2018 届研究生硕士学位论文

分类号：_____ 学校代码：10269

密　级：_____ 学　　号：51153901101

華東師範大學

East China Normal University

硕士学位论文

MASTER'S　DISSERTATION

基于随机森林的长江三角洲 $PM_{2.5}$ 浓度空间模拟及暴露风险评估

院　　系： 地理科学学院

专　　业： 地图学与地理信息系统

研究方向： 地理大数据与空间数据挖掘

指导教师： 徐建华 教授

学位申请人： 赵佳楠

2018 年 4 月

Dissertation for masteral degree in 2018 University code:10269

Student ID: 51153901101

East China Normal University

MASTER'S DISSERTATION

Spatial simulation and exposure risk assessment of $PM_{2.5}$ concentration in the Yangtze river delta based on random forest model

Department: School of Geographic Sciences

Major: Cartography and Geographic Information System

Research direction: Spatial Data Dining Geography

Supervisor: Professor Jianhua Xu

Candidate: Jianan Zhao

April，2018

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《基于随机森林优化模型的长江三角洲 $PM_{2.5}$ 浓度空间模拟》，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：赵佳楠　　日期：2018年5月29日

华东师范大学学位论文著作权使用声明

《基于随机森林优化模型的长江三角洲 $PM_{2.5}$ 浓度空间模拟》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，本论文的著作权归本人所有。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，并向主管部门和学校指定的相关机构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）

（　）1.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。

（√）2.不保密，适用上述授权。

导师签名　　本人签名　赵佳楠

2018年5月29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）。

赵佳楠 硕士学位论文答辩委员会成员名单

姓名	职称	单位	备注
叶超	教授	华东师范大学	主席
陈睿山	教授	华东师范大学	
李治洪	副教授	华东师范大学	

摘要

土地利用回归（LUR）模型是目前模拟 $PM_{2.5}$ 分布的常用方法，但该方法以多元线性回归进行建模，未考虑自变量与 $PM_{2.5}$ 浓度的非线性的复杂关系，常引起多重共线性。为提高模拟的准确性，本文采用随机森林算法训练模型，进行长江三角洲地区 $PM_{2.5}$ 浓度的空间分布的模拟。并根据 $PM_{2.5}$ 浓度空间分布的模拟结果，运用 $PM_{2.5}$ 人口暴露风险评估方法，对长江三角洲地区 $PM_{2.5}$ 人口暴露风险进行评估。研究表明：

（1）本文采用随机森林方法改进的土地利用回归模型（下文简称随机森林优化模型）来对长江三角洲地区 $PM_{2.5}$ 浓度空间分布进行模拟。模拟的效果采用相关系数（R）、均方根误差（Root Mean Squared Error，RMSE）、平均绝对误差（Mean Absolute Error，MAE）以及拟合指数（Index of Agreement，IA）来评价。其中 R 和 IA 越大，RMSE 和 MAE 越小，模型模拟的效果就相对较好。经检验集检验模型效果，该模型的 IA、MAE、RMSE、R 分别为 0.854、4.757、5.871、0.831。为了说明该模型的高效性及准确性，本文将该模型与传统的土地利用回归(LUR)模型及 SVM 改进的土地利用回归模型进行比较。使用 LUR 模型，检验集的 IA、MAE、RMSE、R 分别为 0.702 、5.862、7.58、0.647。经计算，使用 SVM 改进的土地利用回归模型(SVM 优化模型)，其检验集的 IA、MAE、RMSE、R 分别为 0.825、5.521、6.871、0.714。相比之下，随机森林在进行 $PM_{2.5}$ 浓度空间模拟时，检验集的 Ia 和 R 较大，而 MAE 和 RMSE 较小，因此效果更好。

（2）2015 年长江三角洲地区 $PM_{2.5}$ 年平均浓度呈现北高南低，西高东低的总体格局，部分地区高值集聚，具有连片分布的特点；江苏省各市的 $PM_{2.5}$ 浓度明显比浙江省和上海市更高，尤其是泰州、无锡、扬州、常州等苏南、苏中地区;浙江省内的 $PM_{2.5}$ 浓度较高的地市为湖州、绍兴、嘉兴，状况最好的则为舟山和台州；杭州和宁波虽然全市的平均值不高，但也存在明显的高值区。

（3）基于 $PM_{2.5}$ 空气质量浓度及人口分布下的人口 $PM_{2.5}$ 暴露风险，长江三角洲地区人口 $PM_{2.5}$ 暴露风险从南往北，从东往西呈现梯度递增现象，各城市中心城区是人口 $PM_{2.5}$ 暴露风险高值区。从市域分析，以杭州市为例，人口 $PM_{2.5}$ 暴露风险大致具有同心圆结构，中心往外减弱，且西南区域的风险要比东北区域弱；中心城区呈现连片的高风险区，在中心城区外围的副城具有斑块状的高风险

区；城市内部具有轴状的相对高风险区，如主要交通路线及钱塘江沿线带。而基于人口加权平均的 $PM_{2.5}$ 暴露风险可评估单元间空气污染对公众健康的影响强度。泰州、无锡、常州的风险值最大，应加大对泰州、无锡、常州等市的 $PM_{2.5}$ 暴露风险的防控，此外杭州、宁波、绍兴、上海等地的人口高度集中于 $PM_{2.5}$ 浓度高值区，需重点关注。

（4）利用 $PM_{2.5}$ 空气质量浓度、人口暴露强度指标，可对长江三角洲地区的 $PM_{2.5}$ 人口暴露风险进行初步的空间格局分析；人口加权平均的 $PM_{2.5}$ 人口暴露风险评估可对各行政单元的 $PM_{2.5}$ 人口暴露风险进行对比，筛选 $PM_{2.5}$ 人口暴露风险重点防控城市；利用人口暴露强度可进一步监测重点防控城市的重点区域，适合于城市内部的区域 $PM_{2.5}$ 暴露风险评估。综合以上指标进行 $PM_{2.5}$ 人口暴露风险评估，对于大气污染的治理以及人居环境的改善具有实际价值。

总体而言，本文具有以下研究特色：采用多元数据，将 AOD 数据与常用的土地利用回归数据结合；在传统的 LUR 模型中融入了随机森林算法，在方法上具有一定的创新；对于 $PM_{2.5}$ 人口暴露风险，结合已有的多种评价方式进行多角度分析，更具科学性。

关键词： $PM_{2.5}$；随机森林；空间分布模拟；暴露风险；长江三角洲

Abstract

Land use regression (LUR) model is the commonly used method to simulate the distribution of $PM_{2.5}$, but this method takes the multivariate linear regression modeling, not taking into account the complex relations between variables and $PM_{2.5}$ concentrations of nonlinear, and easy to appear multicollinearity. In order to improve the accuracy of the simulation, this paper uses the random forest algorithm training model, the Yangtze river delta to simulate the spatial distribution of $PM_{2.5}$. And Referring to the existing $PM_{2.5}$ exposure risk assessment methods and the previously obtained $PM_{2.5}$ concentration spatial distribution data, we assess the exposure risk of $PM_{2.5}$ in the Yangtze river delta . Research shows that:

(1)In this paper, we use the method of random forest improved regression model of land use model. To evaluate the simulation results , we adopt the correlation coefficient (R), Root Mean square Error (RMSE), Mean Absolute Error (MAE) and Index of Agreement (IA). The larger R and IA, the better effect of the model; the smaller RMSE and MAE, the better effect of the model. Testing the effect of this model, the IA, MAE, RMSE and R of this model are 0.854, 4.757, 5.871, 0.831 by the test sets. To illustrate the efficiency and accuracy of the model, this paper compared the model we adopted with multivariate regression of traditional land use regression model and SVM improved land use regression model . Using multivariate regression of traditional land use regression model in common use, IA, MAE, RMSE and R of test sets are 0.702, 5.862, 7.58, 0.647, respectively.Using the method of SVM improved land use regression model test set of the model calculation of IA, MAE, RMSE and R are 0.825, 5.521, 6.871, 0.825. By contrast, due to the larger test set IA and R, and the smaller MAE and RMSE, the random forest algorithm in $PM_{2.5}$ concentration space simulation effect is better.

(2)Average 2015 $PM_{2.5}$ concentrations of the Yangtze river delta shows the trend that lower from north to south, east to west;Cities in Jiangsu Province of $PM_{2.5}$ concentration significantly higher than in Zhejiang Province and Shanghai, especially in Taizhou, Wuxi, Yangzhou, Changzhou and other Southern Jiangsu, central

Jiangsu;The highest concentrations of $PM_{2.5}$ cities in Zhejiang Province are Huzhou,Shaoxing, Jiaxing, the best is in zhoushan and Taizhou;Hangzhou and Ningbo while the city's average value is not high, but also exist obvious high value area.

(3)Based on the concentration of $PM_{2.5}$ air quality and $PM_{2.5}$ exposure risk population distribution of the population, the population $PM_{2.5}$ exposure risk of Yangtze river delta region from south to north, from east to west to present the phenomenon of increasing gradient.from.The downtown area is high value area population $PM_{2.5}$ exposure risks.Taking Hangzhou as an example, the population of $PM_{2.5}$ exposure risk is roughly with concentric circles structure. Center to the edge, the population of $PM_{2.5}$ exposure risk is weaker.weaker risk is in southwest area.Present continuous high risk area, downtown in the city center of vice city with patches of high risk area.Inside the city has some axis of relatively high risk area, such as belt along the main traffic routes and Qiantang River. While based on weighted average $PM_{2.5}$ exposure population risk, we can evaluate the effects of air pollution on public health between the units.Taizhou, Wuxi and Changzhou as the biggest risk value, should be given more to prevention and control, in addition, Hangzhou, Ningbo, Shaoxing, Shanghai and other places of the population is highly concentrated in $PM_{2.5}$ concentrations high value area, also need to focus on.

(4)Concentration of $PM_{2.5}$ air quality, utilizing the exposure intensity, the population of the Yangtze river delta region exposure risk can be a preliminary analysis of spatial pattern;By the weighted average of $PM_{2.5}$ population exposed to the risk assessment for each administrative unit of $PM_{2.5}$, comparing the population exposure risk screening of $PM_{2.5}$ population exposure risk focus on prevention and control;Using the population exposure intensity can be further monitoring key prevention and control the focus of the urban area, suitable for urban area within the exposure risk assessment.Comprehensive above index for atmospheric pollution control and the improvement of living environment has important practical significance.

In general, this paper has the following research features: the simulation accuracy is higher by combining AOD data with common land use regression data ; In

the traditional LUR model, the random forest algorithm is incorporated into the method, and it has some innovation in the method. It is more scientific to analyze the risk of $PM_{2.5}$ population exposure by combining the existing multiple evaluation methods.

Key words: $PM_{2.5}$; Random Forest; Spatial distribution simulation; risk exposure; Yangtze River delta

目录

图目录

表目录

第一章 绪论

1.1 研究背景与意义

由于工业化及城市化进程的加快，$PM_{2.5}$ 问题已经成为我国日益严峻的环境问题。它对公众健康造成了严重的危害。$PM_{2.5}$ 问题存在四大高发地带，长江三角洲地区是其中的一个高发地。2015 年以来，在我国长江三角洲地区已经检测到多次严重的大气雾霾问题，这对公众健康构成严重威胁。缓解 $PM_{2.5}$ 问题首要的是了解 $PM_{2.5}$ 空间分布特征。$PM_{2.5}$ 空间分布特征的研究对大气污染防治和公众 $PM_{2.5}$ 污染暴露风险评估十分关键，已经成为环境研究领域的热点和重点。目前常规的 $PM_{2.5}$ 研究（包括空间分布研究）的数据来源主要是地面监测，但由于仪器设备、人力以及资金的限制，监测点往往较为分散，仅能够从中获取到有限区域内的空气污染情况，无法获取到区域内连续一致的 PM2.5 空间分布格局（汉瑞英等，2016）。空气污染暴露风险评估需要连续一致的 PM2.5 空间分布数据，因此单纯依靠地面监测的手段很难对其提供有效的指导(Yang et al.，2017)。目前较多使用的方法包括了空间插值、基于卫星遥感的 PM2.5 模拟、大气化学传输模拟以及基于统计模型的 $PM_{2.5}$ 模拟（Robert et al.，2007；王敏等，2013；Zou et al.，2009；Hock et al.，2008）。基于统计模型的 $PM_{2.5}$ 模拟，考虑的要素比较全面、对数据的要求不高、模拟的空间分辨率和精度高，因此应用十分广泛（Zou et al.，2009；Hock et al.，2008）。目前 $PM_{2.5}$ 问题较为严峻的背景下，依靠 $PM_{2.5}$ 监测网络布局的逐步加强以及地理数据共享性和完备性的改善，使用随机森林算法等新型统计模型进行 $PM_{2.5}$ 空间分布模拟，可为 $PM_{2.5}$ 污染时空特征分析、健康效应分析等提供坚实的基础支撑，值得做进一步的深入研究(罗艳青，2014)。

1.2 国内外研究进展

$PM_{2.5}$ 浓度空间模拟作为环境问题研究中的基础性工作，近年来越来越受到相关科研工作者的关注。在改进 $PM_{2.5}$ 浓度空间模拟的效果方面，学者们已取得了丰富的成果，并将其应用于实践当中。研究的主要问题集中在模拟算法的分析与讨论、模拟流程的优化、模拟结果在 $PM_{2.5}$ 暴露风险评估中的应用等。

1.2.1 $PM_{2.5}$ 浓度空间模拟方法研究进展

为获取连续一致的 $PM_{2.5}$ 浓度空间分布格局,目前有空间插值、基于卫星遥感

的 $PM_{2.5}$ 模拟、大气化学传输模拟以及基于统计模型的 $PM_{2.5}$ 模拟三种方式。

（一） 空间插值

空间插值是指根据监测站点的 $PM_{2.5}$ 值，构造拟合函数，估计其他未知点的 $PM_{2.5}$ 值的方法(罗艳青，2014)。该方法具有原理易懂、运算简单、使用范围较广等优势，但是一般需要足够数量的观测点。如果在 $PM_{2.5}$ 观测数据数量不足的区域，预测的效果会不理想。此外这种方式考虑的要素相对简单，容易放大 $PM_{2.5}$ 极端浓度值的变化（Wilson et al.，2006）。

（二） 基于卫星遥感的 $PM_{2.5}$ 模拟

和地面监测的方法比较，卫星遥感的手段所获取的数据具有覆盖范围大，存在动态性的特点。这使得它能够弥补地面监测站点数量上的不足，对于拥有较少地面监测站点的区域，这种方法具有全方位、立体化的优势，具有较好的模拟效果。它可以为大气污染监测提供关键的数据基础(罗艳青，2014)。采用卫星遥感计算 $PM_{2.5}$ 浓度需依靠气溶胶光学厚度（Aerosol Optical Depth， AOD）。林海峰等通过对北京市近地层的 $PM_{2.5}$ 浓度与 AOD 进行相关性计算，结果表明 $PM_{2.5}$ 浓度与 AOD 显著相关（林海峰等，2013）。因此可根据 AOD 数据来计算地表颗粒物浓度。一般依据 AOD 数据与 $PM_{2.5}$ 的相关情况，便可用遥感方式来模拟 $PM_{2.5}$ 空间分布以及时间变化。根据相关研究整理，建立相关关系的方法主要有两种：方法 1 为直接利用 AOD 数据和近地面的 $PM_{2.5}$ 计算相关关系，方法 2 是对 AOD 数据进行湿度和垂直分布等修正，然后再与近地面的 $PM_{2.5}$ 建立相关关系（林海峰等，2013）。从理论上，在模型中引入气溶胶垂直分布和相对湿度的影响可以改善 AOD 与近地面 PM2.5 的相关性(罗艳青，2014)，但这两个因素的时空分异太大，在区域尺度上对这两个因素进行处理会导致结果具有空间局限性。所以通过方法 2 提升区域尺度的估算精度在实现上存在较大困难，在区域尺度的模拟上可直接使用方法 1。基于遥感观测的模拟方法还受遥感观测的基本特征、AOD 反演算法、回归模型计算误差、云覆盖、颗粒物化学组成等因素影响， AOD 与地面 PM2.5 浓度之间的相关关系在不同地区以及不同时节存在明显的差异，在区域性的推广上难度很大（丁冰等，2016）。AOD 数据的可获取性也存在较大局限，它只能在晴空无云的天气获得完整数据，这也限制了利用 AOD 数据来估算地面 $PM_{2.5}$。

（三） 大气化学传输模拟

针对地面监测或遥感方法模拟区域 $PM_{2.5}$ 浓度空间分布时存在的缺陷，许多学者试图建立模型来弥补上述两种方法的不足。一般的，$PM_{2.5}$ 浓度模拟的模型可分为两大类：大气扩散传输模型(Chemical Transport Models，CTMs)和统计模型。大气扩散传输通过大气物理传输、化学反应过程的模拟研究，并将结果与地面、航天等观测值进行对比和验证，来分析污染物的来源，从而完成对污染物浓度分布的模拟。化学传输模型比较复杂，包含了不同的化学机制、化学动力学表达式、反应速度系数、化学物种数和气相反应等内容（丁冰等，2016）。典型的 CTMs 模型有：CMAQ（许建明等，2005；付维雅等，2010；王丽涛等，2012；陈训来等，2008），GEOS-Chem（杨艳，2010；漏嗣佳等，2010），LOTOS-EUROS（Manders et al.，2009），MOZART（Pfister et al.，2008）等。李世广等（2013）使用 CMAQ 模型完成了成渝经济区 $PM_{2.5}$ 浓度的空间预测，但受污染物源清单的影响，与实际结果的偏差较大。此外该方法还受到应用尺度的制约。因此大气化学传输模型虽然比较复杂，且需要大量计算，但并不够准确。Appel、 Mandes 等人都发现：CMAQ 模型、LOTOS-EUROS 模型等受众多独立的因子影响，将这些因素叠加到一起计算就会产生累计误差（Appel et al.，2008; Mandes et al.，2009）。所以大气化学传输模型在局地、短期数字模拟上有优势，但不适合长期的模拟。

（四） 基于统计模型的 $PM_{2.5}$ 模拟

统计模型在 $PM_{2.5}$ 浓度模拟中比大气化学传输模型更具优势，应用更为广泛。随着统计模型在 $PM_{2.5}$ 浓度模拟研究中的深入，其朝着参数选择科学化、算法运用多元化的方向发展。

（1） 传统的 LUR 模型进行 $PM_{2.5}$ 浓度空间模拟

最常见的使用统计模型进行 $PM_{2.5}$ 值预测的方法是土地利用回归（Land Use Regression，LUR）模型。目前该方法被欧美学者广泛应用于欧洲、北美一些城市的 $PM_{2.5}$、NO_2 等污染物浓度的空间模拟(Hoek et al.，2008；ROSS et al.，2007)。国内陈莉等较早使用 LUR 模型对天津市 NO_2、PM_{10} 污染物浓度的空间分布进行了预测（陈莉等，2009）。近年来，邹滨、焦利民、陈健、吴建生等学者先后在不同城市及区域进行了 LUR 模型的应用研究，这些研究的关键点大部分集中在

解释变量的扩展、构建方法的改善以及时空分异的加强等方面（焦利民等，2015；汉瑞英等，2016；吴健生等，2015；江曲图，2017）。LUR 模型早期侧重于利用土地利用情况、气象条件与地面 $PM_{2.5}$ 浓度的相关关系来进行 $PM_{2.5}$ 浓度的空间模拟。随着对 $PM_{2.5}$ 问题认识的深入，区域交通道路状况、人口情况、地形等因素也被引入到 LUR 模型中。但使用 LUR 模型进行 $PM_{2.5}$ 浓度空间模拟也存在明显的不足：用线性回归的方法进行建模忽视了解释变量与 $PM_{2.5}$ 浓度之间存在的非线性复杂关系；模型面临多重共线性的问题；对建模解释变量的空间尺度考虑不充分；样本数据的采集不够精确；模型精度评价体系需要加强等（符立伟等，2015）。

