DOI:10.13203/j.whugis20170042 文章编号:1671-8860(2018)10-1574-06

基于改进 LUR 模型的大区域 $PM_{2.5}$ 浓度空间分布模拟

李 爽[1] 翟 亮[2,3] 桑会勇[2,3] 邹 滨[4] 方 新[4] 甄云鹏[5]

1 深圳市勘察研究院有限公司,广东 深圳,518026
2 中国测绘科学研究院,北京,100830
3 地球观测与时空信息科学国家测绘地理信息局重点实验室,北京,100830
4 中南大学地球科学与信息物理学院,湖南 长沙,410083
5 武汉市测绘研究院,湖北 武汉,430000

摘 要:针对采用传统土地利用回归(land use regression, LUR)模型进行大气污染物浓度模拟时预测变量信息损失的缺陷,将主成分分析(principle component analysis, PCA)与逐步多元线性回归(stepwise multiple line regression, SMLR)相结合,提出了一种改进的 LUR(PCA+SMLR)模型模拟大区域 $PM_{2.5}$ 浓度空间分布的方法。首先采用相关分析筛选与 $PM_{2.5}$ 显著相关的预测变量,然后对筛选出的预测变量进行主成分变换(PCA),最后保留所有主成分变量进行 SMLR 建立回归模型模拟 $PM_{2.5}$ 浓度。并以京津冀为研究区域进行实验验证,对 PCR、SMLR、PCA+SMLR 这 3 种模型的实验结果进行对比分析,结果表明,PCA+SMLR 模型可提高预测变量对回归模型的贡献度,调整后 R^2 达 0.883,并且其精度检验指标及制图效果皆优于传统的 LUR 模型,证明了该模型可有效提高 $PM_{2.5}$ 浓度的模拟精度,对 $PM_{2.5}$ 区域联防联控具有指导意义。

关键词:$PM_{2.5}$;PCA;SMLR;贡献度;空间分布

中图分类号:P208 **文献标志码**:A

大气细小颗粒物 $PM_{2.5}$(直径小于等于2.5 μm)是大气主要污染物之一,与雾霾天气的发生密切相关[1]。根据环保部发布的《环境空气质量标准》中规定的居民区 $PM_{2.5}$ 年均浓度不超过 35 $\mu g/m^3$ 来衡量,2017 年 1 月全国 $PM_{2.5}$ 排行榜中的 114 个城市仅有 8 个城市空气质量达标。研究表明,$PM_{2.5}$ 会导致心血管和呼吸系统等疾病发病率的增加,严重影响人们的身体健康[2-3]。$PM_{2.5}$ 污染已成为严峻的社会问题,并引起了公众及政府环保部门的广泛关注[4]。

$PM_{2.5}$ 浓度模拟可为环保部门治理大气污染提供决策支持,$PM_{2.5}$ 浓度精确模拟已成为当前研究热点[5-6]。土地利用回归(land use regression, LUR)[7]模型是大气污染物浓度模拟的主要研究方法之一,该类研究采用与因变量显著相关的预测变量直接进行逐步多元线性回归(stepwise multiple line regression, SMLR)[8-9];或对预测变量进行主成分变换(principal component analysis, PCA),之后挑选特征根大于 1 或者累计方差贡献率达到 80%的前几个主成分进行主成分回归(principal component regression, PCR)[10-12],建立回归模型。但是,SMLR 方法中的预测变量存在一定的共线性问题,并且在逐步回归时直接从模型中剔除了部分与因变量显著相关的预测变量;而 PCR 方法虽解决了预测变量的共线性问题,但该方法直接采用前几个主成分变量建立回归模型,没有进行主成分变量的筛选。

针对上述不足,本研究将 PCA 与 SMLR 两种方法相结合,首先采用相关分析筛选与 $PM_{2.5}$ 显著相关的预测变量,然后对筛选出的预测变量进行 PCA,最后保留所有主成分变量进行 SMLR 确立最优建模驱动因子,同时构建回归模型进行 $PM_{2.5}$ 浓度空间分布模拟。

收稿日期:2017-02-28

项目资助:2017 年地理国情监测工程(第二批)第一包专题性监测(Q1722);国家自然科学基金(41701213);中国测绘科学研究院基本科研业务费项目(7771716);地理国情监测国家测绘地理信息局重点实验室开放基金(2016NGCMZD03);国家自然科学青年基金(41501192)。

