segmentation using feature selection and random decision forest classification[J]. Information Fusion, 2017, 34: 1-15.

[3] Bart L, Dirk V. Predicting customer retention and profitability by using random forests and regression forests techniques[J]. Expert Systems with Applications, 2005, 29: 472-484.

[4] Bench G. Measurement of contemporary and fossil carbon contents of $PM_{2.5}$ aerosols: results from Turtleback Dome, Yosemite National Park[J]. Environmental Science & Technology, 2004, 38: 2424-2427.

[5] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[6] Bun T, Komei S, Koji Z. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting $PM_{2.5}$[J]. Neural Computing & Applications, 2016, 27: 1553-1566.

[7] Carolin S, Anne L, Thomas K, et al. Conditional variable importance for random forests[J]. BMC Bioinformatics, 2008, 9: 307.

[8] Carolin S, Anne L, Achin Z, et al. Bias in random forest variable importance measures: Illustrations, sources and a solution[J]. BMC Bioinformatics, 2007, 8: 25.

[9] Gaspar C, Jose G, Alberto G, et al. Automatic selection of molecular descriptors using random forest: Appication to drug discovery[J]. Expert Systems With Application, 2017, 72: 151-159.

[10] Gholamreza A, Hossein Z, Shiva H. Predicting $PM_{2.5}$ Concentrations Using Artificial Neural Networks and Markov Chain, a Case Study Karaj City[J].

Asian Journal of Atmospheric Environment, 2016, 10: 67-79.

[11] Haywood J, Boucher O. Estimates of the direct radiative forcing due to tro pospheric aerosols: a review[J]. Reviews of Geophysics, 2000, 38: 513-543.

[12] Jian L, Zhao Y, Zhu Y, et al. An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roa dside in Hangzhou, China[J]. Science of the Total Environment, 2012, 426: 336-345.

[13] Joseph S, Petter L. Standard errors for bagged and random forest estimator s[J]. Computational Statistics and Data Analysis, 2009, 53: 801-811.

[14] Li G, Nesrine C, Clement M, et al. Relevance of airborne lidar and multis pectral image data for urban scene classification using Random Forests[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2011, 66: 56-66.

[15] Liu Y, Tang S, Carlos F, et al. Experimental studay and Random Forest pr ediction model of microbiome cell surface hydrophbicity[J]. Expert Systems With Applications, 2017, 72: 306-316.

[16] Maher E, Nor A, Noor F. Development and comparison of regression mod els and feedforward backpropagation neural network models to predict seas onal indoor $PM_{2.5-10}$ and $PM_{2.5}$ concentrations in naturally ventilated schools [J]. Atmospheric Pollution Research, 2015, 6: 1013-1023.

[17] Onisimo M, Elhadi A, Moses A. High density biomass estimation for wetla nd vegetation using WorldView-2 imagery and random forest regression alg orithm[J]. International Journal of Applied Earth Observation and Geoinfor mation, 2012, 18: 399-406.

[18] Quan J, Zhang X, Zhang Q, et al. Importance of sulfate emission to sulfur deposition at urban and rural sites in China[J]. Atmospheric Research, 200 8, 89: 283-288.

[19] Ramon D, Sara A. Gene selection and classification of microarray data usi

ng random forest[J]. BMC Bioinformatics, 2006, 7:3.

[20] Rodriguez G, Ghimire B, Rogan J, et al. An assessment of the effectivenes s of a random forest classifier for land-cover classification[J]. ISPRS Journ al of Photogrammetry and Remote Sensing, 2012, 67: 93-104.

[21] Seinfeld J, Pandis S. Atmospheric Chemistry and Physics-from Air Pollutio n to Climate Change[M]. John Wiley and Sons, 1998: 1326.

[22] Shadi A, Jamil A. Assessing the Accuracy of ANFIS, EEMD-GRNN, PCR and MLR models in predicting $PM_{2.5}$[J]. Atmospheric Environment, 2016, 1 42: 465-474.

[23] Sun W, Zhang H, Palazoglu A, et al. Prediction of 24-hour-average $PM_{2.5}$ concentration using a hidden Markov model with different emission distribu tions in Northern California[J]. Science of the Total Environment, 2013, 44 3: 93-103.

[24] Tzanis C, Varotsos C, Christodoulakis J, et al. On the corrosion and soilin g effects on materials by air pollution in Athens, Greece[J]. Atmospheric C hemistry and Physics, 2011, 11: 12039-12048.

[25] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forest: A survey and results of new tests[J]. Pattern Recognition, 2011, 44: 330-349.