（2） 融合新算法的 LUR 模型进行 $PM_{2.5}$ 浓度空间模拟

针对 LUR 模型的不足，一些新算法被引入到 $PM_{2.5}$ 浓度空间分布模拟中，例如地理加权回归、人工神经网络、支持向量机等（王飞龙等，2017；张怡文等，2017；喻其炳等，2017）。还有一些算法目前虽较少应用于 $PM_{2.5}$ 浓度空间分布模拟，但已经在其它领域的空间模拟研究中有过一定的应用，且获得了较好的效果。例如随机森林（Random Forest，RF），它与之前的方法不同，属于集成学习方法，具有更好的精度及泛化性，特别适合非线性、小样本数据的预测及模拟，能较好地满足 $PM_{2.5}$ 浓度空间分布模拟的现实需求（王丽爱等，2015）。陈凯等人（2017）基于随机森林来构建元胞自动机来模拟广东佛山市的城市扩展，结果发现与逻辑回归模型相比，精度有了明显的提高，此外还具有运算速度快、解释性好的优点。谭敏等人（2017）则利用随机森林优化模型对对珠江三角洲区域 2010 年的人口分布进行了空间模拟，总体模拟精度较高，但在人口密度偏高或偏低的区域模拟精度较低，该论文认为这与随机森林优化模型建立时选取的因素不能完全反映人口密度偏高或偏低的区域的人口分布特征有关。Wang 等（2017）基于气象站点及遥感数据对黄土高原的地上生物量的空间分布进行模拟，与 SVM 算法进行比较，误差评价参数更加理想。在环境污染领域，也有学者引入了随机森林优化模型。Walsh 等（2017）利用随机森林优化模型对河口沉积物污染的空间分布进行了模拟，发现模拟误差小，可以用来反映污染程度，确定污染的热点区域。因此利用随机森林的统计模型进行 $PM_{2.5}$ 浓度空间分布模拟具有实际意义，且该方法比较成熟，具有可行性，有必要进行进一步的探讨研究。在具

体的实施过程中，应注重特征变量的筛选、模型的检验与评价等。

1.2.2 LUR 模型建模热点问题研究进展

目前已有的 $PM_{2.5}$ LUR 模型众多，虽然不同模型在构建流程和模型精度上存在一定的差异，但总体上都包含以下过程，样本数据选择、建模特征变量选择、模型建立与检验、模型精度评价四个方面。对 $PM_{2.5}$ LUR 模型的研究也主要集中在以上几个方面。

（一） 样本数据选择

样本数据选择的差异主要表现在样本数据来源、样本数量大小与分布情况及采样时间等方面(罗艳青，2014)。

大部分研究基于常规监测网络来获取 $PM_{2.5}$ 浓度数据（Ross et al.，2007；Moore et al.，2007；Liu et al.，2009；Hector et al.，2012），只有少部分研究进行了以模型构建为目的的自主采样监测，以避免常规监测网络的密度不足以满足小尺度室外空气污染浓度变化建模的问题（Clougherty et al.，2008；Hoek et al.，2011），有研究甚至采用其他更精确的数据采集方法用于 LUR 模型的改良(Johnson et al.，2010)。自主采样监测就是自主设置监测网络（控制监测时间及密度，进行科学布点），在设定好的时间点进行 $PM_{2.5}$ 监测，使得样本数据满足模型构建的特定要求。但是这同时意味着自主采样监测费用高、监测的时空覆盖度低。与此相比，由政府负责布局的常规监测网则成本较低，而且所提供的监测数据具有时空覆盖度高的优势。近年来 $PM_{2.5}$ 常规监测网络的检测能力不断提升，绝大多数地区的常规监测网络已经可以很好的满足模型的构建需求。

目前还没有特定的方法用来确定某一地区需要的监测站数量和分布。$PM_{2.5}$ LUR 研究中样本数量常在 10 多个到 300 多个不等 (罗艳青，2014)。在样本数量的确定过程中，可从人口数量和城市规模等角度来判断。

关于监测时间的选择，$PM_{2.5}$ 浓度数据的获取通常选择 1-4 个长为 7-14 天的监测周期(罗艳青，2014)。虽然空气污染程度在时间序列上具有很强的相关关系，但使用周期采样数据可代替年均浓度数据。此外由于天气等不可预计的因素，在样本数据的采集过程中应尽可能选择合适的时间以防采集到影响空气污染情况的事件。

（二）特征变量的选择

根据 $PM_{2.5}$ 的污染源及扩散方式， LUR 模型建模所需的特征要素包含：土地利用类、道路交通类、人口分布类、气候类、地形类、其他类等六大类(罗艳青，2014)。土地利用类包括了地表覆盖、城市形态、污染源排放等影响 $PM_{2.5}$ 分布的因子。它是 LUR 的主要建模要素。其次是道路交通类要素和人口分布类要素，其中道路交通类要素反映的是移动源（例如汽车等）的分布，人口分布类要素则反映人类活动强度对 $PM_{2.5}$ 浓度的影响。很多相关研究在建模中融入了这三种地理要素，但是不同学者在建模中对地理要素类型的具体选择仍有不同见解，Hochadel 等选择了道路交通要素、人口分布要素、其他要素三种地理要素（Hochadel，2006）；Ross 等选择了土地利用要素、道路交通要素、人口分布要素和其他要素等四种地理要素（Ross，2007）；Clougherty 等选择了土地利用要素、道路交通要素、人口分布要素、气象要素、其他等五种地理要素（Clougherty，2008）。

在分析的地理要素类型中，不同研究引入的特征变量同样存在明显差异，如 Hector 等考虑的土地利用类的特征变量包含：农场、森林、铁路、高速路、公路、政府机构、学校、商场、医院等（Hector，2012）；Hoek 等考虑的土地利用类的特征变量包括低密度住宅区、工业、港口、城市绿地、半天然土地利用等，将到道路距离及最近道路的交通强度两类因子代表道路交通因素，在其他类要素里则采用了街道宽度等代表城市形态的因子（Hoek，2011）；Moore 等采用了道路长度表征道路交通，在其他类里面则考虑了距海岸距离、海拔等特征变量（Moore，2007）。总的来说，道路交通类的特征变量有车辆数、交通流量、交通强度、交通密度、车辆里程数、车辆类型、道路长度、到道路距离等；人口分布类特征变量包括：人口/住房数量、人/住房密度；地形要素类特征变量包括有经纬度、高度等；气象要素类特征变量包括有风速、风向、湿度等；其他类特征变量有背景污染物浓度、相关污染物浓度、污染排放、距污染源距离、距海距离、城市形态等；而由于土地利用类型的划分存在不同方式，土地利用类特征变量的选择存在明显差异。

各特征变量的值一般是在特定半径范围内，利用空间分析方法中的缓冲区分析方法获取的(罗艳青，2014)。LUR 的模拟精度及空间分辨率与给定空间范围（缓冲区）的具体形态和大小密切相关。当前看到的 LUR 研究大部分采用圆形缓冲

区。缓冲区的具体半径，绝大多数选择 20m-5000m 的半径范围（Ross et al.，2007；Hochadel et al.，2006）。只有少数考虑了大于 10km 的缓冲区大小，如 Moore 等，确定了 10km、20km 的缓冲区半径来获取人口密度特征变量（Moore et al.，2007）。此外研究表明，道路交通对 $PM_{2.5}$ 浓度的影响是随着到道路距离的增大，$PM_{2.5}$ 浓度减小；当超过主要城市道路约 100 米，或主要高速公路 500 米，则基本维持稳定（Roorda-Knape et al.，1998）。

（三）模型建立与检验

进行模型训练，首先需要对相关地理要素做预处理和优选。其中地理要素的预处理方法有 PCA、聚类分析和对数变换，地理要素优选方法包含二元线性回归方法和双变量相关性分析方法（Moore et al.，2007；Hochadel et al.，2006）。经过预处理和筛选，可以提高预测变量的预测效果，并筛选出预测能力较强的变量，从而提高模型精度。

目前使用较广的 LUR 模型训练算法是多元线性回归方法。逐步回归的原理是从数量较多的地理因子中找出某种关系来拟合模型，使得模型的拟合度（adj R^2）最大 (罗艳青，2014)。最终的模型剩下的预测变量都为代表性强和预测能力佳的变量。相比于 LUR 模型，Liu 等融合多元逐步回归模型及贝叶斯最大熵法(BME)，对台北地区 $PM_{2.5}$ 浓度进行预测，结果显示模拟的平均误差从 2.79ug/m^3 变为 2.16 ug/m^3（Liu，2009）。关于 $PM_{2.5}$ 浓度空间分布模拟，有学者针对多种地理要素尝试了地理加权回归算法，最终的模拟结果较为理想（Levy et al.，2007；王敏等，2013）。随着人工智能时代的到来，数据挖掘方法被陆续引入到 LUR 模型中，成为新的研究热点。

对模型的模拟效果进行验证是进行 LUR 建模的关键步骤。常用的效果较好的验证方法有留一法和交叉验证法。N-1 留一法基于数量为 n-1 个的站点构建模型，将预测的结果与未参与训练的唯一一个站点的真实值对比，重复操作 n 次计算平均值。交叉验证法将监测站点分成两组，一组样本用于建模，另外一组比例小的样本数据集用于模型验证。该方法的优势在于不需要密集的计算处理，劣势在于可能会由于站点的先验划分从而使检验结果出现偏差。

（四）模型精度评价与对比

大量研究已将 LUR 和其它常见的 $PM_{2.5}$ 空间模拟方法进行了模拟效果的对

比。通常，LUR 法比地统计法效果更好。Ross 等人对纽约市 $PM_{2.5}$ 浓度进行了模拟，LUR 模型在对试验区 $PM_{2.5}$ 空间浓度的模拟过程中，预测站点的 RMSE 为 1.15 ug/m^3 和 1.00 ug/m^3，kriging 方法则明显较大，达到 1.30 ug/m^3 和 1.47 ug/m^3（Ross et al.，2007）； Liu 等在模拟台北市的 $PM_{2.5}$ 浓度过程中发现， LUR 模型的平均误差仅为 2.79 ug/m^3，Kriging 模型的平均误差达到了 3.18 ug／m^3（Liu，2009）。但目前缺乏对监测站点以外地区 $PM_{2.5}$ 浓度模拟精度的评价，评价的科学性仍然存在疑问。此外由各类数据挖掘算法改进的 LUR 模型的涌现使得有必要对模型精度评价与对比的过程进行进一步的探讨。

通过对以上流程的分析，LUR 建模研究在以下方面仍需提高：如何选择最优的空间作用尺度以反映各种特征变量对于 $PM_{2.5}$ 浓度的影响；如何评价各特征变量对于 $PM_{2.5}$ 浓度的重要性；需要对各类算法进行系统科学的比较；精度评价的对象仅集中在 $PM_{2.5}$ 浓度监测的样本点，未针对其它地区开展 $PM_{2.5}$ 浓度模拟的精度评价，且未考虑模型的空间分辨率等特性。

1.2.3 $PM_{2.5}$ 暴露风险评估研究进展

2010 年我国因与 $PM_{2.5}$ 污染相关的提前死亡人数为 120 万左右，占我国居民年度死亡总数的 1/9，因此公众健康受高浓度 $PM_{2.5}$ 的威胁很大（Hoek et al.，2011）。寻找评估 $PM_{2.5}$ 污染暴露下的居民健康风险的定量方法，同时对较严重的空气污染事件进行健康预警具有重要的意义。

大气污染物浓度是是目前普遍使用的空气污染暴露风险评价表征方法。提升 $PM_{2.5}$ 浓度空间分布模拟的效果可以改善 $PM_{2.5}$ 暴露风险的评价。但是这种测度方法的准确性仍存在问题，例如在无人区的 $PM_{2.5}$ 污染物高浓度值区域通常无人口 $PM_{2.5}$ 污染暴露风险（伏晴艳等，2004）。因此有学者在 $PM_{2.5}$ 暴露风险评价的过程中应该考虑人口分布的非均一性（Hoek et al.，2002）。同丽嘎等采用克里金插值法和人口 $PM_{2.5}$ 暴露相对风险模型，基于 $PM_{2.5}$ 监测站数据、人口普查数据和土地利用数据，在 $PM_{2.5}$ 时空分布特征分析的基础上完成了包头市居民 $PM_{2.5}$ 暴露风险研究（同丽嘎等，2015）。在今后的研究中，应着力于提高 $PM_{2.5}$ 浓度及人口密度空间数据的准确性。邹滨等集成了大气污染扩散模型、地理信息空间分析技术等方法得到了较高空间分辨率的 $PM_{2.5}$ 污染浓度和人口密度空间分布数据，之后进一步通过构建人口 $PM_{2.5}$ 污染暴露相对风险评价模型完成人口 $PM_{2.5}$

污染暴露程度的空间格局展示（邹滨等，2013）。

但上述研究都仅是利用了单一方法来进行人口 $PM_{2.5}$ 污染暴露程度，存在局限性。目前常用的评价方法有空气质量浓度、人口暴露强度和人口加权浓度大气污染暴露风险评估方法。利用空气质量浓度进行评估的方法，假设评估区域人口分布均一，评估存在一定的理论偏差（Kousa A et al.，2002）；人口暴露强度结合空气质量浓度空间分布和人口空间分布，以栅格运算的方法，理论上提高了评估结果的精度，但在较大空间范围内存在各评估单元间风险值无法比较的问题（Cao Junji et al.，2005；Wang Shuxiao et al.，2007）；人口加权浓度的人口空气污染暴露风险评价指标是一种顾及人口空间分布的暴露评估方法，可在较大空间范围内评估各评估单元间的人口空气污染暴露风险，但需要先确定评估单元，然后进行统计，在空间分辨率上会有欠缺（伏晴艳等，2004）。人口暴露强度、人口加权浓度这两种评估方法最明显的区别在于前者是基于格网的，后者则基于区域（例如城市、城市群等）。因此有必要对上述三种方法进行对比研究，综合的利用上述三种方法来进行人口 $PM_{2.5}$ 污染暴露风险的评估。邹滨等人即针对单一指标的局限性， 根据“空气质量浓度 、人口暴露强度、人口加权浓度 3 种指标，对比了各指标在长沙市 $PM_{2.5}$ 暴露风险区划中的特点。

综上所述，本论文拟采用随机森林改进 LUR 模型（下文简称随机森林优化模型）并与其它统计模型进行系统性的比较研究，旨在说明随机森林优化模型在 $PM_{2.5}$ 浓度空间分布模拟中的优势，获取更高精度的模拟结果。此外，本论文希望通过获取的 $PM_{2.5}$ 浓度空间分布信息及空间化的人口数据，以多指标综合评价的方式对人口 $PM_{2.5}$ 暴露风险进行研究。这一做法可以使人口 $PM_{2.5}$ 暴露风险的评估相比先前具有更高的可靠性，可为大气污染治理和人居环境的提高提供重要的科学依据。

1.3 研究内容

本文的主要研究内容如下：

（1） 地理要素与 $PM_{2.5}$ 的相关性分析

根据 $PM_{2.5}$ 的源解析结果及常用地理相关变量，考虑长江三角洲地区实际情况，选取气溶胶光学厚度数据（Aerosol Optical Depth，AOD）、气象、土地利用、地形、交通、人口密度等变量。以格网为单位，直接提取数据或通过缓冲区方式

提取数据，对数据进行前期处理后，获得地理要素数据。通过双变量相关分析探讨各类地理要素数据与 $PM_{2.5}$ 浓度数据之间的相关关系，将相关性较强的地理要素数据筛选作为建模的目标因子。

（2） 基于随机森林优化模型的长江三角洲地区 $PM_{2.5}$ 空间模拟.

基于随机森林优化模型，建立相应模型，并通过相关的检验方法证实随机森林优化模型相比传统的 LUR 模型、支持向量机改进的 LUR 模型的优势。基于该模型，对长江三角洲地区 $PM_{2.5}$ 浓度进行空间模拟。

（3） 人口 $PM_{2.5}$ 暴露风险评价

利用 $PM_{2.5}$ 浓度及人口密度空间数据，以人口 $PM_{2.5}$ 暴露风险模型对长江三角洲人口 $PM_{2.5}$ 暴露风险进行评价。

1.4 技术路线

本文的研究技术路线如图 1-1 所示。

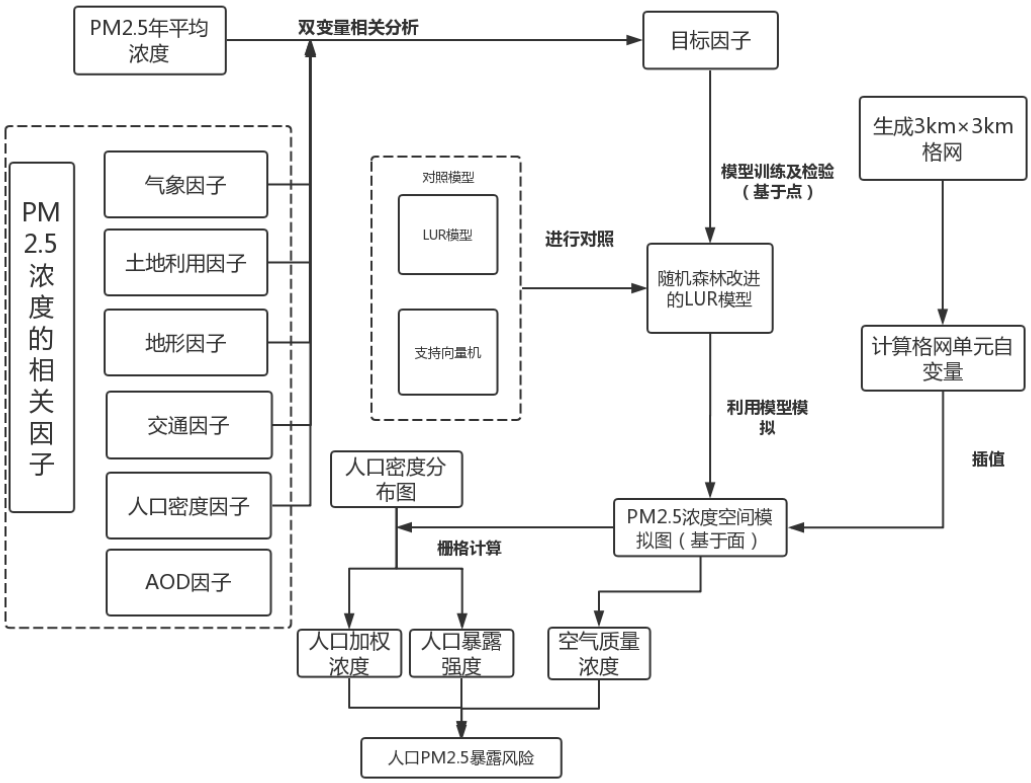


图 1-1 研究技术路线图

Fig. 1-1 research technology roadmap

第二章 数据来源与研究方法

2.1 研究区概况

长江三角洲地区位于我国东部沿海和长江流域的结合处，包含了上海市、江苏省和浙江省两省一市（图 2-1）。该地区的气候类型为亚热带季风气候，雨热同期，夏季高温且多雨，冬季温和而少雨。它以全国 2.2%的土地和 11.7%的人口，占据了全国约 21%的 GDP。长江三角洲地区城市化水平高达 64.7%，城镇连绵分布，是我国经济发展的先行者（上海市统计局，2011；江苏省统计局，2011；浙江省统计局，2011）。但是正是由于高速发展的工业化以及城市化给该地区的生态环境带来了不容忽视的影响，近年来环境事件的发生越加频繁。虽然这一地区的 $PM_{2.5}$ 问题不如京津冀等地区严重，但是同样造成了巨大的危害。该地区的 $PM_{2.5}$ 问题已经日益受到公众的关注，且具有一定的特殊性，但是目前的相关研究还很不充分。所以，本文将研究范围选定为长江三角洲地区的核心区域，包括上海、南京、苏州、杭州等 16 个城市。