第一作者:李爽,硕士,主要从事地理国情信息统计与分析方面的研究。ls02020029@163.com

通讯作者:桑会勇,博士。huiyong.sang@casm.ac.cn

1 研究模型的建立

1.1 改进的LUR模型

本文在传统LUR模型的基础之上构建了改进的LUR模型。相比于传统LUR模型在预测变量信息损失方面的缺陷,改进的LUR模型不仅可以消除预测变量的共线性,从而避免信息冗余,而且可以让所有与$PM_{2.5}$显著相关的预测变量参与到回归建模构建当中,达到提高预测变量对回归模型贡献度的目的。本文提出的改进LUR模型的核心在于结合PCA与SMLR两种方法建立回归模型,即先利用PCA消除预测变量的共线性,之后利用SMLR将变换后的预测变量逐步引入回归模型之中。

PCA的基本思想是将原来众多具有一定相关性的变量重新组合成一组相互无关的新变量来代替原来的变量[10,13]。所选取的新变量被称为主成分变量,选取的原则是尽可能保留原有变量所包含的信息。PCA在数学上的处理是将原来的变量X作线性组合,生成新的综合变量P,模型结构如下[11]:

$$P_i = l_{1i}X_1 + l_{2i}X_2 + \cdots + l_{ni}X_n \quad (1)$$

式中,P_i表示第i个主成分变量;l_{ni}表示预测变量X_n的载荷。由于各预测变量X_n的量纲不同,需要先对其进行0~1标准化处理,之后采用PCA方法将预测变量转换为主成分变量,以消除原预测变量的共线性。不同于以往的PCR方法,本研究不依据特征根或方差贡献率直接选取前几个主成分变量,而是利用SMLR方法对主成分变量进行筛选。

获取P_i之后,利用SMLR方法建立回归模型。SMLR是传统多元线性回归模型的扩展,其基本思想是在向前引入每一个新的自变量之后都要重新对之前已选入的自变量进行检查,以评价其有无继续保留在方程中的价值[14]。SMLR中自变量是否被引入或剔除取决于其偏回归平方和的F检验或校正决定系数R^2(Adjusted R^2,Adj_R^2),自变量的引入和剔除交替进行,直到无具有统计学意义的新变量可以引入,也无失去统计学意义的自变量可以剔除时为止[15]。SMLR的公式如下:

$$Y_i = \beta_0 + \sum_{i=1}^{n}\beta_i X_i + \varepsilon \quad (2)$$

式中,Y_i表示因变量;X_i表示自变量;β_i表示回归系数;ε为模型的随机误差。以P_i作为自变量,利用SPPS 22.0自动实现自变量的引入或剔除,最终建立回归模型。

1.2 技术流程

将PCA与SMLR两种方法相结合构建了一种改进的LUR模型模拟$PM_{2.5}$浓度,整个研究分为预测变量筛选、回归建模、模型检验、$PM_{2.5}$年均浓度空间分布模拟制图4个子过程。首先依据现有研究结果[6,16-17]提取预测变量,进而根据Pearson相关性系数筛选与$PM_{2.5}$显著相关的预测变量;然后对筛选出的预测变量进行PCA,并保留所有P_i进行SMLR建立回归模型;之后统计拟合模型与交叉验证[18]模型下的均方根误差(root mean square error, RMSE)、平均预测误差(mean prediction error, MPE)、平均相对预测误差(mean relative prediction error, MRPE)3个指标[17]来检验模型性能;最后在研究区内建立10 km×10 km的加密点并采用普通克里金插值方法进行整个京津冀地区的$PM_{2.5}$年均浓度空间分布模拟制图。技术路线如图1所示。