[26] Vogels M, De J, Sterk G, et al. Agricultural cropland maping using black-and-white aerial photography, Object-Based Image Analysis and Random Fo rests[J]. International Journal of Applied Earth Observation and Geoinformat ion, 2017, 54: 114-123.

[27] Wang T, Li S, Jiang F, et al. Investigations of main factors affecting tropo spheric nitrate aerosol using a coupling model[J]. China Particuology, 2006, 6: 336-341.

[28] Xie Y, Li X, Ngai E, et al. Customer churn prediction using improved bal anced random forests[J]. Expert System with Applications, 2009, 36: 5445-

5449.

[29] Zoue B, Wilson J, Zhan G, et al. Air pollution exposure assessment methods utilized in epidemiological studies[J]. Journal of Environmental Monitoring, 2009, 11: 475-490.

[30] 崔寒, 庄毅斌, 曹茜, 等. 人工神经网络与逐步回归法对大雾预报对比[J]. 环境科学与技术, 2015, 38(12): 404-407.

[31] 方匡南, 吴见彬, 朱建平, 等. 随机森林研究方法综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.

[32] 付倩娆. 基于多元线性回归的雾霾预测方法研究[J]. 计算机科学, 2016, 43(6): 526-528.

[33] 郭新彪, 魏红英. 大气 $PM_{2.5}$ 对健康影响的研究进展[J]. 科学通报, 2013, 58(12): 1171-1177.

[34] Han J W, Micheline K, Pei J. 数据挖掘：概念与技术[M]. 北京: 机械工业出版社, 2015.

[35] 林成德, 彭国兰. 随机森林在企业信用评估指标体系确定中的应用[J]. 厦门大学学报(自然科学版), 2007, 46(2): 199-203.

[36] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[37] 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报, 2013, 50(4): 1190-1197.

[38] 李毓, 张春霞. 基于 out-of-bag 样本的随机森林算法的超参数估计[J]. 系统工程学报, 2011, 26(4): 566-572.

[39] 孙云海, 张财涛, 杨洪斌. 基于逐步回归分析方法的 PM_{10} 浓度预测模型[J]. 环境科学与技术, 2009, 32(5): 100-102.

[40] 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版), 2014, 44(1): 137-141.

[41] 尹华, 胡玉平. 基于随机森林的不平衡特征选择算法[J]. 中山大学学报(自然科学版), 2014, 53(5): 59-65.

[42]余辉, 袁晶, 于旭耀, 等. 基于 ARMAX 的 $PM_{2.5}$ 小时浓度跟踪预测模型[J]. 天津大学学报(自然科学与工程技术版), 2017, 50(1): 105-111.

[43]张华伟, 王明文, 甘丽新. 基于随机森林的文本分类模型研究[J]. 山东大学学报(理学版), 2006, 41(3): 139-143.

[44]甄亿位, 郝敏, 陆宝宏, 等. 基于随机森林的中长期降水量预测模型研究[J]. 水电能源科学, 2015, 33(6): 6-10.

致谢

时光荏苒，三年的硕士研究生生活即将结束，回想在师大的日日夜夜、点点滴滴，内心充满了留恋与不舍。在这即将毕业的时刻，有太多的感谢。

首先，感谢我的导师过仲阳老师三年来对我的亲切关怀和悉心指导。过老师平易近人的性格、严谨的治学态度、认真的作风、渊博的学识，使我受益良多。在过老师的教导和鼓励下，我摆脱了刚入校时的懵懂，成长为一名对专业知识有一定见解并且能够学以致用的合格研究生。三年来，过老师为我提供了许多把学术价值转化为社会价值的机会，不仅锻炼了我理论结合实际的能力，也切实提升了在校期间的生活品质。能遇到一位在学习和生活中都对我帮助巨大的老师是我的幸运，在此对过老师表达最诚挚的谢意。

同样，感谢 119 实验室的各位小伙伴，感谢王媛媛、王细元、曹琼珊、瞿丽、纵清华师姐，感谢马品、展洪强、姚艳豪师兄，感谢闫密巧、陈亦辉、郑旭曼、许晓宁、常恬君、王志宇。因为有你们，面对挑战时有了互相支持、共同前进的伙伴，因为有你们，实验室的氛围永远是那样轻松愉快、温馨和谐，也是因为你们，使我对校园更加怀念。希望我们的 119 越来越好。

感谢家人的支持与付出、理解与照顾，为我提供了坚实的后盾，使我能够在求学期间心无旁骛。

最后，还要感谢那个努力的自己，感谢迷茫时把我拉回来共同奋斗 365 天的老范和小爽，感谢那些相信我的人。师大将成为我的烙印，希望自己多年后依然：

O ever youthful, O ever weeping.

王雨晨

丁酉年夏于樱桃河畔