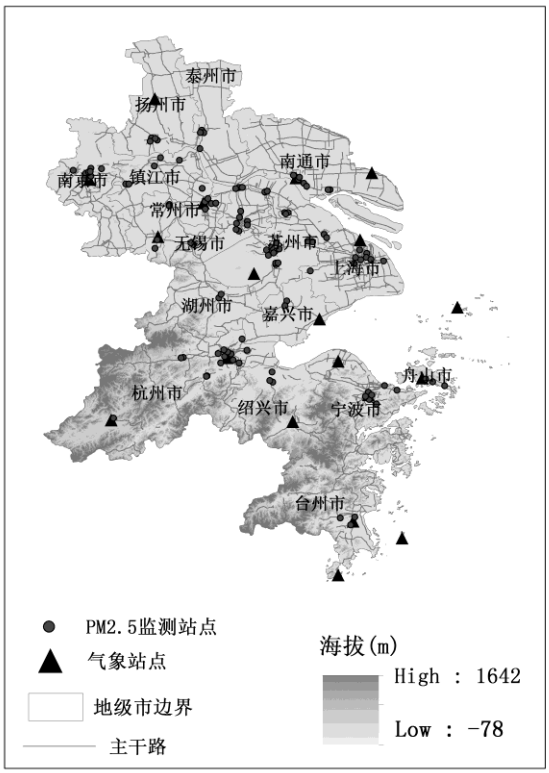


图 2-1 长江三角洲核心区域图

Fig. 2-1 map of the core area of the Yangtze river delta

2.2 数据来源与处理

相关因子的选择是 $PM_{2.5}$ 浓度空间模拟的关键性问题。参考已有研究，常用的地理因子包括土地利用、自然地理要素、气象要素、人口密度以及交通等（Hoek et al.，2008；Gulliver et al.，2011）。$PM_{2.5}$ 污染的污染源包括工业排放、机动车尾气、土壤扬尘以及二次污染。其中二次污染与气象因素有关（成海容等，2012）。根据长江三角洲地区的实际情况，结合 $PM_{2.5}$ 源解析的经验及常用地理相关因子，本研究共选取了 AOD、土地利用、气象、地形、交通、人口密度等 5 类相关因子。在监测点的缓冲区内计算耕地、林地、草地、水体、建设用地等土地利用类型的比例来代表土地利用状况；缓冲区范围内道路长度代表道路交通状况；缓冲区内人口密度代表人口密度状况。

2.2.1 AOD 数据

目前使用最为广泛的 AOD 产品是 MODIS 中的 AOD 产品。MODIS 传感器搭载在美国国家航空航天局的地球观测系统（Earth Observing System，EOS）上的 Terra 和 Aqua 卫星。其中 Terra 为上午星，Aqua 为下午星。本文所使用的 AOD 数据下载自 https://neo.sci.gsfc.nasa.gov/，该产品的反演采用深蓝算法和暗像元算法融合的方式。产品的空间分辨率为 1 °×1 °，时间分辨率为月均数据，时间跨度从 2015 年 1 月至 2015 年 12 月，共包含 12 幅图像。

2.2.2 $PM_{2.5}$ 数据

该数据下载自中国环境监测总站开放的城市空气质量实时发布平台（http://http://www.cnemc.cn/）。数据集共包含 214 个监测站，数据的时间范围为 2015 年 1 月 2 日~2015 年 12 月 31 日（1 月 1 日数据缺失）。源数据是每小时观测值，日均值经均值计算得到，年均值在日均值的基础上求均值获取。经描述性统计，数据集的最大值为 69.86 ug/m^3，最小值值为 28.66 ug/m^3，变异系数达到 13.7%，归为中等变异程度。利用 SPSS19 中的非参数检验模块（K-S 检验）对 $PM_{2.5}$ 浓度观测值进行检验，P 值小于 0.05，属于非正态分布。

2.2.3 土地利用数据

土地利用数据来源于全球 30 m 地表覆盖信息服务系统（http://www.globallandcover.com/），空间分辨率为 30 m。该数据的总体精度为 83.51%，kappa 系数为 0.78。本文以监测站为中心做 5 类不同半径的缓冲区（1km、

3km、5km、7km、10km），然后借助 ArcGIS 中的 Zonal Histogram 工具分别统计不同缓冲区内各种土地利用类型的比例。

2.2.4 气象数据

该数据在中国气象科学数据共享服务网（http://data.cma.cn/）上获取，数据集包括日均气温、相对湿度、风速、气压以及 24h 累计降水量数据。本文对各项数据求年均值，然后进行空间插值以获取区域内气象要素的连续表面。

2.2.5 地形数据

地形数据采集自地理空间数据云(http://www.gscloud.cn/search)，以 SRTM 全球 90m 分辨率高程数据来获取地形状况。通过监测站点可提取高程数据对应位置的海拔高度以及领域内的起伏度，本文仅考虑海拔高低和起伏度与 $PM_{2.5}$ 浓度之间的关系。

2.2.6 道路交通数据

道路交通数据获取自 Wiki 世界地图数据库（http://www.openstreetmap.org/）。道路对 $PM_{2.5}$ 浓度的影响通过一定缓冲区内道路长度表示。本文以不同值作为搜索半径（0.5km、1km、3km、5km、7km、10km），利用线密度法生成道路密度图，然后利用监测站点提取所在位置的道路密度状况。

2.2.7 人口密度数据

人口密度数据获取自 Worldpop 项目（http://www.worldpop.org.uk/）。该项目的 WorldPop 数据集是高空间分辨率的，较新的世界人口分布数据。本文选取的数据为 2015 年数据，空间分辨率为 100m，单位为每公顷人数。通过监测站点可直接提取对应位置的人口密度数据。

2.3 研究方法

2.3.1 LUR 建模核心方法

（1）缓冲区分析

缓冲区分析（buffer analysis）是在点、线、面等地理要素的基础上，设定距离值进行扩展，从而建立要素周围特定距离范围内缓冲区图形的一种地理信息方法。LUR 建模过程中，缓冲区分析主要用于提取监测站点附近一定范围内地理要素的相关因子，例如 3km 范围内建设用地占比。缓冲区的形态和半径需要考虑 $PM_{2.5}$ 在大气中的扩散形式及各特征变量对 $PM_{2.5}$ 浓度的影响而决定。本文采

用圆形缓冲区，初步选用 1km、3km、5km、7km、10km 等距离进行实验。

（2）相关分析

相关分析（correlation analysis）用于探索客观事物（变量）间相关关系及其强弱的计量方法。进行相关分析，可以计算得到特征变量间的相关系数。相关系数表示要素之间相关性的大小。

$$R = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (2\text{-}1)$$

式中x_i、y_i为样本 i 的 x、y 变量观测值，$\bar{x}$、$\bar{y}$是所有样本的均值。R 是相关系数，当 R 为正数时，表示变量 x 和 y 呈现正相关关系，相反的，R 为负数时，呈现为负相关。R 的取值范围在$[-1, 1]$之间，如果 R 的绝对值越大，表示变量之间的相关性越强。

在 LUR 建模过程中运用双变量相关分析，目的在于分析相关因子与 $PM_{2.5}$ 浓度之间的相关性，进一步挑选出相关性较强的目标因子，以用于模型构建。

（3）回归分析

在 LUR 建模时需要定量的表示目标因子与 $PM_{2.5}$ 之间的关系，因此回归分析在其中起到了关键的作用。LUR 建模常用多元线性回归方法建立 LUR 模型。传统的 LUR 模型通常采用的是多元逐步回归方法，也可利用随机森林、支持向量机、GWR 等算法对 LUR 模型进行优化。

该方法在 LUR 建模中的基本思想是将 $PM_{2.5}$ 分解为受可量化因素（即目标因子）影响的规律性部分，以及难量化因素影响的非规律部分。规律性部分可以用多元线性回归方程来表示，非规律部分即为方程的残差。多元回归方程的计算值即为 $PM_{2.5}$ 的估算值。具体步骤如下：

首先，将站点的 $PM_{2.5}$ 浓度数据作为因变量，各站点对应的目标因子数据作为自变量，建立它们之间的多元线性关系，获得回归方程；然后将每个栅格的目标因子数据代入回归方程，即可得到各栅格对应的 $PM_{2.5}$ 浓度。

2.3.2 模型精度评价方法

对模型进行精度评价可反映模型的拟合效果及误差。回归方法是对已给出的数据求得最优值，在新数据集上不一定有很好的表现。交叉验证（cross-validation，CV）是一种精度评价方法，它将大部分原始数据作为训练样本，剩余的样本作为验证集。先由训练集进行训练获得回归方程，然后利用验证集对所得模型进行检验，评价模型效果的性能。该方法可以做到模型的偏差和方差的平衡。本文采用的就是交叉验证方法，预留 20%的监测站点作为验证集。通过判断验证集的模拟结果和实际的观测值之间的接近程度，来反映模型的模拟精度。本文采用的评价指标包括相关系数（R）、均方根误差（Root Mean Squared Error，RMSE）、平均绝对误差（Mean Absolute Error，MAE）以及拟合指数（Index of Agreement，IA）。其中，R 表示模型的拟合效果，RMSE 和 MAE 表示预测值与真实值之间的偏差，而 IA 表示预测值和真实值之间的一致性（朱亚杰等，2016）。结合 4 个指标可较全面的评价所得的模型。

$$R^2 = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}} \tag{2-2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1}} \tag{2-3}$$

$$MAE = \frac{\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|}{n} \tag{2-4}$$

$$IA = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(\left|\hat{y}_i - \bar{y}_i\right| + \left|y_i - \bar{y}_i\right|)^2} \tag{2-5}$$

式中：y_i 为实际观测值；$\hat{y}_i$ 为模型的预测值；$\bar{y}_i$ 为观测值的平均值；n 为预测样本数。

此外本文也引入信息丰富度的概念进行模型的评价。在信息论中常用熵来表示信息的丰富程度。设有矢量 V= $\{x_1, x_2, \ldots, x_3\}$，如果

$x_i \in V$的概率$p_i = P（x_i）$，则 V 的信息熵可以按如下公式表示：

$$E（V）=-\sum_i^n P(x_i)\log_2(P(x_i)) \quad （2\text{-}6）$$

由于 $PM_{2.5}$ 年平均值在区域内具有明显的空间变异特性，所以特定区域内的 $PM_{2.5}$ 年平均浓度表面图像常含丰富的信息，图像的信息熵小。本文根据这一特性，尝试以图像信息丰富度来评价模型精度。如果今后对模拟方法需进一步改进，可作为对比指标之一。

2.3.3 随机森林算法

随机森林(random forest)属于集成学习算法，它本质上是由多棵决策树组合而成，可用于分类、聚类、回归等。对所有树的结果取平均可以进行回归预测（崔东文等，2014）。其基本原理如下：如果有 M 个输入变量，在每个节点随机选择 m（m<M）个特定变量，运用这 m 个变量作为决策树分裂的候选变量，从这些变量中选择信息含量最丰富的变量来进行节点分裂，且不对决策树进行剪枝，使其尽可能的进行生长。最后可以通过对所有决策树做加总，预测新数据（马玥等，2016）。预测时，如果用于分类时，多采用多数投票，而回归则经常采用平均法计算。总体而言，随机森林具有以下优点：运算量小，但预测精度高；可以高效的处理非线性过程；预测结果对非平衡数据和缺失数据较稳健（朱蕾等，2007）。因此随机森林是目前性能最好的机器学习算法之一。

本文用 Python 的 Scikit-Learn 模块对随机森林进行回归建模。建模过程中需要使用 ntree 和 ntry 两个自定义参数，以优化模型。ntree 指组成随机森林的决策树数量，也就是重抽样次数；mtry 为所用特征变量的数目，通常是所有输入变量的 1/3。

随机森林算法支持以下常用功能：

（1） OOB 估计：随机森林算法用 bootstrap 法进行，训练集的样本数 N 趋向于无穷大时，各样本未被抽中的概率为（1-1/N）N，此概率值将收敛到 1/e，约为 0.368，也就是说有 36.8%的样本被作为袋外数据（out-of-bag）。该类数据能够估算随机森林的泛化误差。

（2） 变量重要性度量：回归变量之间存在相互作用，因此变量的重要性较难定义。随机森林利用 OOB 来对要素的重要性测度。每棵决策树都具有一个

误分率。在随机森林算法中对袋外数据的某一变量进行随机调整顺序或者加入一定的噪声，同时保持其它变量不变，经过修改 OOB，并由分类树计算新的预测值，分析由袋外数据改变所引起的误差增加来估算某一变量的重要性。修改后的 OOB 误分率减去原始 OOB 误分率，再除以标准差即为变量的重要性。OOB 的变化越显著，则该变量越重要。

（3） 随机特征选取：随机森林算法引入了随机性。它对每棵树的每个节点的一部分变量进行了分割，树的生长只取决于所选的部分输入变量，这一做法可解决数据高维度的问题。

2.3.4 支持向量机

支持向量机（SVM）是 Vapnik（1995）开发的一种机器学习方法。该算法满足结构最小风险准则，把训练误差作为约束条件使得置信范围最小。SVM 算法由较少的样本信息对模型的复杂度与学习力进行折中处理。SVM 最终将转化为二次规划问题进行求解，解为全局优化解。支持向量机不需要进行模型假设，其最大优势在于能够较好的避免维数过高和过拟合，总体而言精度较高，运算的速度快（Vapnik，1998）。

SVM 建模的思想为：利用非线性转换函数 Φ 把低维空间中的 x 映射到高维空间 Φ（x），从而在特征空间中求出回归超平面，解决了高度非线性的问题。

2.3.5 随机森林优化模型

在 2.3.1 中已经对传统的土地利用回归模型进行了详细的阐述。传统的土地利用回归模型，在模拟过程中采用的是多元线性回归。多元线性回归不可避免的忽视了要素之间的交互效应。因此有必要采用其它的算法来替代多元线性回归方法。而随机森林已被证明可以高效的处理非线性过程，预测的精度很高。因此本研究采用双变量相关分析提取建模所需要的目标因子（其中包括 AOD 数据），将训练集中的 $PM_{2.5}$ 浓度以及相应的目标因子作为数据进行随机森林回归建模。

第三章 基于随机森林的长江三角洲 $PM_{2.5}$ 浓度空间模拟

3.1 样本数据划分及目标变量提取

3.1.1 样本数据划分

长江三角洲地区共有 115 个站点存在有效的 $PM_{2.5}$ 浓度年均值，将其作为总体样本。随机选择 20%的监测站点作为测试数据，其余数据作为训练数据。因此训练集的样本数量为 92 个，测试集的样本数量为 23 个。各类样本的基本信息见表 3-1。

表 3-1 各类样本的基本信息表

Table 3-1 basic information table of various samples

样本类别	样本容量	最小值	最大值	平均值
总体样本	115	28.66 ug/m^3	69.86 ug/m^3	54.54 ug/m^3
训练集	92	28.66 ug/m^3	69.86 ug/m^3	54.08 ug/m^3
测试集	23	39.08 ug/m^3	63.49 ug/m^3	55.65 ug/m^3

3.1.2 目标变量提取

不同区域，相关地理因子与 $PM_{2.5}$ 浓度之间的相关关系不同。因此在正式建立模型之前，对各相关地理因子与 $PM_{2.5}$ 浓度做双变量相关分析，计算相关系数。据双变量相关分析的结果，识别和 $PM_{2.5}$ 年平均浓度显著相关的目标因子。对于同类因子，选择其中相关性最大的因子作为目标因子。

(1)气溶胶光学厚度(AOD)与 $PM_{2.5}$

气溶胶光学厚度(AOD)，一般是指整层气溶胶的消光系数在垂直方向上的积分,表示的是气溶胶对光的衰减作用（王静等，2010）。$PM_{2.5}$ 是气溶胶的重要组成部分，它与 AOD 存在某种关系。分别提取监测站点处的 AOD 数据（Terra，Aqua）与相应监测站点的 $PM_{2.5}$ 浓度做相关分析，结果显示 AOD(Terra)与 $PM_{2.5}$ 的相关系数为 0.422，通过 0.01 水平的显著性检验，相比之下 AOD(Aqua)的相关系数较低，为 0.261。因此，可以选择使用 AOD(Terra)对 $PM_{2.5}$ 进行反演。在统计模型中，引入 AOD 数据可以增强模型效果。

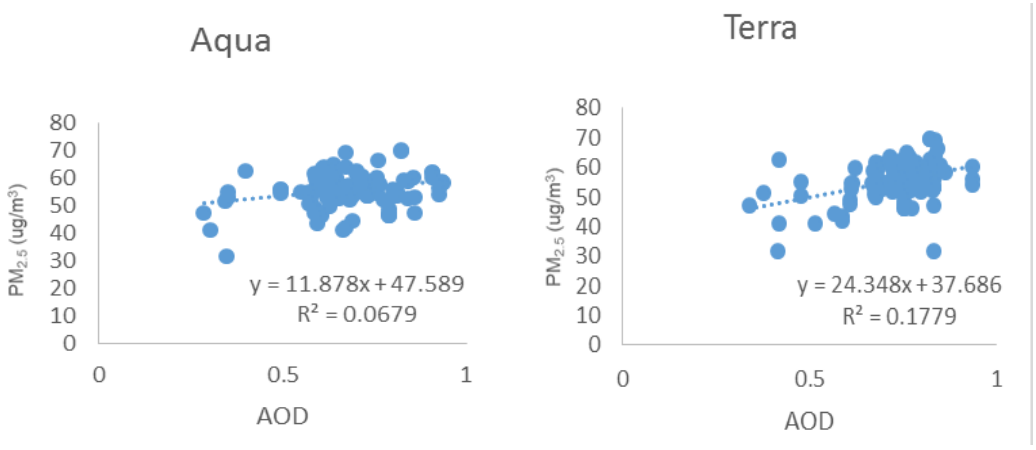


图 3-1 AOD(Aqua，Terra)与 $PM_{2.5}$ 的散点图

Fig. 3-1 scatter diagram of AOD(Aqua, Terra) and $PM_{2.5}$

(2) 地形与 $PM_{2.5}$

地形对 $PM_{2.5}$ 也有较明显的影响，例如三面环山，类似于簸箕状的地形，容易形成静风逆温的不利条件，使得大气污染物难以扩散（车瑞俊等，2007）。由于具体的地形类型与 $PM_{2.5}$ 的关系难以寻找，本文选择了起伏度、高程以及坡度三个地形指标与 $PM_{2.5}$ 浓度进行相关分析。其中起伏度指在某一范围内，最高点高度与最低点的海拔差距，可表示某地区地形的宏观特征。本文起伏度的区域尺度定义为 10×10 的邻域。经计算，高程、起伏度与 $PM_{2.5}$ 的相关系数分别为-0.257与-0.241，它们都在 0.01 的置信度水平下通过检验，坡度的相关系数为-0.107，相关性不显著。

(3) 气象条件与 $PM_{2.5}$

气象条件对于 $PM_{2.5}$ 浓度的变化有很大的贡献。温度、相对湿度、气压、降水与 $PM_{2.5}$ 的相关性均在 0.01 的显著性水平下通过检验，而风速则在 0.05 的显著性水平下通过检验。气温与 $PM_{2.5}$ 浓度的相关系数为-0.441，这可能是由于温度越高，越不易形成逆温，有利于污染物的扩散（王琪等，2014）。相对湿度与 $PM_{2.5}$ 浓度的相关系数为-0.439。一般认为，相对湿度对 $PM_{2.5}$ 的作用存在阈值，当小于阈值时，由于相对湿度增大时，空气中的水分增加，起到凝结作用，使气溶胶颗粒物浓度增大，但当大于阈值时，空气中水分能够使细颗粒物湿沉降，减小了 $PM_{2.5}$ 浓度（徐杰等，2017）。长江三角洲地区空气湿度大，当湿度增大时，超过阈值能够起到湿沉降的作用。气压与 $PM_{2.5}$ 浓度的相关系数为 0.375，这与温度的作用原理相似，气压升高，大气的稳定性强，污染物扩散能力弱，从而导致 $PM_{2.5}$ 浓度增大。风速与 $PM_{2.5}$ 浓度为负相关的关系，相关系数为-0.210，风速

越快越有利于污染物的扩散，从而降低 $PM_{2.5}$ 浓度。

（4）人口密度与 $PM_{2.5}$

人口密度与 $PM_{2.5}$ 浓度的相关系数为 0.114，相关性不显著。人类活动是空气污染的主要原因，但并非是指人口增长，这是一个综合性的人类活动，包括能源消耗量、汽车数量等因素都会造成空气质量下降，因此单纯将人口密度与空气污染划等号是不准确的。需要寻找更具代表性的社会经济数据来表示人类活动强度。