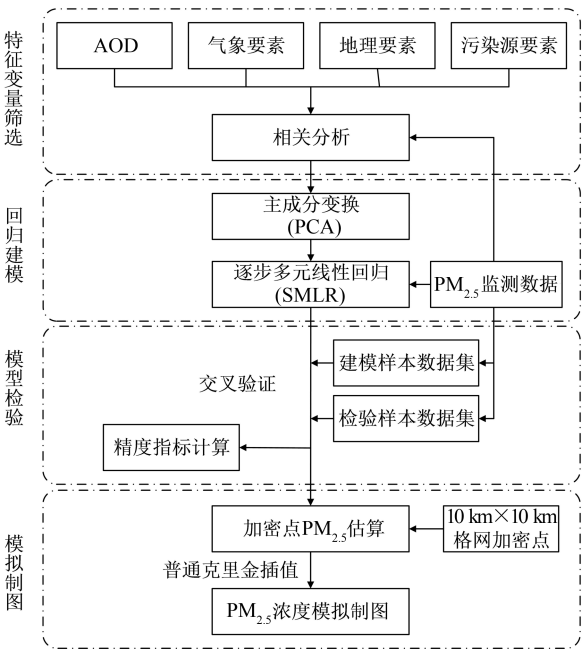

图1 技术路线图

Fig.1 The Technique Flow Chart

2 数据分析与处理

2.1 研究区概况及数据收集

京津冀地区东临渤海,西为太行山地,北为燕山山地,地势西北高东南低,面积约21.6万km^2(见图2)。该地区经济发展迅速,加之三面环山的地形条件,使其成为国内大气污染最严重的地

区之一。

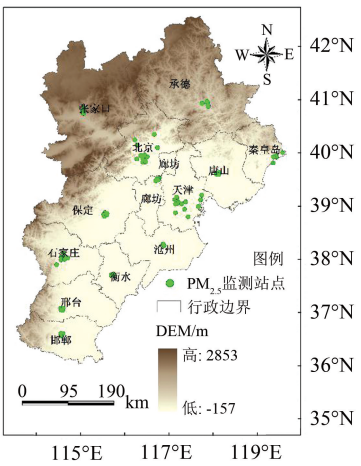

图 2　研究区域示意图

Fig.2　Sketch Map of the Study Area

本研究采用的数据可分为 5 大部分:$PM_{2.5}$ 监测站点实时浓度数据、气溶胶光学厚度(aerosol optical depth, AOD)数据、气象要素数据、地理要素数据、污染源要素数据。$PM_{2.5}$ 监测站实时浓度数据来自中国环境监测总站城市空气质量实时发布平台;AOD 数据采用从美国航空航天宇航局数据中心网站下载的 MOD04_L2 大气气溶胶数据产品;气象要素数据包括风速、气压、温度、降水、湿度,皆来源于中国地面气候资料日值数据集;地理要素数据包括 DEM、道路数据和地表覆盖数据;污染源要素数据包括采用高分辨率遥感影像或航空正射影像获取的扬尘地表污染源数据和从企业法人数据库整理得到的工业企业污染源数据。

2.2　结果与分析

本文对 PCR、SMLR 及 PCA+SMLR 这 3 种方法的实验结果进行了对比分析。

2.2.1　回归模型构建结果

对于传统的 PCR 方法,回归模型拟合优度与主成分变量个数之间的关系如图 3 所示。从图 3 可以看出:主成分变量个数达到 8 个以后,回归模型的拟合优度趋于平稳,当所有主成分变量全部进入回归模型时,其拟合优度最高,达到 0.880。但研究表明[19],回归模型中预测变量个数过多会导致模型的过拟合问题,当因变量与自变量之比为 10～15 时模型较为合理。本文共 78 个 $PM_{2.5}$ 监测站点,选取 5～7 个变量作为建模回归因子为宜。因此,本文选取特征根大于 1 的 6 个主成分变量构建回归模型 PCR1,同时为了验证过拟合问题,构建含有 17 个主成分变量的回归模型 PCR2,并构建了 SMLR 及 PCA+SMLR 模型。

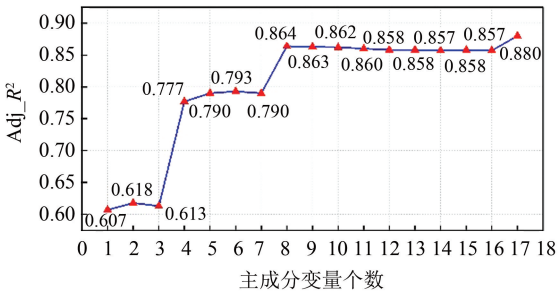

图 3　回归模型拟合优度折线图

Fig.3　The Line Chart of Goodness of Fit for Regression Model

上述 4 种模型的参数及拟合度如表 1 所示。从表 1 可知:SMLR 模型只保留了 5 个与 $PM_{2.5}$ 相关的预测变量,其余 12 个与 $PM_{2.5}$ 强相关的预测变量对模型无贡献;PCR1 与 PCR2 模型的建模驱动因子为主成分变量,因此 17 个与 $PM_{2.5}$ 强相关的预测变量都对回归模型有所贡献,但 PCR1 模型的拟合度较差,而 PCR2 模型的变量太多,可能导致模型的过拟合问题;相比较而言,PCA+SMLR 模型通过 SMLR 逐步引入或剔除主成分变量,其调整后的 R^2 为 0.883,较 PCR 模型(0.793/0.880)和 SMLR 模型(0.832)有明显提升。