（5）交通与 $PM_{2.5}$

交通导致的人为源会对大气污染物浓度产生影响。以 0.5~10km 为搜索半径，产生道路密度分布图，提取交通指标。提取的指标为道路密度。由图 3-2 可知长三角整体的道路密度偏大，尤其是该地区的中心城市周边道路密度非常大。经过相关系数的计算，各搜索半径下，道路密度与 $PM_{2.5}$ 浓度的相关性都不显著。这与长江三角洲地区道路密度普遍较大，而监测点多分布于城区，交通差异更小，因此对 $PM_{2.5}$ 空间分异的贡献不足，不能成为主导因素。例如 1km 搜索半径下的长江三角洲道路密度见图 3-2。

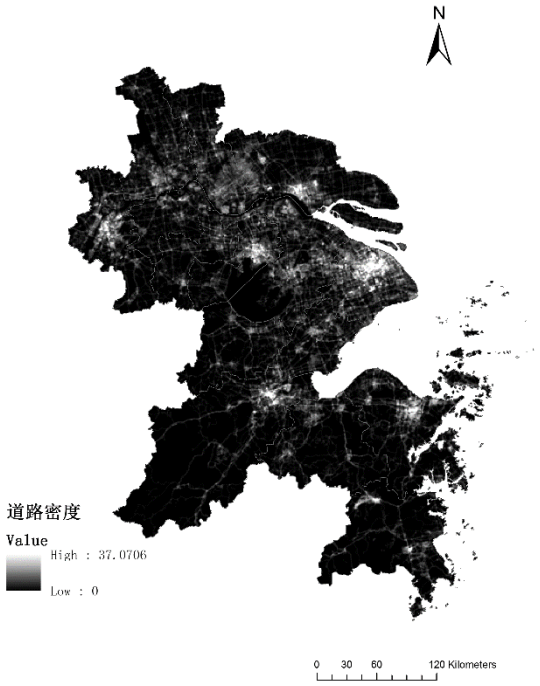


图 3-2 1km 搜索半径下的长江三角洲道路密度图

Fig. 3-2 Road density map of the Yangtze river delta under the search radius of 1km

(6)土地利用与 $PM_{2.5}$

表 3-2 土地利用因子相关分析结果

Table 3-2 correlation analysis results of land use factors

	目标因子	与 $PM_{2.5}$ 浓度相关系数 R
1km	耕地	0.039
	林地	-0.358*
	水体	-0.215*
	建设用地	0.199*
	草地	-0.072
3km	耕地	0.05
	林地	-0.495 **
	水体	-0.072
	建设用地	0.266 **
	草地	-0.239 **
5km	耕地	-0.015
	林地	0.493 **
	水体	0.078
	建设用地	0.267 **
	草地	-0.329**
7km	耕地	0.008
	林地	-0.511**
	水体	0.156
	建设用地	0.260**
	草地	-0.432**
10km	耕地	0.073
	林地	0.507**
	水体	0.148
	建设用地	0.244**
	草地	-0.482**

注：***、**、*分别表示在 0.01、0.05、0.1 的水平上显著

土地利用对大气污染具有重要影响。经过相关分析，林地在 5km 的缓冲区下与 $PM_{2.5}$ 的浓度的相关系数为-0.511，在 0.05 的显著性水平下通过检验。这表明林地对 $PM_{2.5}$ 浓度具有明显的负效应。林地面积越大，$PM_{2.5}$ 浓度越低。草地在 10km 的缓冲区下与 $PM_{2.5}$ 浓度的相关系数为-0.482，同样在 0.05 的显著性水平下通过检验。表明草地对 $PM_{2.5}$ 浓度的降低有显著作用，但草地在大尺度下作用才较明显。碎片化的草地分布并不会显著降低 $PM_{2.5}$ 浓度，连片的大范围分布的草地才会对 $PM_{2.5}$ 浓度产生明显的负效应。总体而言，公园绿化用地对 $PM_{2.5}$ 浓度的降低是有较好的改善作用的。建设用地在 5km 的缓冲区下与 $PM_{2.5}$ 的浓度的相

关系数为 0.267，在 0.05 的显著性水平下通过检验。这说明城镇化扩张导致的建筑扬尘以及工业用地带来污染气体，能够加重雾霾天气。水体在 1km 的缓冲区下与 $PM_{2.5}$ 的浓度的相关系数为-0.215，在 0.1 的显著性水平下通过检验。水体能够缓解小范围内的 $PM_{2.5}$ 污染，对于较远距离的地区改善作用有限。而耕地与 $PM_{2.5}$ 浓度的相关性不显著，这与长江三角洲地区耕地的分布较为分散有关，对于 $PM_{2.5}$ 污染的缓解作用有限。

综合各类因子，本研究最终选择的目标因子为 AOD(Terra)、起伏度、高程、降水、气温、气压、相对湿度、风速、7km 缓冲区内的林地比例、5km 缓冲区内的建设用地比例、1km 缓冲区范围内的水体比例、10km 缓冲区范围内的草地比例。

3.2 随机森林优化模型的训练及拟合

3.2.1 随机森林优化模型训练

通过 Python3.5 进行随机森林优化模型的训练，主要采用的模块为 Scikit-Learn 模块（赵佳楠等，2018）。将相关分析选出的目标因子作为模型的特征，监测站点的 $PM_{2.5}$ 浓度作为监督值，进行随机森林算法的训练。本文使用交叉检验进行误差的估计，以测试该模型的有效性。长江三角洲地区共有 115 个存在有效的 $PM_{2.5}$ 浓度年均值的监测站点。利用随机抽样选取 20%的监测站点作为测试数据，其余监测站点为训练数据。

利用 Scikit-Learn 模块中的 RandomForestRegressor 函数进行算法的训练，从而实现随机森林优化模型的构建。训练所需的数据来自于训练集，样本容量为 92。计算过程中需要考虑两个关键参数，包含随机森林算法中的回归树数量以及每次建立回归树选入的自变量数量，分别表示为 n_estimators 及 max_features（赵佳楠等，2018）。通过测试不同回归树数量所得模型的均方根误差(RMSE)以及 n_estimators 与 RMSE 的关系图，确定 n_estimators 的具体值；max_features 的值使用系统默认设置（变量数的开方）。

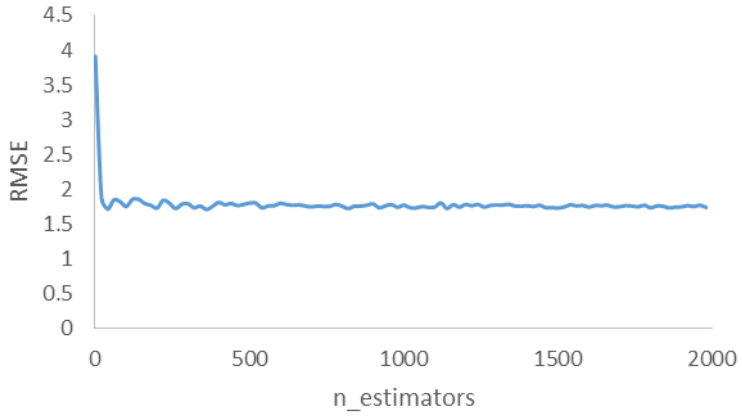


图 3-3 n_estimators 与 RMSE 关系图

Fig.3-3 n estimators and RMSE diagram

根据图 3-3，RMSE 的值随着回归树的增加先降低，后趋于稳定。随机森林回归树的 n_estimators 超过 500 之后，RMSE 的波动几乎很小。为保证模型的误差趋于稳定，并且运算效率可接受的情况下，选择回归树数量为 1500。

随机森林的抽样为有放回抽样，且设置了袋外数据。基于袋外数据可计算两种指标用来评价自变量的影响力（赵佳楠等，2018）：一是自变量在袋外时模型均方误差(Mean Squared Error, MSE)增量（%IncMSE）；自变量在袋外时对模型树节点纯度的影响(IncNodePurity)。上述两项指标越大都表示变量的影响越大。

表 3-3 自变量影响力评价

Table 3-3 influence evaluation of independent variables

目标因子	%IncMSE	IncNodePurity
AOD	16.100	455.409
起伏度	11.362	290.512
海拔	9.735	190.118
降水	13.696	382.497
气温	23.745	520.798
气压	13.449	221.496
相对湿度	36.783	1532.858
风速	21.733	431.211
林地_7km	18.175	738.624
建设用地_5km	4.547	347.538
水体_1km	1.181	110.575
草地_10km	20.558	849.768

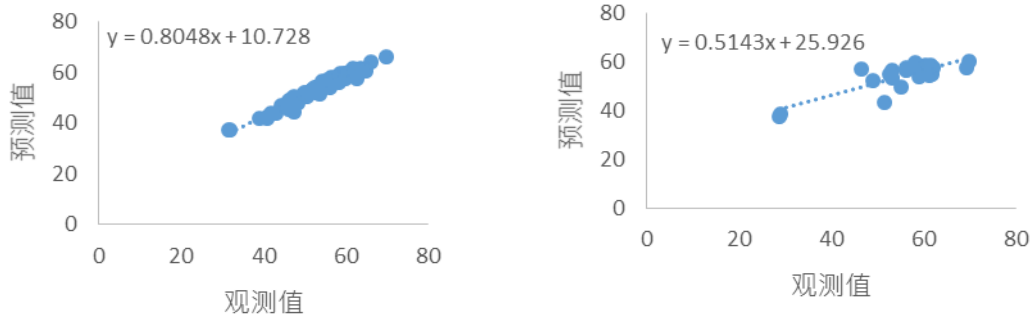
由表 3-3，相对湿度、气温、10km 缓冲区内的草地比例、7km 缓冲区内的林地比例、风速、AOD 等目标因子对模型的影响力大，对精度的贡献大。其中

相对湿度的%IncMSE 达到 36.783，IncNodePurity 为 1532.858，两项指标在各目标因子中均为最大，因此其对于 $PM_{2.5}$ 浓度预测的精度贡献最大。

3.2.2 随机森林优化模型拟合效果评价

本文使用的模拟方法为随机森林优化模型，为表现随机森林优化模型的模拟的精度，分别计算随机森林优化模型、支持向量机优化模型、LUR 模型的拟合指标，并将随机森林优化模型与另两类模型进行比较。

（1）随机森林优化模型拟合效果



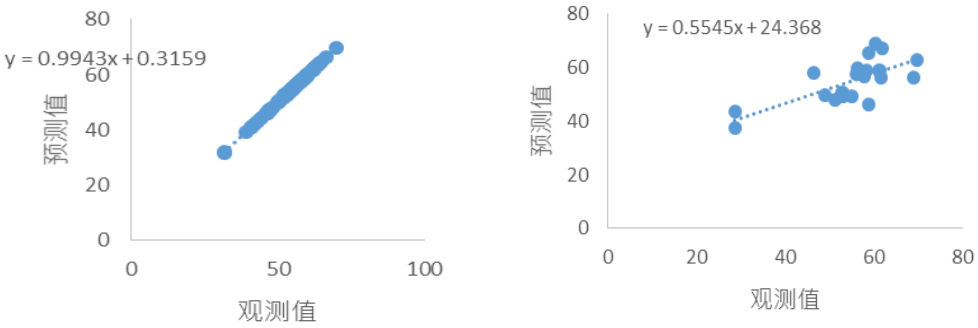
a 训练集　　b 验证集

图 3-4 随机森林优化模型拟合结果散点图

Fig. 3-4 random forest model fitting results scatter diagram

利用交叉检验的方法，对随机森林优化模型进行训练。图 3-4 是随机森林优化模型的预测值与观测值的散点图（包括训练集及验证集），无论是训练集，还是验证集，观测值与预测值之间都存在较好的拟合。进一步计算有关于预测精度的指标，训练集的相关系数（R）、均方根误差（RMSE）、平均绝对误差（MAE）以及拟合指数（IA）分别为 0.977、1.761、1.311、0.979，而验证集对应的指标值分别为 0.831、5.871、4.757、0.854。

（3） 支持向量机优化模型拟合效果



a 训练集　　b 验证集

图 3-5 支持向量机拟合结果散点图

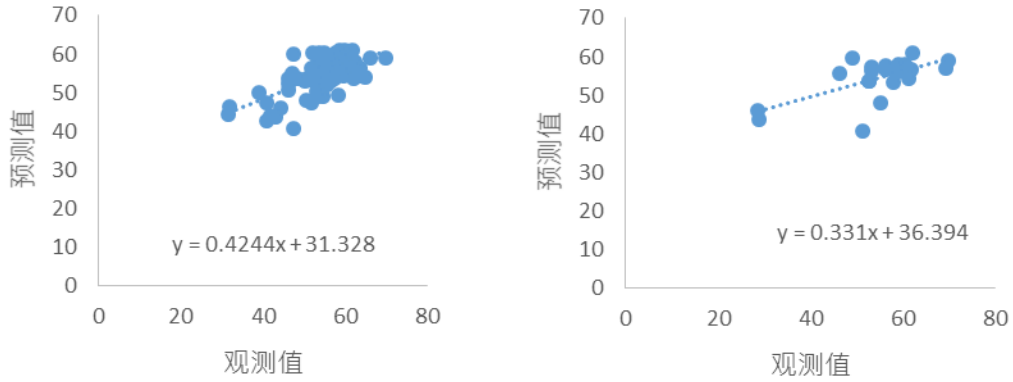
Fig. 3-5 support vector machine fitting results scatter diagram

为反映随机森林森林模型的效果，利用相同的目标因子进行支持向量机优化模型的训练，模型计算通过 Python 的 sklearn 完成。图 3-5 是支持向量机优化模型的预测值与观测值的散点图（包括训练集及验证集），训练集的拟合度非常好，但是验证集的效果欠佳，这表明使用支持向量机模型会带来明显的过拟合。它的验证集代入模型，计算所得 R、RMSE、MAE 以及 IA 分别为 0.714、6.871、5.521、0.825。

（3）传统土地利用回归模型（多元回归）拟合效果

利用 SPSS 的逐步回归模块，拟合土地利用回归模型，将其作为对照组。最终气温、风速及 7km 尺度范围下的林地比例作为目标因子进入模型。多元回归模型的方程为：

$$PM_{2.5}=-4.339\times气温-0.277\times林地\_7km-6.805\times风速+148.637$$



a 训练集　　b 验证集

图 3-6 传统土地利用回归拟合结果散点图

Fig. 3-6 LUR fitting results scatter diagram

经计算， R、RMSE、MAE 以及 IA 分别为 0.651、5.100、3.922、0.759，而验证集对应的指标值分别为 0.647、7.580、5.862、0.702。

（4） 评估指标对比

表 3-4 随机森林优化模型与其它模型的比较(基于验证集)

Table 3-4 comparison of random forest model with other models (based on validation set)

模型名称	R	RMSE	MAE	IA
随机森林	0.831	5.871	4.757	0.854
支持向量机	0.714	6.871	5.521	0.825
土地利用回归	0.647	7.580	5.862	0.702

根据对各参数的了解，R 及 IA 越大，表示拟合的越好，模拟的效果越好；RMSE、MAE 越小，表示模拟的误差越小，模拟的效果越好。

由上表可知，随机森林优化模型的 R 值比支持向量机优化模型和土地利用回归模型分别大 16.4%、28.4%；随机森林优化模型的 IA 值比支持向量机优化模型和土地利用回归模型分别大 3.5%、21.7%；随机森林优化模型的 RMSE 值比支持向量机优化模型和土地利用回归模型分别小 14.6%、22.5%；随机森林优化模型的 MAE 值比支持向量机优化模型和土地利用回归模型分别小 13.8%、18.9%。因此从各项参数来看，随机森林优化模型都要优于另两类模型。利用随机森林优化模型进行长江三角洲 $PM_{2.5}$ 浓度空间模拟具有更好的性能。因此可以证实利用随机森林模型改进 LUR 模型的方法是可取的，该方法模拟的数据从理论上来说具有更高精度，从而可将模拟结果应用于人口 $PM_{2.5}$ 暴露风险评估。

3.2.3 长江三角洲 $PM_{2.5}$ 浓度空间模拟

利用 ArcGIS 的渔网功能生成长江三角洲区域内 3km 边长的格网,同时产生各格网的中心点。以格网中心点（共 11549 个点）为参照点提取各类地理要素，包括 AOD(Terra)、起伏度、高程、降水、气温、气压、相对湿度、风速、7km 缓冲区内的林地比例、5km 缓冲区内的建设用地比例、1km 缓冲区范围内的水体比例、10km 缓冲区范围内的草地比例等 12 种参数。提取方式与训练模型时基本一致。土地利用类的数据在提取过程中由于数据量较大，Fragstats 无法完成该任务，因此利用 GDAL 来进行计算，代码见附录 1。

将 11549 组数据代入训练完成的随机森林优化模型中，预测各点的 $PM_{2.5}$ 浓

度。通过将预测的点的数据链接到对应的格网处，然后将预测的 $PM_{2.5}$ 利用分层设色法进行可视化，即可获得长江三角洲地区 $PM_{2.5}$ 浓度空间分布的基本格局（如图 3-7）。

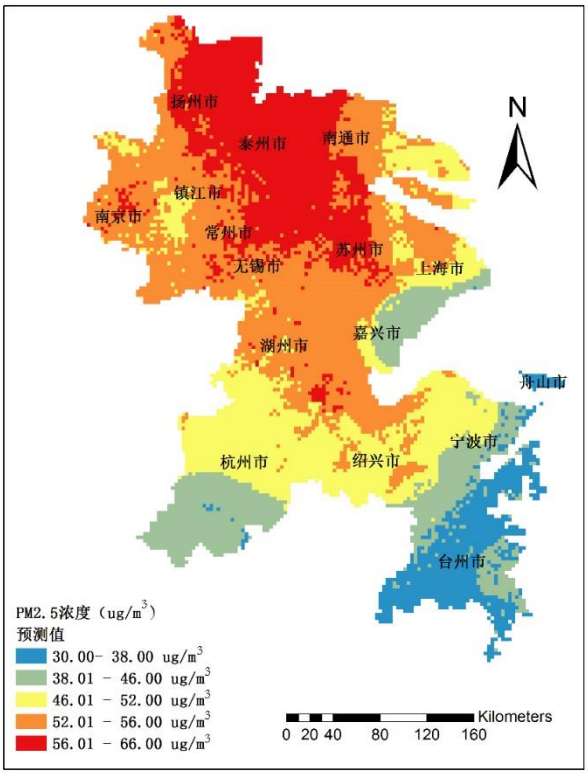


图 3-7 2015 年长江三角洲地区 $PM_{2.5}$ 浓度空间格局

Fig. 3-7 spatial pattern of $PM_{2.5}$ concentration in the Yangtze river delta region in 2015

由图 3-7 可知，模拟的结果具有很明显的空间格局规律，效果较理想。长江三角洲地区 $PM_{2.5}$ 年平均浓度的格局的总体态势为北高南低，西高东低，$PM_{2.5}$ 浓度的高值集聚在部分地区，具有连片分布的特点（赵佳楠等，2018）。这种现象和自然气象因素及人类活动强度有很强的联系。东部沿海地区降水丰富、风速大，利于大气污染物的湿沉降、稀释以及扩散。而西部内陆地区的气象条件不如东部，湿沉降、稀释以及扩散作用弱。南部地区的 $PM_{2.5}$ 浓度状况良好，则一方面是由于植被覆盖好，另一方面是由于人为活动较弱，大气污染颗粒物的排放低。

为了检验模拟的数据是否具有良好的空间信息丰富度，同时定量化的计算地区内各行政单元的 $PM_{2.5}$ 年平均浓度状况，需要对原数据进行数据格式的转换。利用格网中心点的 $PM_{2.5}$ 浓度数据进行插值，获取区域内连续一致的 $PM_{2.5}$ 浓度

数据（图 3-8）。

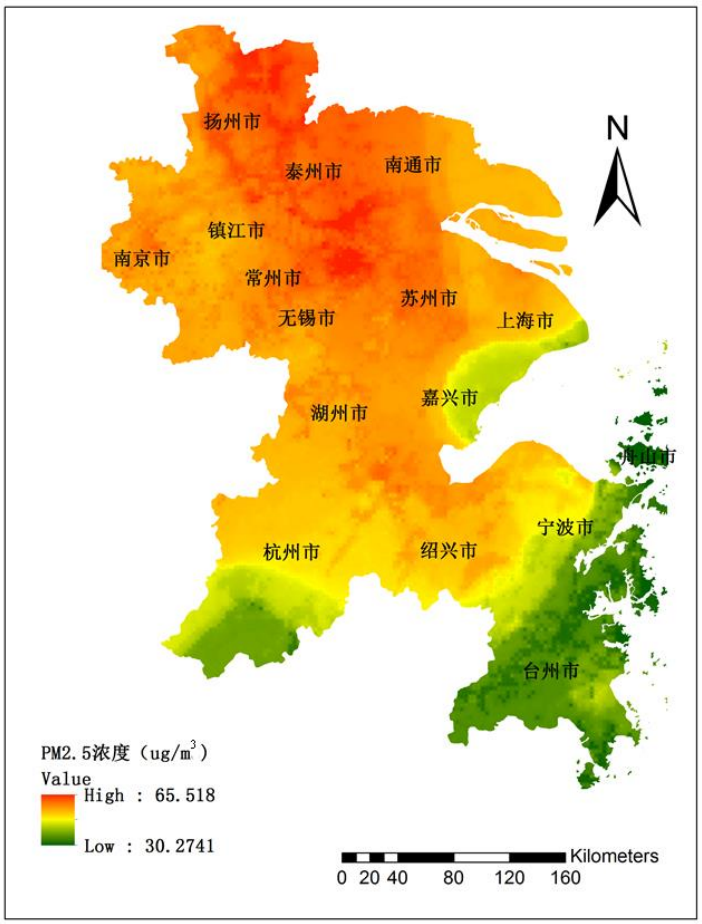


图 3-8 插值后 2015 年长江三角洲地区 $PM_{2.5}$ 浓度分布

Fig. 3-8 the distribution of $PM_{2.5}$ concentration in the Yangtze river delta region in 2015 after the interpolation

空间信息丰富度利用信息熵来评价。信息熵的计算需要一个离散的矢量数据 V，但模型模拟的值为连续的浮点数，需要离散化，因此对 $PM_{2.5}$ 值以四舍五入的方式取整。求解信息熵时利用香农信息熵公式计算，相关代码见附录 1。最终计算得到 2015 年长江三角洲地区 $PM_{2.5}$ 浓度分布数据的信息熵为 4.46。根据信息熵的原理，熵值越大，信息的分异越大。因此该图像具有较明显的分异特征，表明长江三角洲地区的 PM2.5 浓度分布的分异性强，包含的空间信息丰富。

利用 Zonal Statistics as Table 进行分区统计，分区的单位为地级市，下表为各地级市的统计结果。

表 3-5 长江三角洲地区各市 $PM_{2.5}$ 浓度统计量

Table 3-5 statistics of $PM_{2.5}$ concentrations in various cities in the Yangtze river delta region

行政区	最小值（ug/m^3）	最大值（ug/m^3）	平均值（ug/m^3）	标准差（ug/m^3）
舟山市	30.27	42.85	33.05	2.57
台州市	31.41	46.06	36.95	2.37
宁波市	31.41	54.54	42.65	6.37
杭州市	33.84	58.56	47.68	5.11
嘉兴市	41.98	56.48	48.95	4.33
绍兴市	40.99	55.81	50.18	2.95
上海市	36.73	57.62	50.20	3.69
湖州市	50.74	58.43	53.09	1.14
南京市	50.52	58.87	53.43	1.06
镇江市	50.34	63.53	54.27	2.34
南通市	50.90	60.13	54.77	2.45
苏州市	50.96	64.01	55.43	1.70
常州市	51.70	62.76	55.65	2.15
扬州市	51.74	61.21	56.22	2.03
无锡市	50.81	65.52	56.76	2.66
泰州市	55.47	64.89	58.81	1.28

由表可知，江苏省各市的 $PM_{2.5}$ 浓度明显比浙江省和上海市更高，尤其是泰州、无锡、扬州、常州等苏南、苏中地区， $PM_{2.5}$ 年平均浓度的平均值分别达到 58.81 ug/m^3、56.76 ug/m^3、56.22 ug/m^3、55.65 ug/m^3。这主要是由于产业结构偏重，聚集了大量燃煤电厂、石化工业、塑料制造等“三高”企业，是造成 $PM_{2.5}$ 浓度较高的主要原因（戴昭鑫等，2016）。从空间上看，虽然江苏省 $PM_{2.5}$ 浓度总体较高，但部分地区的 $PM_{2.5}$ 浓度呈现明显的“凹地”现象。例如位于镇江市下辖县级市句容，该地区南北分别倚靠茅山及长江，赤山湖位于其城西，宁镇山脉透迤境内，拥有优越的气候条件、良好的植被覆盖以及丰富的森林资源（赵佳楠等，2018）。优良的生态环境使句容的 $PM_{2.5}$ 浓度相比周边地区明显较低。

浙江省内的 $PM_{2.5}$ 浓度较高的地市为湖州、绍兴、嘉兴，$PM_{2.5}$ 年平均浓度的平均值分别达到 53.09 ug/m^3、50.18 ug/m^3、48.95ug/m^3、55.65 ug/m^3，最低的两市则为舟山和台州，均低于 40 ug/m^3。杭州和宁波虽然全市的平均值不高，但由于市内地理要素差异大，也存在高值区，体现在具体数据上表现为标准差较大。杭州东部地区 $PM_{2.5}$ 浓度高，而西部地区的值较低。这是因为杭州的主要工业区分布在杭州东部，如下沙经济技术开发区，大江东产业集聚区，西部地区则生态环境较为优良。宁波杭州湾新区也明显比宁波其它地区 $PM_{2.5}$ 污染严重，原因也

与产业布局有关。该模拟图空间分辨率高，可进行城市内部 $PM_{2.5}$ 空间格局的分析。以杭州市为例，往西南其 $PM_{2.5}$ 浓度变低；$PM_{2.5}$ 浓度的高值区分布在东部地区， $PM_{2.5}$ 浓度最高值出现在中心城区，下沙、萧山及余杭东北部等地 $PM_{2.5}$ 浓度也较高。

已有学者对长江三角洲地区 $PM_{2.5}$ 浓度空间格局进行研究。戴昭鑫等人利用克里金插值来分析长江三角洲地区 $PM_{2.5}$ 浓度空间分布格局，研究结果表明 2013-2015 年期间，长江三角洲三省市当中，江苏省 $PM_{2.5}$ 浓度最高，上海次之，浙江最低。从总体空间变化趋势来看，长江三角洲地区 $PM_{2.5}$ 浓度呈现北部高南部低，局部地区略有突出的特点。毛婉柳等人利用 2015 年长三角地区监测数据，研究长江三角洲地区城市 $PM_{2.5}$，结果显示 2015 年长三角地区城市 $PM_{2.5}$ 年均浓度从江苏到浙江呈减少趋势，特征上表现为北高南低，局部地区突出。浓度值还具有集聚现象，低值集聚在浙江沿海地区，高值则在苏南地区（毛婉柳等，2017；Vapnik，1998；Yang et al.，2017；Lu et al.，2018）。本文的模拟结果与已有研究相符，具有可靠性；且在空间分辨率上较高，可表现长三角 $PM_{2.5}$ 浓度空间格局的更多细节。

3.3 本章小结

综上分析，得到以下结论：

（1）建立随机森林优化模型，需要提取主要的目标因子。在对备选地理因子与 $PM_{2.5}$ 浓度的双变量相关分析中，发现 AOD(Terra)、起伏度、高程、降水、气温、气压、相对湿度、风速、7km 缓冲区内的林地比例、5km 缓冲区内的建设用地比例、1km 缓冲区范围内的水体比例、10km 缓冲区范围内的草地比例与 $PM_{2.5}$ 浓度的相关性较显著。因此利用随机森林优化模型构建时，选用以上目标因子生成模型是合适的。

（2）本文所要使用的随机森林优化模型，经检验集检验模型效果，模型的 IA、MAE 、RMSE 、R 分别为 0.854、4.757、5.871、0.831。而多元回归构建的 LUR 模型，检验集的 IA、MAE 、RMSE 、R 分别为 0.702 、5.862、7.58、0.647。随机森林优化模型比多元回归构建的模型在 IA 和 R 两个参数上要明显更大，而 MAE 、RMSE 两个参数则明显更小，因此随机森林优化模型效果更好。对比于使用广泛的支持向量机算法，随机森林优化模型同样具有更好的效果。支

持向量机优化模型的检验集计算的 IA、MAE 、RMSE 、R 分别为 0.825、5.521、6.871、0.714，而训练集的 IA、MAE 、RMSE 、R 则表现为误差极小，因此支持向量机优化模型存在较明显的过拟合现象。

（3）经过模型模拟及基本的统计量计算，长三角 $PM_{2.5}$ 浓度空间分布具有下列特征：呈现北部高南部低，西部高东部低的总体分布格局；高值集聚在部分地区，具有连片分布的特点；江苏省各市的 $PM_{2.5}$ 浓度明显比浙江省和上海市更高，尤其是泰州、无锡、扬州、常州等苏南、苏中地区；浙江省内的 $PM_{2.5}$ 浓度较高的地市为湖州、绍兴、嘉兴，最低的两市则为舟山和台州；杭州和宁波虽然全市的平均值不高，但也存在高值区，例如杭州东部地区 $PM_{2.5}$ 浓度高，宁波杭州湾新区也明显比宁波其它地区 $PM_{2.5}$ 污染严重。

第四章 长江三角洲人口 $PM_{2.5}$ 暴露风险评价

4.1 人口 $PM_{2.5}$ 暴露风险评价体系

目前，常用的大气污染防控分区主要依赖空气质量浓度、人口暴露强度以及人口加权浓度三类大气污染暴露风险评估方法。其中空气质量假设评估区域人口分布均一，评估存在一定的理论偏差；人口暴露强度同时考虑空气质量浓度空间分布和人口空间分布，使用栅格运算的方法，理论上提高了评估结果的精度，但在较大空间范围内存在各评估单元间风险值无法比较的问题；人口加权浓度的人口空气污染暴露风险评价指标是一种顾及人口空间分布的暴露评估方法，可在较大空间范围内评估各评估单元间的人口空气污染暴露风险，但需要先确定评估单元，然后进行统计，在空间分辨率上会有欠缺。人口暴露强度以及人口加权浓度这两种方法最显著的区别在于前者是基于格网的，后者则基于区域（例如城市、城市群等）。

三种方法都具有各自的优势与劣势，因此结合三种指标可弥补单一指标的问题，利于全面分析长江三角洲人口 $PM_{2.5}$ 暴露风险问题。空气质量浓度、人口暴露强度可以分析市域内的细节，人口加权浓度主要用于比较地市之间的风险值差异。

4.1.1 $PM_{2.5}$ 空气质量浓度指标

$PM_{2.5}$ 空气质量浓度指标以 $PM_{2.5}$ 浓度来表示居民暴露于其中的风险（式 4-1）。一般借助于空间插值的方法从监测站点获取的 $PM_{2.5}$ 浓度离散值来得到覆盖整个研究区的连续的 $PM_{2.5}$ 浓度表面。

$$E_i = \rho_i \tag{4-1}$$

式中：i 表示空间单元（栅格单元）编号，ρ_i 为空间单元 i 处的 $PM_{2.5}$ 浓度，E_i 为空间单元 i 处的人口暴露风险。

本文已对 $PM_{2.5}$ 浓度的区域模拟方法进行了改进，具有较高的精度及分辨率，因此采用模拟的结果代替较传统的克里金插值来进行 $PM_{2.5}$ 浓度的空间化。数据为 100m×100m 空间分辨率的基于 $PM_{2.5}$ 空气质量浓度的人口 $PM_{2.5}$ 暴露风险图。

4.1.2 $PM_{2.5}$ 人口暴露强度指标

$PM_{2.5}$ 人口暴露强度指标采用同一空间尺度，对人口密度和 $PM_{2.5}$ 浓度进行

乘积来表示居民暴露于 $PM_{2.5}$ 污染的风险。$PM_{2.5}$ 人口暴露强度指标是基于格网的，每个格网都存在一个暴露风险值，该风险值由该网格的人口密度及与 $PM_{2.5}$ 空气质量浓度相乘得到（式 4-2）。

$$E_i = P_i \rho_i \tag{4-2}$$

式中：P_i 表示空间格网 i 处的人口密度。本文利用栅格计算器求长江三角洲人口密度图与 $PM_{2.5}$ 空气质量浓度图之积，获取 100m×100m 空间分辨率的长江三角洲人口 $PM_{2.5}$ 暴露风险图。

4.1.3 $PM_{2.5}$ 人口加权暴露浓度指标

$PM_{2.5}$ 人口加权暴露浓度指标指的是计算区域内居民暴露于 $PM_{2.5}$ 污染下的风险一种方式，能够直观的比较区域间人口 $PM_{2.5}$ 暴露风险的严重程度。同一空间尺度下，空间格网人口占空间单元整体的比重和 $PM_{2.5}$ 浓度之积来表示单个格网的暴露风险贡献值，将区域内各格网的贡献值求和，即为区域的 $PM_{2.5}$ 人口加权暴露浓度（式 4-3），

$$E_i = \frac{Pop_i \rho_i}{\sum_{i=1}^{n} Pop_i} \tag{4-3}$$

式中：Pop_i 为空间单元 i 内的人口数。

利用人口加权的方式，以市为区域单元，可计算得到各市的 $PM_{2.5}$ 人口加权暴露浓度，从而对各市人口 $PM_{2.5}$ 暴露风险的严重程度进行比较。

4.2 长江三角洲人口 $PM_{2.5}$ 暴露风险

4.2.1 $PM_{2.5}$ 浓度下人口 PM2.5 暴露风险

图 4-1 为基于 $PM_{2.5}$ 浓度的长江三角洲地区人口 $PM_{2.5}$ 暴露风险空间分布（同 $PM_{2.5}$ 浓度空间模拟图）。从该图来看，长江三角洲区域内年均 $PM_{2.5}$ 浓度的平均值为 50.42 ug/m^3。区域内绝大部分地级市及上海市的 $PM_{2.5}$ 浓度均超过我国 2012 版年平均二级标准值为 35 ug/m^3，仅舟山市在此范围内，空气质量优良。其中苏州市、常州市、扬州市、无锡市、泰州市等 5 个地市均超过了 55 ug/m^3，是最为严重的地市。从长江三角洲区域整体来看，地区人口 PM2.5 暴露绝对风险从南往北，从东往西呈现梯度递增的现象。

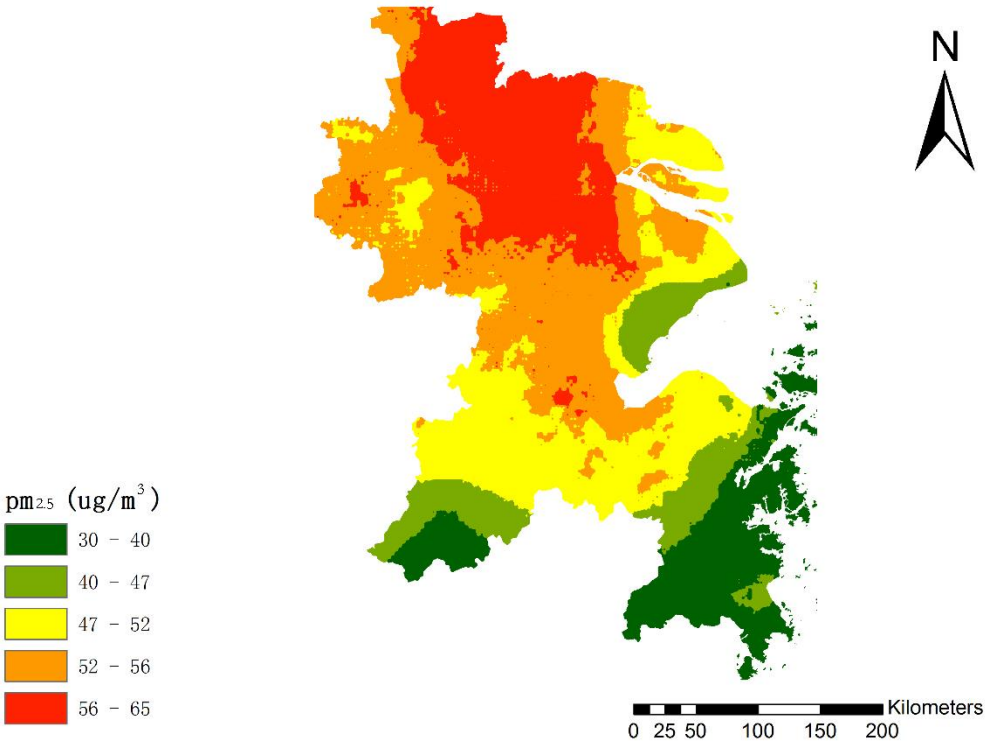


图 4-1　$PM_{2.5}$ 浓度下人口 $PM_{2.5}$ 暴露风险

Fig. 4-1 Exposure risk of $PM_{2.5}$ concentration under $PM_{2.5}$ concentration

因为模拟的数据的空间分辨率高，因此可以对市域内的人口 $PM_{2.5}$ 暴露风险进行分析。基于 $PM_{2.5}$ 浓度的空间分布情况，杭州市人口 $PM_{2.5}$ 暴露风险大致具有中心城区向外围递减的趋势，且西南区域的风险要比东北方向弱（图 4-1）。这与西南部的淳安、建德、临安、桐庐等地生态环境好，城市化、工业化水平相对较低有关。该区域植被覆盖度高，可滞尘，消减 $PM_{2.5}$；分布着新安江水库，富春江等优质水体，可有效较低 $PM_{2.5}$。

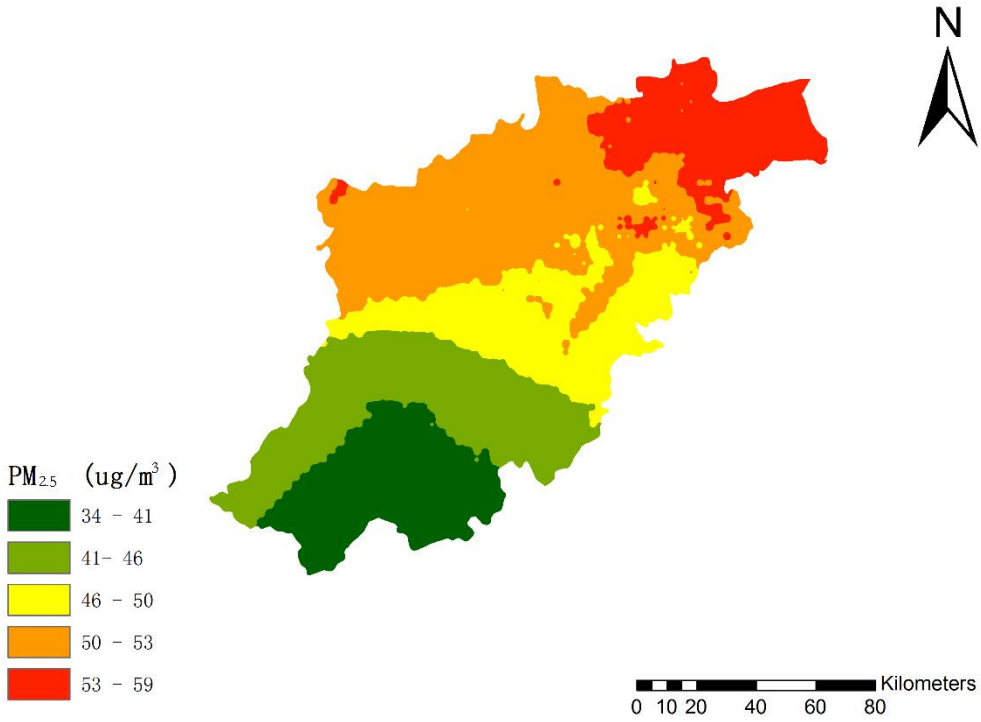


图 4-2 杭州市范围内 $PM_{2.5}$ 浓度下人口 $PM_{2.5}$ 暴露风险

Fig. 4-2 the risk of PM2.5 exposure at the concentration of PM2.5 in Hangzhou

对杭州市区（包括萧山、余杭）各区的 $PM_{2.5}$ 浓度下人口 $PM_{2.5}$ 暴露风险进行统计，结果如表 4-1。各城区中西湖区的人口 $PM_{2.5}$ 暴露风险最低，但标准差却最大，说明行政区内存在较大的差异，部分区域风险小，但另一些区域仍然存在较大风险。下城区人口 $PM_{2.5}$ 暴露风险最大，其次是拱墅区。江干区人口 $PM_{2.5}$ 暴露风险排在八区中的第三，且标准差最小，说明江干区整体而言风险较大。这表现了各城区人口 $PM_{2.5}$ 暴露风险存在明显的差异性，可针对性的开展 $PM_{2.5}$ 污染的防控。