表 1　模型参数及拟合度对比结果

Tab.1　Comparison of Parameterization and Model Fitting for Four Models

模型	参数	Adj_R^2
PCR1	P_1、P_2、P_3、P_4、P_5、P_6	0.793
PCR2	P_1、P_2、P_3、P_4、P_5、P_6、P_7、P_8、P_9、P_{10}、P_{11}、P_{12}、P_{13}、P_{14}、P_{15}、P_{16}、P_{17}	0.880
SMLR	X_1、X_2、X_3、X_4、X_5	0.832
PCA+SMLR	P_1、P_2、P_4、P_5、P_8、P_{17}	0.883

注:P_i 为第 i 个主成分;X_1 为气溶胶光学厚度;X_2 为降水;X_3 为监测站 8 000 m 缓冲区内耕地面积占比;X_4 为监测站 8 000 m 缓冲区内房屋建筑面积占比;X_5 为监测站 5 000 m 缓冲区内的露天采掘场面积占比。

2.2.2　模型精度对比结果

图 4 展示了 4 种回归模型拟合结果与实测结果的散点图,表 2 直观地对比了 4 种模型的精度检验指标,其结果均在合理范围内。就模型的拟合精度而言,PCR2 模型的拟合结果最好,并且其 RMSE、MPE、MRPE 均优于其他 3 种模型;PCA+SMLR 模型的拟合精度次之,并且与 PCR2 模型的精度相差不大。但就模型的交叉验证精度而言,PCA+SMLR 模型的验证精度最优,并且相比拟合精度来说浮动很小,证明了该模型的可靠性与稳定性;PCR1 与 SMLR 模型的验证精度与拟合精度结果也较为接近;相反,PCR2 模型的验

证精度浮动相对较大，并且MRPE精度在4个模型里最差，表明该模型存在过拟合问题。

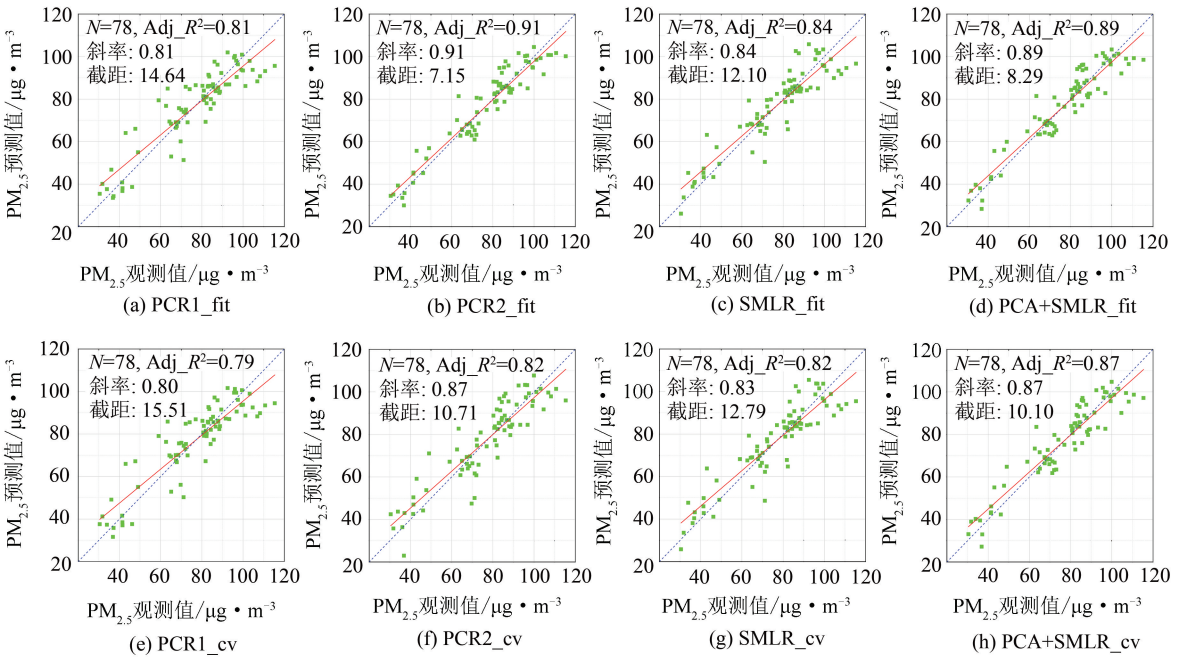

图4 模型拟合结果与实测结果散点图

Fig.4 Scatter Plots of Fitting and Measured Results

表2 精度检验指标对比结果

Tab.2 Comparison of Accuracy Indicators for Four Models

模型	拟合精度			验证精度		
	RMSE /μg·m⁻³	MPE /μg·m⁻³	MRPE /%	RMSE /μg·m⁻³	MPE /μg·m⁻³	MRPE /%
PCR1	8.942	6.869	9.920	10.394	7.102	9.300
PCR2	6.248	5.000	6.983	8.780	6.950	10.266
SMLR	8.104	6.225	8.628	9.303	6.628	8.509
PCA+SMLR	6.721	5.278	7.389	7.391	5.912	8.419

2.2.3 $PM_{2.5}$浓度模拟结果

图5为基于4种模型的$PM_{2.5}$年均浓度模拟空间分布图。从图5可以看出：虽然4种模型的$PM_{2.5}$浓度均呈现由东南至西北区域递减的趋势，但SMLR模型的模拟效果较差(见图5(c))，北京、唐山等城市$PM_{2.5}$浓度整体偏低，沧州、天津等城市中心$PM_{2.5}$浓度低，与实际情况完全相反。其他3种模型的模拟效果相近，均以太行山—燕山山脉为界限，东南地区浓度高，西北地区浓度低。相比PCR1模型(见图5(a))，PCR2(见图5(b))与PCA+SMLR模型(见图5(d))中城市中心至城市边界$PM_{2.5}$浓度逐渐降低的变化趋势更加明显。此外，张家口中心城区$PM_{2.5}$浓度较高，与之前研究中张家口$PM_{2.5}$浓度较低的结论相反，这主要是由于张家口地区筹备2022年冬奥会而产生的影响。

2.2.4 综合对比分析

从上述实验结果可以看出，传统的SMLR模型预测变量的贡献度较低且$PM_{2.5}$浓度模拟结果相对较差；PCR1模型的拟合精度相对较差；PCR2模型采用主成分变量个数过多导致模型过拟合；本文提出的改进的LUR(PCA+SMLR)模型在模型精度及$PM_{2.5}$浓度模拟上都取得了较好的结果。此外，构建好PCA+SMLR模型之后，可以通过主成分逆变换确定$PM_{2.5}$浓度与17个原始强相关特征变量之间的相关关系，进而确定研究区内的$PM_{2.5}$浓度主要受到哪些变量的影响。逆变换结果表明，本研究区内的气温、气压、降水等气象要素对$PM_{2.5}$浓度影响较大，高程、污染企业、道路次之，各类地表覆盖等短期无明显变化的地理要素对$PM_{2.5}$浓度的影响较小。