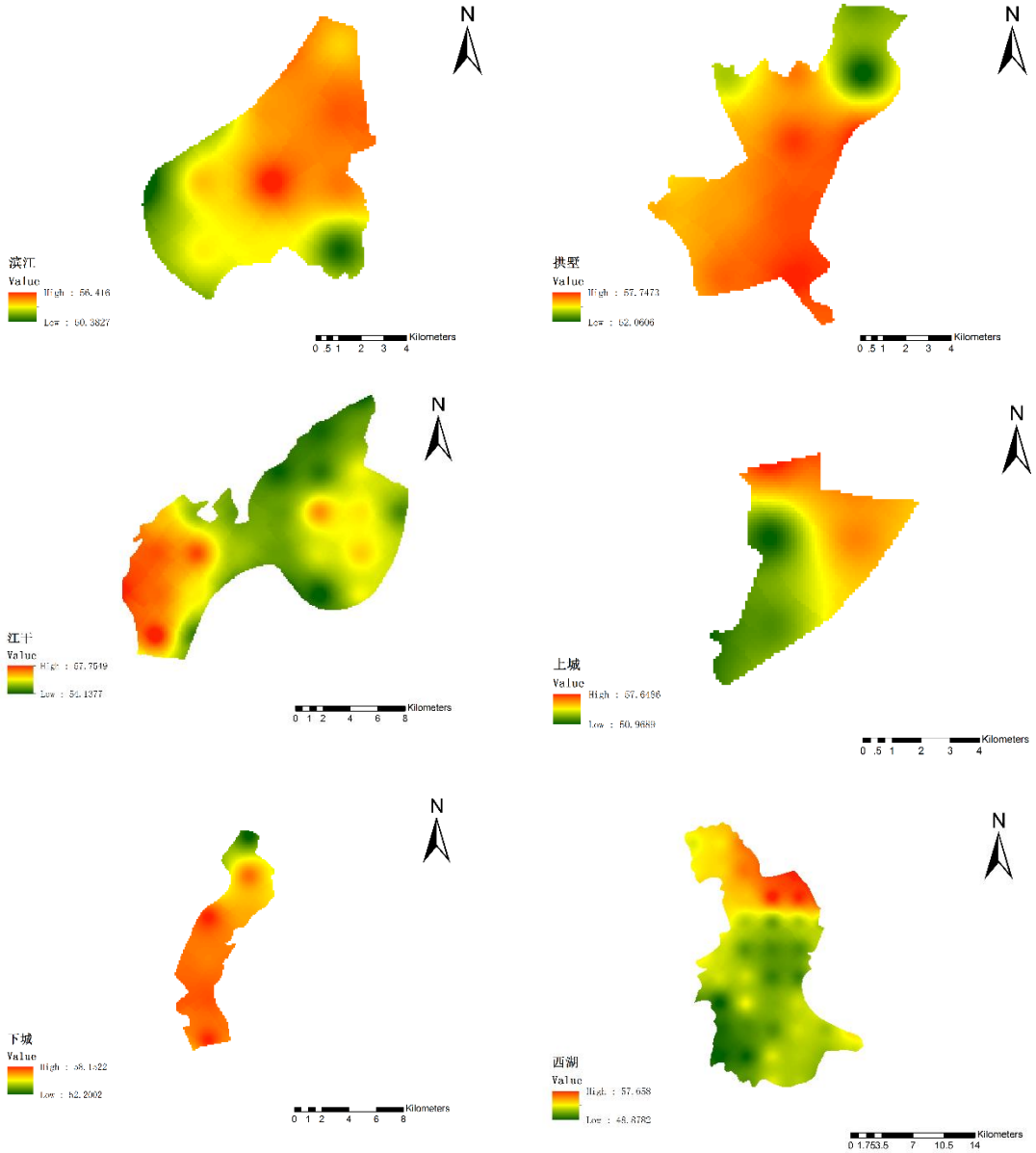
表 4-1 杭州城区各区 $PM_{2.5}$ 浓度下人口 $PM_{2.5}$ 暴露风险统计

Table 4-1 statistics of exposure risk of PM2.5 concentration in various districts of Hangzhou city

行政区	最小值(ug/m^3)	最大值(ug/m^3)	平均值(ug/m^3)	标准差
西湖区	48.88	57.66	52.28	1.85
萧山区	48.63	58.56	53.30	1.74
余杭区	50.05	58.44	53.31	1.47
上城区	50.97	57.78	54.01	1.66
滨江区	50.32	56.42	54.15	1.14
江干区	54.14	57.75	55.65	0.80
拱墅区	52.06	57.75	56.10	1.24

下城区	52.20	58.15	56.63	1.13

为了解各行政区内部的人口暴露风险分布，对各区的人口暴露风险分别制图（图 4-3）。在了解城市形态及周边地理环境的状况下，可对此做更细致的分析。例如萧山区人口 $PM_{2.5}$ 暴露风险具有明显的空间规律，中心城区是主要的风险高值区，其次为沿江的部分地区，上述地区应该重点监测及控制 $PM_{2.5}$ 污染。总体上，东北地区相比西南地区的暴露风险要明显高。对两地的地理环境进行分析可知西南地区生态环境远好于东北地区，因此应该着力提高东北地区的生态环境以降低人口的 $PM_{2.5}$ 暴露风险。



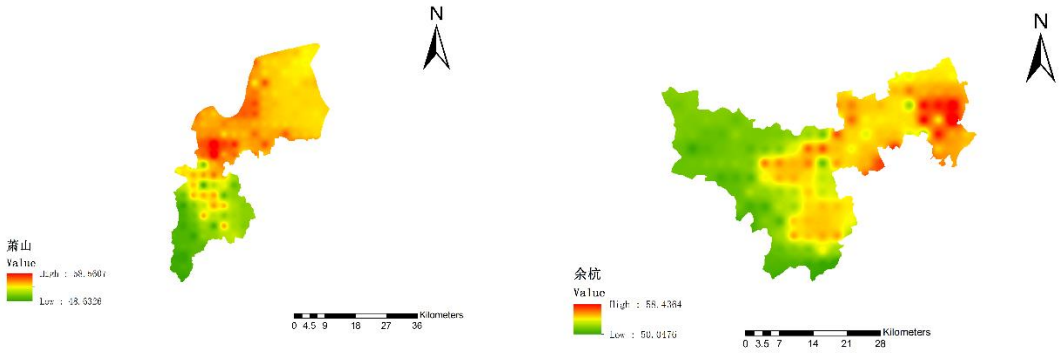


图 4-3 杭州城区各区内 $PM_{2.5}$ 浓度下人口 $PM_{2.5}$ 暴露风险

Figure 4-3 the exposure risk of $PM_{2.5}$ concentration in the urban areas of Hangzhou

4.2.2 人口暴露强度下人口 PM2.5 暴露风险

图 4-4 中右图为基于人口分布下的人口 $PM_{2.5}$ 暴露风险。该图表明各城市中心城区是人口 $PM_{2.5}$ 暴露风险高值区。大部分地区的暴露风险值低于 800 人.ug.10^{-4}.m^{-5}，而中心城区则超过 1600 人.ug.10^{-4}.m^{-5}。整体上看，人口 $PM_{2.5}$ 暴露风险与人口分布呈现高度的空间一致性。但由于区域范围较大，进行长三角尺度下各个地区之间的比较，可能与实际情况存在偏差。

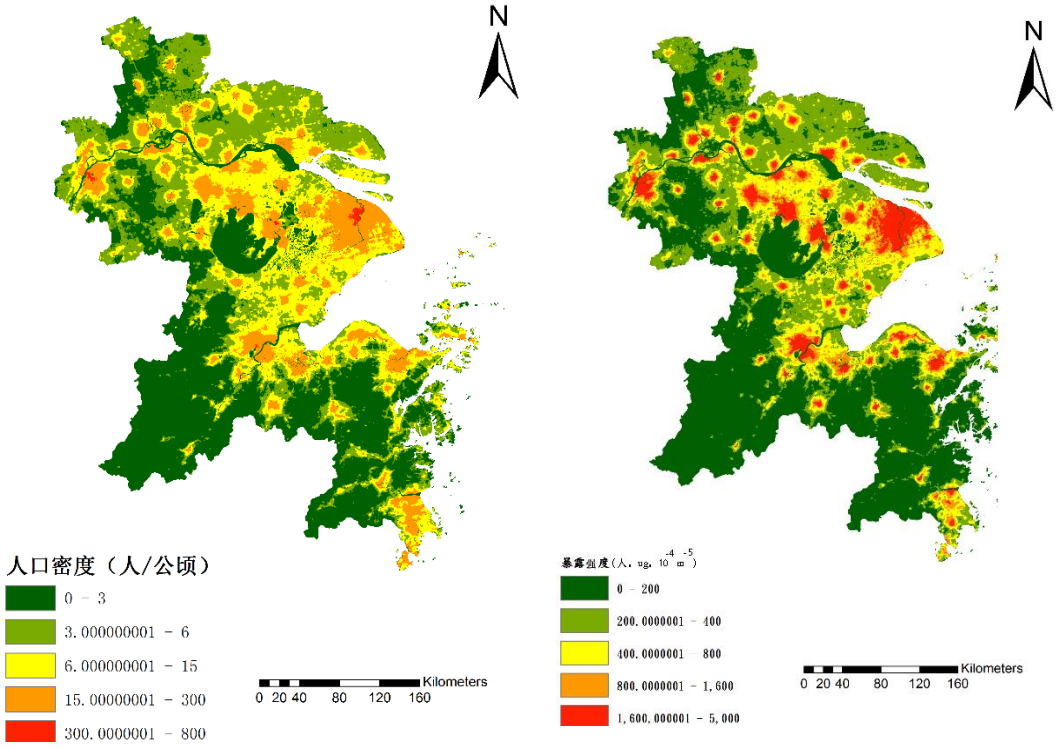


图 4-4 长江三角洲人口密度分布及基于人口分布下的人口 $PM_{2.5}$ 暴露风险

Fig. 4-4 population density distribution in the Yangtze river delta and the risk of PM2.5 exposure based on population distribution

将尺度缩小到市域内，在杭州市范围内呈现以下特征（图 4-5）：人口 $PM_{2.5}$ 暴露风险近似呈现同心圆结构，中心往外减弱；中心城区呈现连片的高风险区；在中心城区外围的副城具有斑块状的高风险区；城市内部具有一些轴状的相对高风险区，如中心城区与各副城之间的主要交通路线及钱塘江沿线带。

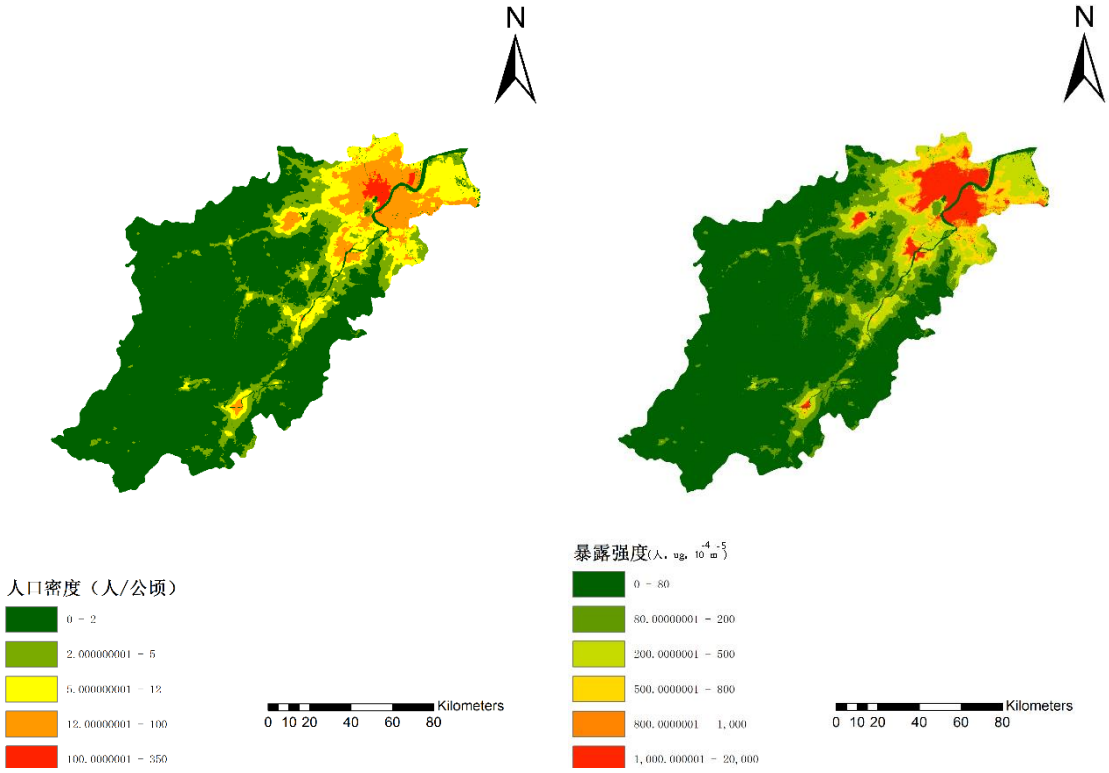


图 4-5 杭州市人口密度分布及基于人口分布下的人口 $PM_{2.5}$ 暴露风险

Fig. 4-5 population density distribution in hangzhou and the risk of PM2.5 exposure based on population distribution

4.2.3 人口加权下人口 $PM_{2.5}$ 暴露风险

人口加权平均 $PM_{2.5}$ 暴露风险适合评估空气污染对公众健康的影响,并为单个城市群及跨区域合作污染减排政策研究提供研究视角。

利用前面所述的人口加权视角的人口 $PM_{2.5}$ 暴露风险计算方法，以城市为基本单元计算各单元人口 $PM_{2.5}$ 暴露风险。相关代码见附录 1，最终计算结果整理为表 4-2。

表 4-2 基于人口加权的人口 $PM_{2.5}$ 暴露风险

Table 4-2 is based on the population weighted population PM2.5 exposure risk

城市	人口加权下人口 $PM_{2.5}$ 暴露风险(ug/m^3)	排名	与 $PM_{2.5}$ 平均浓度的差(ug/m^3)
泰州	59.05	1	0.24
无锡	58.44	2	1.68
常州	57.35	3	1.7
扬州	56.35	4	0.13
苏州	56.04	5	0.61
南通	55.09	6	0.32
南京	54.99	7	1.54
镇江	54.93	8	0.66
湖州	53.83	9	0.74
杭州	53.27	10	5.59
绍兴	52.22	11	2.04
上海	52.12	12	1.92
嘉兴	49.01	13	0.06
宁波	46.03	14	3.38
台州	37.71	15	0.76
舟山	33.24	16	0.19

泰州、无锡、常州为风险值最大的三个地市，宁波、台州、舟山三个地市为风险值最小的。人口加权下，风险值大，说明该地区 $PM_{2.5}$ 浓度整体较高，且人口集中于地区内 $PM_{2.5}$ 浓度高的区域，对当地居民造成的健康危害更大。对各城市 2015 年人口加权的人口 $PM_{2.5}$ 暴露风险与 2015 年该城市 $PM_{2.5}$ 平均浓度求差，各城市所得的值均大于 0。其中杭州、宁波、绍兴、上海位列前四，绝对值分别为 5.59、3.38、2.04、1.92。这说明这几个地区风险值虽不如泰州、无锡、常州等地，但是人口更集聚于区域内 $PM_{2.5}$ 高值区。对这些区域内的高值区 $PM_{2.5}$ 采取防控措施，对于降低地区的 $PM_{2.5}$ 人口暴露风险成效更加明显。

结合 $PM_{2.5}$ 空气质量浓度、人口暴露强度及人口加权平均，既有基于区域的人口暴露风险评估，又有高空间分辨率的基于格网的人口暴露风险评估。利用 $PM_{2.5}$ 空气质量浓度、人口暴露强度指标，对长江三角洲地区的人口暴露风险可初步进行空间格局的分析；通过人口加权平均的 $PM_{2.5}$ 人口暴露风险的评估可对各行政单元的 $PM_{2.5}$ 人口暴露风险进行对比，筛选 $PM_{2.5}$ 人口暴露风险重点防控城市，此外结合 $PM_{2.5}$ 空气质量浓度，可提取人口高度集中于 $PM_{2.5}$ 高值区的城市；利用人口暴露强度可进一步监测重点防控城市的重点区域，适合于城市内部的区域暴露风险评估。

4.3 本章小结

综合以上分析，可得出以下结论：

（1）空气质量假设评估区域内人口分布均一，评估存在一定的理论偏差；人口暴露强度同时考虑空气质量浓度空间分布和人口空间分布，使用栅格运算的方法，理论上提高了评估结果的精度，但在较大空间范围内存在各评估单元间风险值无法比较的问题；人口加权浓度的人口空气污染暴露风险评价指标是一种顾及人口空间分布的暴露评估方法，可较大空间范围内评估各评估单元间的人口空气污染暴露风险，但需确定评估单元进行统计，在空间分辨率上会有欠缺。结合三种指标可弥补单一指标的问题，利于全面分析长江三角洲人口 $PM_{2.5}$ 暴露风险问题。

（2）基于 $PM_{2.5}$ 空气质量浓度，长江三角洲区地区人口 $PM_{2.5}$ 暴露风险从南往北，从东往西呈现梯度递增的现象。该模拟数据精度高，以杭州为例进行市域的分析，杭州市人口 $PM_{2.5}$ 暴露风险大致具有中心城区向外围递减的趋势，且西南区域的风险要比东北方向弱。

（3）基于人口分布下的人口 $PM_{2.5}$ 暴露风险，从区域尺度分析，各城市中心城区是人口 $PM_{2.5}$ 暴露风险高值区；大部分地区的暴露风险值低于 800 人.ug.10^{-4}.m^{-5}，而中心城区超过 1600 人.ug.10^{-4}.m^{-5}；人口 $PM_{2.5}$ 暴露风险与人口分布呈现高度的空间一致性。从市域分析，以杭州为例，人口 $PM_{2.5}$ 暴露风险近似呈现同心圆结构，中心往外减弱；中心城区呈现连片的高风险区；在中心城区外围的副城具有斑块状的高风险区；城市内部具有一些轴状的相对高风险区，如中心城区与各副城之间的主要交通路线及钱塘江沿线带。

（4）人口加权平均 $PM_{2.5}$ 暴露风险适合评估单元间空气污染对公众健康的影响强度。泰州、无锡、常州为风险值最大的三个地市，宁波、台州、舟山三个地市为风险值最小的。应加大对泰州、无锡、常州等市的 PM2.5 暴露风险的防控，此外杭州、宁波、绍兴、上海的人口高度集中于 $PM_{2.5}$ 浓度高值区，需要重点关注。

（5）利用 $PM_{2.5}$ 空气质量浓度、人口暴露强度指标，对长江三角洲地区的人口暴露风险可初步进行空间格局的分析；以人口加权平均的 $PM_{2.5}$ 人口暴露风险的评估可对各行政单元的 $PM_{2.5}$ 人口暴露风险进行对比，筛选 $PM_{2.5}$ 人口暴露