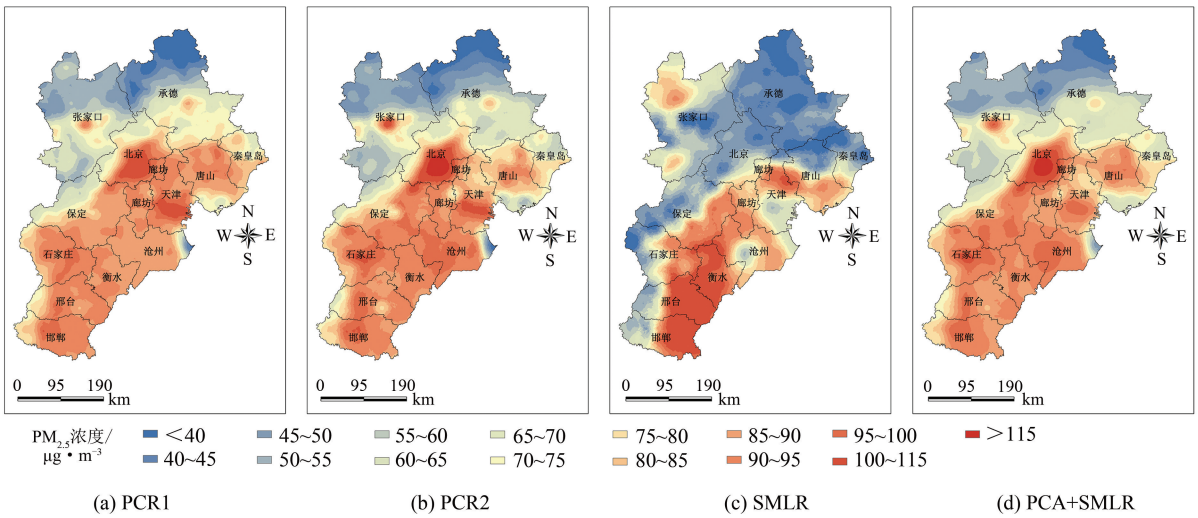

图 5　$PM_{2.5}$ 年均浓度模拟空间分布图

Fig.5　Spatial Distribution of $PM_{2.5}$ Annual Concentrations Estimated

3　结　语

本研究将 PCA 与 SMLR 方法相结合,建立改进的 LUR 模型以实现 $PM_{2.5}$ 年均浓度模拟空间分布制图。实例分析表明:PCA+SMLR 模型不仅解决了预测变量的共线性问题,而且弥补了传统 LUR 模型在预测变量信息损失方面的缺陷,其拟合度、精度检验指标及浓度模拟效果皆优于传统 LUR 模型。此外,通过本研究得到了京津冀地区 $PM_{2.5}$ 浓度的空间分布规律,为 $PM_{2.5}$ 区域联防联控提供了有力的信息支撑。

然而,本文仅选取本地区的污染源作为预测变量,未考虑外来污染的迁移因素,后续研究可综合考虑本地区污染源及输入性污染源,从而对 $PM_{2.5}$ 浓度模拟进行更加深入的探讨。

参　考　文　献

[1] Zou B, Pu Q, Bilal M, et al. High-Resolution Satellite Mapping of Fine Particulates Based on Geographically Weighted Regression [J]. *IEEE Geoscience and Remote Sensing Letters*, 2016, 13(4): 495-499

[2] Silva R A, West J J, Zhang Y, et al. Global Premature Mortality Due to Anthropogenic Outdoor Air Pollution and the Contribution of Past Climate Change[J]. *Environmental Research Letters*, 2013, 8(3):034005

[3] Lim J M, Jeong J H, Lee J H, et al. The Analysis of $PM_{2.5}$ and Associated Elements and Their Indoor/Outdoor Pollution Status in an Urban Area[J]. *Indoor Air*, 2011, 21(2):145-155

[4] Zou Bin, Xu Shan, Zhang Jin. Spatial Variation Analysis of Urban Air Pollution Using GIS: A Land Use Perspective [J]. *Geomatics and Information Science of Wuhan University*, 2017, 42(2): 216-222(邹滨, 许珊, 张静. 融合空间尺度特征的时空序列预测建模方法 [J]. 武汉大学学报·信息科学版, 2017, 42(2):216-222)

[5] Deng Min, Chen Ti, Yang Wentao. A New Method of Modeling Spatio-temporal Sequence by Considering Spatial Characteristics[J]. *Geomatics and Information Science of Wuhan University*, 2015, 40(12):1 625-1 632(邓敏, 陈倜, 杨文涛. 融合空间尺度特征的时空序列预测建模方法[J]. 武汉大学学报·信息科学版, 2015, 40(12):1 625-1 632)

[6] Zou B, Luo Y, Wan N, et al. Performance Compari-son of LUR and OK in $PM_{2.5}$ Concentration Mapping: A Multidimensional Perspective [J]. *Sci Rep*, 2015, 5(5): 8 698

[7] Briggs D J, Collins S, Elliott P, et al. Mapping Urban Air Pollution Using GIS: A Regression-based Approach[J]. *International Journal of Geographical Information Science*, 1997, 11(7):699-718

[8] Jiao Limin, Xu Gang, Zhao Suli, et al. LUR-based Simulation of the Spatial Distribution of $PM_{2.5}$ of Wuhan [J]. *Geomatics and Information Science of Wuhan University*, 2015, 40(8):1 088-1 094(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):1 088-1 094)