风险重点防控城市；利用人口暴露强度可进一步监测重点防控城市的重点区域，适合于城市内部的区域暴露风险评估。综合这三类指标，可为大气污染治理和人居环境的提高提供重要的科学依据。

第五章 结论与展望

5.1 结论

基于目前对 $PM_{2.5}$ 空间分布模拟及 $PM_{2.5}$ 人口暴露风险的研究,本文综合利用多源数据，借助于随机森林改进的土地利用回归模型（随机森林优化模型）对长江三角洲地区的 $PM_{2.5}$ 浓度进行空间模拟且对该地区的 $PM_{2.5}$ 人口暴露风险进行评估。

（1）对备选地理因子与 $PM_{2.5}$ 浓度进行双变量相关分析， AOD(Terra)、起伏度、高程、降水、气温、气压、相对湿度、风速、7km 缓冲区内的林地比例、5km 缓冲区内的建设用地比例、1km 缓冲区范围内的水体比例、10km 缓冲区范围内的草地比例与 $PM_{2.5}$ 浓度的相关性较显著。因此进行随机森林优化模型构建时，可选用以上目标因子。

（2）本文使用随机森林优化模型进行长江三角洲 $PM_{2.5}$ 浓度空间模拟，经检验集检验模型效果，模型的 IA、MAE 、RMSE 、R 分别为 0.854、4.757、5.871、0.831。为凸显随机森林优化模型的优势，本文也同时对 LUR 模型、支持向量机（SVM）优化模型进行效果检验。使用多元回归的传统土地利用回归模型常用，检验集的 IA、MAE 、RMSE 、R 分别为 0.702 、5.862、7.58、0.647。经计算，使用 SVM 改进的土地利用回归模型(SVM 优化模型)，其检验集的 IA、MAE 、RMSE 、R 分别为 0.825、5.521、6.871、0.714。由于随机森林优化模型的检验集的 IA 和 R 较大，而 MAE 和 RMSE 较小，因此进行 $PM_{2.5}$ 浓度空间模拟时效果更好，本文使用随机森林优化模型来进行模拟是可取的。

（3）分析模型模拟结果，长三角 $PM_{2.5}$ 浓度空间分布具有下列特征：其分布格局呈现北高南低，西高东低的特征；高值集聚，连片分布；江苏省各市的 $PM_{2.5}$ 浓度明显比浙江省和上海市更高，尤其是泰州、无锡、扬州、常州等苏南、苏中地区；浙江省内的 $PM_{2.5}$ 浓度较高的地市为湖州、绍兴、嘉兴，最低的两市则为舟山和台州；杭州和宁波整体均值不高，但存在明显的相对高值区，例如杭州东部地区 $PM_{2.5}$ 浓度高，宁波杭州湾新区则显著的比宁波其它地区 $PM_{2.5}$ 污染严重。

（4）基于 $PM_{2.5}$ 空气质量浓度及人口分布下的人口 PM2.5 暴露风险，长江三角洲区地区人口 $PM_{2.5}$ 暴露风险从南部往北部，从东部往西部呈现梯度递增的

现象，各城市中心城区是人口 $PM_{2.5}$ 暴露风险高值区。从市域分析，以杭州为例，人口 $PM_{2.5}$ 暴露风险大致具有同心圆结构，中心往外减弱，且西南区域的风险要比东北方向弱；中心城区呈现连片的高风险区，在中心城区外围的副城具有斑块状的高风险区；城市内部具有一些轴状的相对高风险区，如主要交通路线及钱塘江沿线带。而基于人口加权平均 $PM_{2.5}$ 暴露风险，可评估单元间空气污染对公众健康的影响强度。泰州、无锡、常州为风险值最大的三个地市，应加大对泰州、无锡、常州等市的 $PM_{2.5}$ 暴露风险的防控，此外杭州、宁波、绍兴、上海的人口高度集中于 $PM_{2.5}$ 浓度高值区，需要重点关注。

（5）利用 $PM_{2.5}$ 空气质量浓度、人口暴露强度指标，对长江三角洲地区的人口暴露风险可初步进行空间格局的分析；以人口加权平均的 $PM_{2.5}$ 人口暴露风险的评估可对各行政单元的 $PM_{2.5}$ 人口暴露风险进行对比，筛选 $PM_{2.5}$ 人口暴露风险重点防控城市；利用人口暴露强度可进一步监测重点防控城市的重点区域，适合于城市内部的区域暴露风险评估。综合以上三类指标，对于大气污染治理和人居环境的提高具有重要的实际意义。

5.2 特色与展望

本文主要有以下特色：

（1）从研究数据来看，本文采用多源数据进行 $PM_{2.5}$ 浓度空间模拟，除土地利用回归模拟常用的气象数据、土地利用数据、气象数据外，引入 AOD 数据。即利用了 AOD 数据对于 PM2.5 模拟的良好效果，同时又避免了使用 AOD 单一模拟数据源时空间覆盖度不够的问题。

（2）从研究方法来看，本文对 $PM_{2.5}$ 浓度空间模拟引入了随机森林方法。基于随机森林方法训练土地利用回归模型（随机森林优化模型），模型运算效率高，具有更高的精度，此外还可避免多重共线性及过拟合问题。对 $PM_{2.5}$ 人口暴露风险，本文利用随机森林优化模型模拟的高精度 $PM_{2.5}$ 浓度空间分布数据进行评估，且利用多指标评价方法从多个角度对长江三角洲地区的 $PM_{2.5}$ 人口暴露风险进行评估。该评估方式的可信度高。

但本文也有不足之处，主要为以下两点：

（1）由于难以获得高空间分辨率的社会经济类的指标，在 PM2.5 浓度空间模拟模型中该类指标较少，对模型的精度会有一定的影响。在今后的研究中有必

要引入。

（2）本研究模拟结果为年均值，未考虑季节因素，忽视了不同季节的 $PM_{2.5}$ 浓度模拟效果。此外时间跨度上不足，仅模拟了 2015 年的长江三角洲 $PM_{2.5}$ 浓度空间分布，无法进一步进行分析 $PM_{2.5}$ 浓度空间分布的时间变化。下一步需进行时间尺度上的拓展。

参考文献

[1]Appel K W，Bhave P V，Gilliland A B，et al． Evaluation of the community multiscale air quality(CMAQ) model version 4.5: Sensitivities impacting model performance; Part Ⅱ—Particulate matter[J]． Atmospheric Environment，2008，42(24) : 6057－6066．

[2]Brunsdon C, Fotheringham A S, Charlton M E. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity[J]. Geographical Analysis, 1996, 28(4):281-298. [61] Vapnik V.1998. Statistical Learning Theeory[M]. New York, John Wiley.

[3]Cao Junji, Li Shuncheng, Cao Junji, et alResearch progress on air pollution exposure［J］. Environmental Pollution ＆ Control，2005，27(2) : 118－122.

[4]Clougherty J E, Wright R J, Baxter L K, et al. Land use regression modeling of intra-urban residential variability in multiple traffic-related air pollutants[J]. Environmental Health, 2008, 7(1):1-14.

[5]Fotheringham A S, Charlton M E, Brunsdon C. Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis[J]. Environment & Planning A, 1998, 30(11):1905-1927.

[6]Gulliver J, Hoogh K D, Fecht D, et al. Comparative assessment of GIS-based methods and metrics for estimating long-term exposures to air pollution[J]. Atmospheric Environment, 2011, 45(39):7072-7080.

[7]Hochadel M, Heinrich J, Gehring U, et al. Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information[J]. Atmospheric Environment, 2006, 40(3):542-553.

[8]Hoek G， Beelen R， De Hoogh K， et al．A review of land - use regression models to assess spatial variation of outdoor air pollution [J]. Atmospheric Environment， 2008，42(33)：7561-7578．

[9]Hoek G, Beelen R, Kos G, et al. Land use regression model for ultrafine particles in Amsterdam.[J]. Environmental Science & Technology, 2011, 45(2):622.

[10]Hoek G, Brunekreef B, Goldbohm S, et al. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study.[J]. Lancet, 2002, 360(9341):1203-1209.

[11]Hoek G, Hoogh B K D, Vienneau D, et al. A Review Of Land-use Regression Models To Assess Spatial Variation Of Outdoor Air Pollution[J]. Atmospheric Environment, 2008, 42(33):7561-7578.

[12]Hoek G, Hoogh B K D, Vienneau D, et al. A Review Of Land-use Regression Models To Assess Spatial Variation Of Outdoor Air Pollution[J]. Atmospheric Environment, 2008, 42(33):7561-7578.

[13]Johnson M, Isakov V, Touma J S, et al. Evaluation of land-use regression models used to predict air quality concentrations in an urban area[J]. Atmospheric Environment, 2010, 44(30):3660-3668.

[14]Kousa A, Oglesby L, Koistinen K, et al. Exposure chain of urban air $PM_{2.5}$—associations between ambient fixed site, residential outdoor, indoor, workplace and personal exposures in four European cities in the EXPOLIS -study[J]. Atmospheric Environment, 2002, 36(18):3031-3039.

[15]Lim S, Vos T, Bruce N. 'The burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions 1990-2010: a systematic analysis'[J]. Lancet, 2012, 380(9859):2224-60.

[16]Liu Y, Paciorek C J, Koutrakis P. Estimating Regional Spatial and Temporal Variability of $PM_{2.5}$ Concentrations Using Satellite Data, Meteorology, and Land Use Information[J]. Environ Health Perspect, 2009, 117(6):886-892.

[17]Lu D, Mao W, Yang D, et al. Effects of land use and landscape pattern on $PM_{2.5}$, in Yangtze River Delta, China[J]. Atmospheric Pollution Research, 2018.

[18]Lu D, Xu J, Yang D, et al. Spatio-temporal variation and influence factors of $PM_{2.5}$ concentrations in China from 1998 to 2014[J]. Atmospheric Pollution Research, 2017.

[19]Manders A，Schaap M，Hoogerbrugge R．Testing the capability of the chemistry transport model LOTOS-EUROS to forecast PM_{10} levels in the Netherlands[J]．Atmospheric Environment，2009，43(26) : 4050－4059．

[20]Manders A，Schaap M，Hoogerbrugge R．Testing the capability of the chemistry transport model LOTOS-EUROS to forecast PM10 levels in the Netherlands[J]．Atmospheric Environment，2009，43(26) : 4050－4059．

[21]Olvera H A, Garcia M, Li W W, et al. Principal component analysis optimization of a $PM_{2.5}$ land use regression model with small monitoring network[J]. Science of the Total Environment, 2012, 425(3):27-34.

[22]Pfister G G，Emmons L K，Hess P G，et al．Contribution of isoprene to chemical budgets: A model tracer study with the NCAR CTM MOZART-4[J]．Journal of Geophysical Research: Atmospheres(1984－2012) ，2008，113(D5) : 308－328．

[23]Remen A L, Chambless D L, Steketee G, et al. Second-generation Operational Algorithm：Retrieval of Aerosol Properties over Land from Inversion of Moderate Resolution Imaging Spectroradiometer Spectral Reflectanc[J].Journal of Geophysical Research，2007，112：doi：10．1029／2006JD007811．

[24]Roorda-Knape M C, de Hartog J J, Phn V V, et al. Air pollution from traffic in city districts near major motorways[J]. Atmospheric Environment, 1998, 32(97):1921–1930.

[25]Ross Z, Jerrett M, Ito K, et al. A land use regression for predicting fine particulate matter concentrations in the New York City region[J]. Atmospheric Environment, 2007, 41(11):2255-2269.

[26] Moore D K, Jerrett M, Mack W J, et al. A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA[J]. J Environ Monit, 2007, 9(3):246-252.

[27] Ross Z， Jerrett M， Ito K， et al．A land use regression for predicting fine particulate matter concentrations in the New York City region [J]. Atmospheric Environment，2007，41 (11)： 2255-2269．

[28] Walsh E S, Kreakie B J, Cantwell M G, et al. A Random Forest approach to predict the spatial distribution of sediment pollution in an estuarine system[J]. Plos One, 2017, 12(7):e0179473.

[29] Wang Shuxiao, Zhao Yu, Chen Gangcai, et al. Assessment of population exposure to particulate matter pollution in Chongqing, China[J] . Environmental Pollution . 2007 (1).

[30] Wang Y, Wu G, Deng L, et al. Prediction of aboveground grassland biomass on the Loess Plateau, China, using a random forest algorithm[J]. Scientific Reports, 2017, 7(1):6940.

[31] Wilson J G, Zawar-Reza P. Intraurban-scale dispersion modelling of particulate matter concentrations: Applications for exposure estimates in cohort studies[J]. Atmospheric Environment, 2006, 40(6):1053-1063.

[32] Yang D, Lu D, Xu J, et al. Predicting spatio-temporal concentrations of $PM_{2.5}$, using land use and meteorological data in Yangtze River Delta, China[J]. Stochastic Environmental Research & Risk Assessment, 2017(9):1-12.

[32]Yang, D., Lu, D., Xu, J. et al. Stoch Environ Res RiskAssess(2017).https://doi.org/10.1007/s00477-017-1497-6.

[33]Zou，B.，Wilson，J．G.，Zhan，F．B.，et a1．Air Pollution Exposure Assessment Methods Utilized in Epidemiological Studies[J]． Journal of Environmental Monitoring，2009，11：475-490．

[34]车瑞俊,刘大锰,袁杨森.北京冬季大气颗粒物污染水平和影响因素研究[J].中国科学院研究生院学报,2007(05):556-563.

[35]陈凯, 刘凯, 柳林, 朱远辉. 2015. 基于随机森林的元胞自动机城市扩展模拟——以佛山市为例[J]. 地理科学进展, 34(8): 937-946

[36]陈莉， 白志鹏， 苏笛， 等. 利用 LUR 模型模拟天津市大气污染物浓度的空间分布[J]. 中国环境科学， 2009， 29（7）： 685-691.

[37]陈训来，冯业荣，范绍佳，等．离岸型背景风和海陆风对珠江三角洲地区灰霾天气的影响[J]． 大气科学，2008，3: 530－542．

[38]成海容， 王祖武， 冯家良，等. 武汉市城区大气 $PM_{2.5}$ 的碳组分与源解析[J].

生态环境学报， 2012(9):1574-1579.

[39]崔东文，金波. 基于随机森林回归算法的水生态文明综合评价[J]. 水利水电科技进展，2014，34（5）：56-61.

[40]戴昭鑫，张云芝，胡云锋，等. 基于地面监测数据的 2013~2015 年长三角地区 $PM_{2.5}$ 时空特征[J]. 长江流域资源与环境，2016，25（5）：813-822.

[41]丁冰, 陈健, 王彬,等. 城市环境 $PM_{2.5}$ 空间分布监测方法研究进展[J]. 地球与环境, 2016, 44(1):130-138.

[42]伏晴艳,阚海东.城市大气污染健康危险度评价的方法第四讲大气污染的暴露评价第二节大气扩散模型及人口加权的大气污染暴露评价(续四)[J].环境与健康杂志,2004,(6):414-416. DOI:10.3969/j.issn.1001-5914.2004.06.024.

[43]符立伟，郭秀锐. 国内空气污染暴露水平评价方法研究进展[J]. 环境科学与技术， 2015， 38（12Q）： 226-230.

[44]付维雅. 第三代空气质量模型的研究与应用[D]. 西安: 陕西师范大学，2010.

汉瑞英， 陈健， 王彬， 等. 利用 LUR 模型模拟杭州市 $PM_{2.5}$ 质量浓度空间分布[J]. 环境科学学报， 2016， 36（9）： 3379-3386.

[45]汉瑞英，陈健，王彬，等. 利用 LUR 模型模拟浙江省 $PM_{2.5}$ 质量浓度空间分布[J]. 科技通报，2016，35（8）：215-221.

[46]江曲图，何俊昱，叶观琼， 等. 基于 LUR/BME 的海岸带地区 $PM_{2.5}$ 时空特性研究[J]. 中国环境科学， 2017， 37（2）： 424-431.

[47]江苏省统计局，国家统计局江苏调查总队.江苏统计年鉴 2006-2010[M].北京：中国统计出版社，2006-2011.

[48]焦利民， 许刚， 赵素丽， 等. 基于 LUR 的武汉市 $PM_{2.5}$ 浓度空间分布模拟[J]. 武汉大学学报（信息科学版）， 2015， 40（8）： 1088-1094.

[48]李世广，蒋厦，俘洪金，等． 基于空气质量模型 CMAQ 的成渝经济区(四川) PM2.5 浓度数值模拟研究[J]．大气环境，2013，32: 109－113.

[49]林海峰， 辛金元，张文煜，等. 北京市近地层颗粒物浓度与气溶胶光学厚度相关性分析研究[J]. 环境科学，2013，34（3）：826-834.

[50]漏嗣佳，朱彬，廖宏．中国地区臭氧前体物对地面臭氧的影响[J]．南京气象学院学报，2010，33(4) : 451－459.

[51]罗艳青. $PM_{2.5}$ 浓度土地利用回归建模关键问题研究[D]. 中南大学, 2014.

[52]马玥，姜琦刚，孟治国，等. 基于随机森林算法的农耕区土地利用分类研究[J].农业机械学报，2016，47（1）：297-303.

[53]毛婉柳，徐建华，卢德彬，等. 2015 年长三角地区城市 $PM_{2.5}$ 时空格局及影响因素分析[J]. 长江流域资源与环境，2017，26（2）：264-272.

[54]上海市统计局，国家统计局上海调查总队.上海统计年鉴 2006-2010[M].北京：中国统计出版社，2006-2011.

[55]谭敏, 刘凯, 柳林, 等. 2017. 基于随机森林优化模型的珠江三角洲 30 m 格网人口空间化[J]. 地理科学进展, 36 (10): 1304- 1312.

[56]同丽嘎，李雪铭，黄哲，等. 包头市人口 $PM_{2.5}$ 暴露风险研究[J].农业机械学报，2015，46（1）：259-266.

[57]王飞龙. 基于机器学习的北京 $PM_{2.5}$ 预测算法[D].天津工业大学,2017.

[58]王静,杨复沫,王鼎益,贺克斌.北京市 MODIS 气溶胶光学厚度和 $PM_{2.5}$ 质量浓度的特征及其相关性[J].中国科学院研究生院学报,2010,27(01):10-16.

[59]王丽爱，马昌，周旭东，等. 基于随机森林回归算法的小麦叶片 SPAD 值遥感估算[J].农业机械学报，2015，46（1）：259-266.

[60]王丽涛，潘雪梅，郑佳，等．河北及周边地区霾污染特征的模拟研究[J]．环境科学学报，2012，4: 925－931．

[61]王敏，邹滨，郭宇，等．基于 BP 人工神经网络的城市 $PM_{2.5}$ 浓度空间预测[J]．环境污染与防治，2013，35(9)：63-66．

[62]王琪,孙巍,张新宇.北京地区 $PM_{2.5}$ 质量浓度分布及其与气象条件影响关系分析[J].计算机与应用化学,2014,31(10):1193-1196.

[63]吴健生，廖星，彭建， 等. 重庆市 $PM_{2.5}$ 浓度空间分异模拟及影响因子[J]. 环境科学， 2015， 36（3）： 759-768.

[64]徐杰, 匡汉祎, 王国强, 等.$PM_{2.5}$ 与空气相对湿度间关系浅析[J].农业与技术,2017,(9):148-149,157.

[65]许刚， 焦利民， 肖丰涛， 等. 土地利用回归模型模拟京津冀 $PM_{2.5}$ 浓度空间分布[J]. 干旱区资源与环境， 2016， 30（10）： 116-121.

[66]许建明，徐祥德，刘煜，等．CMAQ-MOS 区域空气质量统计修正模型预报

途径研究[J].中国科学(地球科学) ，2005，35(z1) : 131－144.

[67]杨艳．基于 GEOS-Chem 模型的大气二氧化碳循环模拟研究[D]．北京：中国地质大学，2010．

[68]喻其炳,李勇,白云,姚行艳,成志伟,李川．基于聚类分析与偏最小二乘法的支持向量机 $PM_{2.5}$ 预测[J]．环境科学与技术,2017,40(06):157-164.

[69]张怡文,敖希琴,时培俊，等．基于 Pearson 相关指标的 BP 神经网络 $PM_{2.5}$ 预测模型[J]．青岛大学学报(自然科学版),2017,30(02):83-87.

[70]赵佳楠,徐建华,卢德彬,杨东阳,毛婉柳.基于 RF-LUR 模型的 $PM_{2.5}$ 空间分布模拟——以长江三角洲地区为例[J].地理与地理信息科学,2018(01):18-23.

[71]浙江省统计局，国家统计局浙江调查总队.浙江统计年鉴 2006-2010[M].北京：中国统计出版社，2006-2011.

[72]朱蕾，黄敬峰．山区县域尺度降水量空间插值方法比较[J]．农业工程学报，2007，23（7）：80-86.

[73]朱亚杰，李琦，侯俊雄，等．运用贝叶斯方法的 $PM_{2.5}$ 浓度时空建模与预测[J].测绘科学，2016，41（2）：44-49.

[74]邹滨，彭芬，焦利民,等．高分辨率人口空气污染暴露 GIS 空间区划研究[J].武汉大学学报(信息科学版), 2013, 38(3):334-338.

[75]邹滨，蒲强，罗岳平,等．城市 $PM_{2.5}$ 污染防控多指标空间区划研究[J]．安全与环境学报, 2016(1):337-342.

附录 1：相关代码

（1）生成 Fragstats 4.2 所需文件

```
import os
fp = r'f:\landuse\land7'
fpath = open('f:/land7.txt','w')
b=[]
for dirpath,filename,filenames in os.walk(fp):
    for filename in filenames:
        if os.path.splitext(filename)[1] == '.tif':#判断是否为 tif 格式
            filepath = os.path.join(dirpath,filename)
            b.append(filepath)
for name in b:
    #fpath.write(name+', x, 999, x, x, 1, x, IDF_GeoTIFF'+'\n')
    fpath.write(name+', x, 999, x, x, 1, x, IDF_GeoTIFF'+'\n')
    #print(name)
```

（2）矢量文件分割栅格图像

```
import sys
reload(sys)
sys.setdefaultencoding( "utf-8" )
import arcpy
import string
from arcpy.sa import *
try:
    raster = arcpy.GetParameterAsText(0)
    clip_feat = arcpy.GetParameterAsText(1)
    field = arcpy.GetParameterAsText(2)
    outworkspace = arcpy.GetParameterAsText(3)
    i=0
    for row in arcpy.SearchCursor(clip_feat):
        mask=row.getValue("Shape")
        outPath=outworkspace+"\\"+str(row.getValue(field))
        outExtractByMask = ExtractByMask(raster,mask)
        outExtractByMask.save(outPath+'.tif')
        i=i+1
        arcpy.AddMessage("Record number: " + str(i) + " writen to files")
except arcpy.ExecuteError:
    print arcpy.GetMessages()
```

（3）计算误差参数

```
#计算误差参数
```

```
from sklearn.cross_validation import train_test_split
from sklearn.ensemble import RandomForestRegressor
import pandas as pd
import numpy as np
import time
import csv
from sklearn import svm
from sklearn.svm import SVR
from sklearn import preprocessing

train = pd.read_csv(u"c:/train1.csv")
test = pd.read_csv(u"c:/test1.csv")
yucey=[]
rmsexl=0
maexl=0
avgxl=train['obj'].values.mean()
ia1xl=0
ia2xl=0
iaxl=0
errorxl=[]
for (x,y) in zip(train['yuce'],train['obj']):
    print(x)
    print(y)
    errorxl.append(y-x)
    rmsexl=rmsexl+(y-(x))**2
    maexl=maexl+abs(y-(x))
    ia1xl=ia1xl+(y-(x))**2
    ia2xl=ia2xl+(abs((x)-avgxl)+abs(y-avgxl))**2
rmsexl=(rmsexl/(len(train)))**0.5
maexl=maexl/(len(train))
iaxl=1-(ia1xl/ia2xl)
```

（4）大文件分割

```
limit=1000
file_count=0
url_list=[]
with open(r'f:\land1.txt') as f:
    for line in f:
        url_list.append(line)
        if len(url_list)<limit:
            continue
        file_name=r'f:\1km'+'\\'+str(file_count)+'.txt'
        with open(file_name,'w') as file:
```

```
                for url in url_list[:-1]:
                    print(url_list)
                    file.write(url)
                file.write(url_list[-1].strip())
                url_list=[]
                file_count+=1
if url_list:
    file_name=r'f:\1km'+'\\'+str(file_count)+'.txt'
    with open(file_name,'w') as file:
        for url in url_list:
            file.write(url)
print('done')
```

（5）计算栅格图像的平均值

```
import numpy as np
import osgeo.gdal as gdal
import pandas as pd
import math
#栅格图像平均值
file1='F:/pm_pop1.tif'
data=gdal.Open(file1,gdal.GA_ReadOnly)
raster_array1 = data.ReadAsArray()
raster_array1=np.round(raster_array1)
data_list1=raster_array1.tolist()
data_fl1=[i    for item in data_list1 for i in item    ]
data_fl1=[i for i in data_fl1 if i != 999999]
print sum(data_fl1)/len(data_fl1)
```

（6）统计地类比率

```
#计算序号为 30 的地类
import os
import numpy as np
import osgeo.gdal as gdal
import pandas as pd

def getFileName(path):
    ''' 获取指定目录下的所有指定后缀的文件名 '''
    tif_list=[]
    f_list = os.listdir(path)
    # print f_list
    for i in f_list:
        # os.path.splitext():分离文件名与扩展名
```

```
            if os.path.splitext(i)[1] == '.tif':
                tif_list.append(i)
        return tif_list
if __name__ == '__main__':
        path='F:/landuse/land10/'
        tiff=getFileName(path)
        res=[]
        nums=[]
        pers=[]
        i=0
        for t in tiff:
            nums.append(os.path.splitext(t)[0])
            data=gdal.Open(path+t,gdal.GA_ReadOnly)
            raster_array = data.ReadAsArray()
            per=float(np.sum(raster_array==30))/np.sum(raster_array!=0)*100
            pers.append(per)
            i=i+1
            print i
        s1=pd.Series(np.array(nums))
        s2=pd.Series(np.array(pers))
        df=pd.DataFrame({"fid":s1,"per":s2})
        df.to_csv('f:/10per.csv')
```

（7）随机森林训练

```
from sklearn.cross_validation import train_test_split
from sklearn.ensemble import RandomForestRegressor
import pandas as pd
import numpy as np
import time
import csv
from sklearn import svm
from sklearn.svm import SVR
from sklearn import preprocessing

#创建 csv
def createListCSV(fileName="", dataList=[]):
        with open(fileName, "w") as csvFile:
            csvWriter = csv.writer(csvFile)
            for data in dataList:
                csvWriter.writerow(data)
            csvFile.close()

#训练 rf
```

```
datama=pd.read_excel(u"I:/xunlian.xlsx")
train=datama.iloc[:,2:]
target=datama.iloc[:,1]
train_X,test_X, train_y, test_y = train_test_split(train,target,test_size = 0.2,random_state = 0)
rf=RandomForestRegressor(n_estimators=1500,oob_score=True)
rf.fit(train_X.values,train_y.values)
```

（8）rf 模型效果参数计算

```
#训练 rf 效果
yucey=[]
rmsexl=0
maexl=0
avgxl=train_y.values.mean()
ia1xl=0
ia2xl=0
iaxl=0
errorxl=[]
for (x,y) in zip(train_X.values,train_y.values):
    print(x)
    print(y)
    yucey.append(rf.predict(x))
    errorxl.append(y-rf.predict(x))
    rmsexl=rmsexl+(y-(rf.predict(x)))**2
    maexl=maexl+abs(y-(rf.predict(x)))
    ia1xl=ia1xl+(y-(rf.predict(x)))**2
    ia2xl=ia2xl+(abs((rf.predict(x))-avgxl)+abs(y-avgxl))**2
rmsexl=(rmsexl/(len(train_y)))**0.5
maexl=maexl/(len(train_y))
iaxl=1-(ia1xl/ia2xl)
ac=train_X
ac['obj']=list(train_y.values)
ac['yuce']=[i[0] for i in yucey]
ac.to_csv("c:/jieguo11.csv",encoding="utf-8")

#rf 监测点效果
yucey1=[]
rmsexl1=0
maexl1=0
avgxl1=test_y.values.mean()
ia1xl1=0
ia2xl1=0
iaxl1=0
errorxl1=[]
```

```
for (x,y) in zip(test_X.values,test_y.values):
        print(x)
        print(y)
        yucey1.append(rf.predict(x))
        errorxl1.append(y-rf.predict(x))
        rmsexl1=rmsexl1+(y-(rf.predict(x)))**2
        maexl1=maexl1+abs(y-(rf.predict(x)))
        ia1xl1=ia1xl1+(y-(rf.predict(x)))**2
        ia2xl1=ia2xl1+(abs((rf.predict(x))-avgxl1)+abs(y-avgxl1))**2
rmsexl1=(rmsexl1/(len(test_y)))**0.5
maexl1=maexl1/(len(test_y))
iaxl1=1-(ia1xl1/ia2xl1)
ac1=test_X
ac1['obj']=list(test_y.values)
ac1['yucey']=[i[0] for i in yucey1]
ac1.to_csv("c:/jieguo1.csv",encoding="utf-8")
```

（9）svm 训练

```
datama = pd.read_excel(u"I:/xunlian.xlsx")
train = datama.iloc[:,2:]
target = datama.iloc[:,1]
train_X,test_X, train_y, test_y = train_test_split(train,target,test_size = 0.2,random_state = 0)
train_X_scaled = preprocessing.scale(train_X)
svr = svm.SVR(C = 1000)
svr.fit(train_X_scaled, train_y)
```

（10）svm 模型效果参数计算

```
#训练 svm 效果
yucey2=[]
rmsexl2=0
maexl2=0
avgxl2=train_y.values.mean()
ia1xl2=0
ia2xl2=0
iaxl2=0
errorxl2=[]
for (x,y) in zip(list(train_X_scaled),train_y.values):
    print(x)
    print(y)
    yucey2.append(svr.predict(x))
    errorxl2.append(y-svr.predict(x))
    rmsexl2=rmsexl2+(y-(svr.predict(x)))**2
```

```
        maexl2=maexl2+abs(y-(svr.predict(x)))
        ia1xl2=ia1xl2+(y-(svr.predict(x)))**2
        ia2xl2=ia2xl2+(abs((svr.predict(x))-avgxl2)+abs(y-avgxl2))**2
rmsexl2=(rmsexl2/(len(train_y)))**0.5
maexl2=maexl2/(len(train_y))
iaxl2=1-(ia1xl2/ia2xl2)
ac2=train_X
ac2['obj']=list(train_y.values)
ac2['yuce']=[i[0] for i in yucey2]
ac2.to_csv("c:/jieguo3.csv",encoding="utf-8")

#svm 监测点效果
yucey3=[]
rmsexl3=0
maexl3=0
avgxl3=test_y.values.mean()
ia1xl3=0
ia2xl3=0
iaxl3=0
errorxl3=[]
test_X_scaled=preprocessing.scale(test_X)
for (x,y) in zip(list(test_X_scaled),test_y):
        print(x)

        print(svr.predict(x))
        z=svr.predict(x)



        yucey3.append(z)
        errorxl3.append(y-z)
        rmsexl3=rmsexl3+(y-(z))**2
        maexl3=maexl3+abs(y-(z))
        ia1xl3=ia1xl3+(y-(z))**2
        ia2xl3=ia2xl3+(abs((z)-avgxl3)+abs(y-avgxl3))**2
rmsexl3=(rmsexl3/(len(test_y)))**0.5
maexl3=maexl3/(len(test_y))
iaxl3=1-(ia1xl3/ia2xl3)
df=pd.DataFrame({'obj':list(test_y.values),'yuce':[i[0] for i in yucey3]})
df.to_csv("c:/jieguo4.csv",encoding="utf-8")
```

（11）rf 参数探索

```
#参数探索
nlist= [a for a in range(1,2000,20)]
```

```
datalist=[]
rmseL=[]
for n in nlist:
    print('**************************',n)
    time.sleep(3)
    rf=RandomForestRegressor(n_estimators=n,oob_score=True)
    rf.fit(train_X.values,train_y.values)
    #训练效果
    rmsexl=0
    maexl=0
    avgxl=train_y.values.mean()
    ia1xl=0
    ia2xl=0
    iaxl=0
    errorxl=[]
    for (x,y) in zip(train_X.values,train_y.values):
        errorxl.append(y-rf.predict(x))
        rmsexl=rmsexl+(y-(rf.predict(x[:])))**2
        maexl=maexl+abs(y-(rf.predict(x[:])))
        ia1xl=ia1xl+(y-(rf.predict(x[:])))**2
        ia2xl=ia2xl+(abs((rf.predict(x[:]))-avgxl)+abs(y-avgxl))**2
    rmsexl=(rmsexl/(len(train_y)))**0.5
    rmsexl=rmsexl.tolist()[0]
    maexl=maexl/(len(train_y))
    iaxl=1-(ia1xl/ia2xl)
    rmseL.append(rmsexl)
datalist.append(nlist)
datalist.append(rmseL)
createListCSV("c:/result.csv",datalist)
```

（12）PM2.5 浓度空间模拟

```
from sklearn.cross_validation import train_test_split
from sklearn.ensemble import RandomForestRegressor
import pandas as pd
import numpy as np
import time
import csv
from sklearn import svm
from sklearn.svm import SVR
from sklearn import preprocessing

#训练 rf
datama=pd.read_excel(u"F:/train.xlsx")
```

```
train=datama.iloc[:,1:-1]
target=datama.iloc[:,-1]
rf=RandomForestRegressor(n_estimators=1500,oob_score=True)
rf.fit(train,target)
datamb=pd.read_excel(u"F:/final.xlsx")
moni=datamb.iloc[:,1:]
pre=rf.predict(moni.values)
pre=pd.Series(pre)
pre.to_csv('F:/pre.csv')
```

（13）信息熵计算

```
from sklearn.cross_validation import train_test_split
from sklearn.ensemble import RandomForestRegressor
import pandas as pd
import numpy as np
import time
import csv
from sklearn import svm
from sklearn.svm import SVR
from sklearn import preprocessing

#训练 rf
datama=pd.read_excel(u"F:/train.xlsx")
train=datama.iloc[:,1:-1]
target=datama.iloc[:,-1]
rf=RandomForestRegressor(n_estimators=1500,oob_score=True)
rf.fit(train,target)
datamb=pd.read_excel(u"F:/final.xlsx")
moni=datamb.iloc[:,1:]
pre=rf.predict(moni.values)
pre=pd.Series(pre)
pre.to_csv('F:/pre.csv')
```

（14）人口加权 PM2.5 暴露风险计算

```
import os
import numpy as np
import osgeo.gdal as gdal
import pandas as pd

def getFileName(path):
    ''' 获取指定目录下的所有指定后缀的文件名 '''
    tif_list=[]
```

```
    f_list = os.listdir(path)
    # print f_list
    for i in f_list:
        # os.path.splitext():分离文件名与扩展名
        if os.path.splitext(i)[1] == '.tif':
            tif_list.append(i)
    return tif_list
if __name__ == '__main__':
    path='F:/POPF/'
    path1='F:/PMF/'
    tiff=getFileName(path)
    tiff1=getFileName(path1)
    names=[]
    tols=[]
    for t,j in zip(tiff,tiff1):
        if t==j:
            print t
            names.append(os.path.splitext(t)[0])
            pop=gdal.Open(path+t,gdal.GA_ReadOnly)
            pop = pop.ReadAsArray()
            pop[pop < -999999999]=0
            pm=gdal.Open(path1+j,gdal.GA_ReadOnly)
            pm = pm.ReadAsArray()
            pm[pm < -999999999]=0
            tol=pop.sum()
            pwel=pop*pm/tol
            pweltol=pwel.sum()
            tols.append(pweltol)

    s1=pd.Series(np.array(names))
    s2=pd.Series(np.array(tols))
    df=pd.DataFrame({"fid":s1,"tols":s2})
    df.to_csv('f:/jiaquan.csv')
```

附录 2：硕士期间科研情况

发表学术论文：

1、**赵佳楠**，徐建华，卢德彬，杨东阳，毛婉柳.基于 RF-LUR 模型的 PM_(2.5)空间分布模拟——以长江三角洲地区为例[J].地理与地理信息科学,2018(01):18-23.

2、毛婉柳，徐建华，卢德彬，杨东阳，**赵佳楠**. 2015 年长三角地区城市 PM2.5 时空格局及影响因素分析[J]. 长江流域资源与环境，2017，26（2）：264-272.

3、Yang D, Lu D, Xu J, **Zhao J**. Predicting spatio-temporal concentrations of PM2.5 using land use and meteorological data in Yangtze River Delta, China[J]. Stochastic Environmental Research & Risk Assessment,2017(9):1-12.

4、Lu, Debin & Xu, Jianhua & Yang, Dongyang & **Zhao, Jianan**. (2017). Spatio-temporal variation and influence factors of PM 2.5 concentrations in China from 1998 to 2014. Atmospheric Pollution Research. 10.1016/j.apr.2017.05.005.

5、Lu, Debin & Mao, Wanliu & Yang, Dongyang & **Zhao, Jianan** & Xu, Jianhua. (2018). Effects of land use and landscape pattern on PM 2.5 in Yangtze River Delta, China. Atmospheric Pollution Research. 10.1016/j.apr.2018.01.012.

参与科研项目：

1. 2015.10-2016.06　　“上海市农业布局专项规划”

2. 2015.10-2016.05　　“基于 WebGIS 的上海城市建设用地数据库”建设

致谢

三年前,初次踏入华师校园的画面仿佛仍在眼前,转眼已近毕业。回想过往三年心中感慨颇多，既有对即将离别的不舍，又有对未来的憧憬。

首先，我必须由衷的对我的导师徐建华教授表示感激。求学之路不易，所幸的是能遇到徐老师。老师不仅学识渊博，言行举止更是处处体现着对待学术一丝不苟的态度以及对于学生的的殷殷期待。您将我领进了研究生学习的大门。三年间，我从科研学习中慢慢体悟到坚持不懈、精益求精的重要，人生亦如此。您对教学的热爱也深深的影响着我，在协助老师教学的过程中愈发理解教学相长的含义。

感谢李治洪老师、周坚华老师、唐熙老师等各位老师对我进行的专业指导。尤其是李治洪老师，是您提供了我项目实践的机会，让我进一步接触了程序设计，使我能够更好的理解并掌握地理信息技术。

感谢 334 这个友爱的大家庭。三年间，一路走来，经历了不少的困顿，所幸的是有这个大家庭的鼓励与支持，让我能够自信的趟过去。感谢已毕业的陈忠生、王祖静、韩乐乐、马亮旭师兄以及柏玲、徐艺文、张影、毛婉柳师姐，感谢卢德彬、王充、杨东阳、杨旭师兄及朱妮娜、杨海清师姐，感谢尹梁明师弟和左京平、刘薇、魏钰烨师妹，感谢同级的郝玉以及胡亚丹。你们无一不在专业知识上给我进行了指导。尤其要感谢卢德彬、杨东阳师兄，本文除了徐老师的悉心指导以外，你们在数据、方法等多方面给我提供了很多帮助。334 承载了我太多美好时光：一起下馆子吃火锅；一起分享生日的喜悦；一起感受户外的美景……有幸遇见，不舍离别。

感谢我的父母，一路走来谢谢你们对我的关爱。是你们的爱，让我在学业上，一步一步走的踏实。三年前，我带着你们的期待走入了新的学习阶段；三年后，我即将整理心情，带着满满的收获离开校园。你们额头的每一根白发都是对我爱的诠释，希望我不辜负你们的期望，在今后的日子里承担起“家”的责任。

最后感谢这三年来每一个在我时光中相逢的朋友，愿我们的生活里满是阳光。

赵佳楠

2018 年 4 月 27 日于资环楼 334 室