[9] Zou B, Xu S, Sternberg T, et al. Effect of Land Use and Cover Change on Air Quality in Urban Sprawl [J].*Sustainability*, 2016, 8(7):677

[10] 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]. *Sci Total Environ*, 2012, 425:27-34

[11] Ul-Saufie A Z, Yahaya A S, Ramli N A, et al. Future Daily PM_{10} Concentrations Prediction by Combining Regression Models and Feedforward Backpropagation Models with Principle Component Analysis (PCA) [J]. *Atmospheric Environment*, 2013, 77:621-630

[12] Li S, Zhai L, Zou B, et al. A Generalized Additive Model Combining Principal Component Analysis for $PM_{2.5}$ Concentration Estimation [J]. *ISPRS International Journal of Geo-Information*, 2017, 6: 248

[13] Ghosh D, Manson S M. Robust Principal Component Analysis and Geographically Weighted Regression Urbanization in the Twin Cities Metropolitan Area of Minnesota [J]. *J Urban Reg Inf Syst Assoc*, 2008, 20(1):15-25

[14] Zhu Jianping, Yin Ruifei. Application of SPSS in Statistical Analysis [D]. Beijing: Tsinghua University Press, 2007(朱建平, 殷瑞飞. SPSS在统计分析中的应用 [D]. 北京:清华大学出版社, 2007)

[15] Zheng Yongmei, Zhang Jun, Chen Xingdan, et al. Reasearch on Model and Wavelength Selection of Near Infrared Spectral Information [J]. *Spectrosc Spect Anal*, 2004, 24(6):675-678(郑咏梅, 张军, 陈星旦,等. 基于逐步回归法的近红外光谱信息提取及模型的研究[J]. 光谱学与光谱分析, 2004, 24(6):675-678)

[16] Zhai L, Zou B, Fang X, et al. Land Use Regression Modeling of $PM_{2.5}$ Concentrations at Optimized Spatial Scales [J]. *Atmosphere*, 2017, 8(1):1-15

[17] Fang X, Zou B, Liu X, et al. Satellite-based Ground $PM_{2.5}$ Estimation Using Timely Structure Adaptive Modeling [J]. *Remote Sensing of Environment*, 2016, 186:152-163

[18] Rodriguez J D, Perez A, Lozano J A. Sensitivity Analysis of Kappa-fold Cross Validation in Prediction Error Estimation [J]. *IEEE Trans Pattern Anal Mach Intell*, 2010, 32(3):569-575

[19] Olvera H A, Garrcia 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 Environment*, 2012, 425(3):27-34

An Improved LUR-based Spatial Distribution Simulation for the Large Area $PM_{2.5}$ Concentration

LI Shuang[1] *ZHAI Liang*[2,3] *SANG Huiyong*[2,3] *ZHOU Bin*[4] *FANG Xin*[4] *ZHEN Yunpeng*[5]

1 Shenzhen Investigation & Research Institute, Shenzhen 518026, China

2 Chinese Academy of Surveying & Mapping, Beijing 100830, China

3 Key Laboratory of Earth Observation and Geospatial Information Science of NASG, Beijing 100830, China

4 School of Geosciences and Info-physics, Central South University, Changsha 410083, China

5 Wuhan Geomatics Institute, Wuhan 430000, China

Abstract: There exists the shortage of traditional land use regression (LUR) model in losing information of predictor variables when simulating the air pollutant concentration. An improved model which combined principal component regression (PCR) and stepwise multiple line regression (SMLR)-LUR (PCA+SMLR) was developed to simulate the spatial distribution of $PM_{2.5}$ in large area. Firstly, the correlation analysis was conducted to screen out effective predictor variables. Secondly, principal component analysis (PCA) was employed to transform effective predictor variables to principle components. Finally, all principal components were used to conduct SMLR to simulate the spatial distribution of $PM_{2.5}$. Meanwhile, the reliability of the improved model was tested in Beijing-Tianjin-Hebei urban agglomeration. Experimental results of three models (PCR, SMLR and PCA+SMLR) were compared and analyzed. The results indicated that the PCA+SMLR model has an adjusted R^2 of 0.883 by improving the contribution of the predictor variables. Besides, it is better than the traditional model for accuracy index and the mapping results. Therefore, it can be concluded that the PCA+SMLR is

(下转第1587页)