

空间回归分析在土壤属性预测制图中的应用^①

刘晓冰¹, 程道全², 刘鹏飞¹, 宋 轩¹, 陈 杰^{1*}

(1 郑州大学水利与环境学院, 郑州 450001; 2 河南省土壤肥料站, 郑州 450002)

摘 要:以河南省孟津县为研究区,选取坡度、高程、地面曲率和复合地形指数(CTI)作为表层土壤缓效钾含量空间预测的环境协变量,系统探讨了空间回归分析技术在土壤属性预测制图中的应用。结果表明:土壤缓效钾的空间自相关距离阈值约为 10 000 m,与坡度、高程和地面曲率存在显著相关性;尽管空间回归模型的预测精度和普通回归模型相近,但前者可以更加准确地表征土壤缓效钾的空间分布格局及空间分异细节特征。

关键词:土壤属性;土壤缓效钾;环境协变量;空间回归模型;土壤预测制图

中图分类号: S159-3

定量化、高精度的土壤属性空间分布信息不仅是现代农业发展的客观需要,也是生态环境建模重要的数据基础。基于不同技术途径的土壤属性预测制图日益成为土壤学相关分支学科普遍关注的热点领域^[1]。相关研究表明,回归分析手段在揭示土壤属性空间分异规律、进行数字化土壤预测制图方面具有简单便捷的技术特点。目前,回归分析模型在土壤学领域中的应用主要集中在 3 个方面:①土壤属性的空间变异特征:基于土壤属性的空间自相关性分析,揭示其自相关距离、块金方差和结构方差等空间变异结构信息^[2];②土壤空间变异影响因素:通过建立回归方程,揭示土壤属性与环境因素之间的相关关系,从而阐明空间变异的主要影响因素^[3];③土壤属性预测制图:基于土壤属性与环境协变量之间的回归关系,对土壤属性实施空间预测与数字化制图。

以往应用于土壤属性预测制图的回归分析模型多为普通多元回归和逐步回归模型,实施过程中包括考虑残差影响和不考虑残差影响两种技术途径^[4-5]。需要指出的是,普通回归分析是一种基于经典统计学理论的常规统计分析技术,其前提假设是数据本身在统计学上的独立性。然而土壤属性普遍存在空间自相关性,这在很大程度上制约了普通回归分析在土壤预测制图中应用的合理性^[6]。

本研究以河南省孟津县为研究区,基于土壤属性空间自相关性及其与环境协变量之间的关系,运用空间回归分析模型实施表层土壤缓效钾预测制图,并将

输出结果与普通回归模型进行对比分析,诠释空间回归分析应用于土壤属性预测制图中的技术优势。

1 材料与方法

1.1 研究区概况

孟津县位于河南省中西部丘陵山区,地理位置为 112°12' ~ 112°49'E, 34°43' ~ 34°57'N,南北长 26.9 km,东西宽 55.5 km,总面积 758.7 km²。海拔范围 110 ~ 481 m,平均海拔 262 m,地形西高东低,中部高,南北低,东部南北两侧为洛河、黄河阶地。孟津县属亚热带和温带的过渡地带,季风环流影响明显,年降雨量为 650.2 mm。县域内成土母质主要为次生黄土母质,分布的主要土壤类型为褐土、潮土、风沙土,其中褐土面积占全县土壤面积的 93%左右。

1.2 样品与分析

在综合考虑研究区成土母质、地形特征和土地利用方式等因素的基础上,采用随机布点的方式,共采集表层土壤样品 2 674 个(孟津县土肥站提供)。采样深度 20 cm,采用 5 点混合采样法,样品经充分混合,在自然状态下风干、磨碎、过筛。土壤缓效钾含量在实验室内用 1 mol/L 硝酸煮沸法测定。

1.3 研究方法

将获取的 2 674 个样点数据按 80% 和 20% 的比例随机分为 2 145 个训练数据和 529 个验证数据两组(图 1),其中训练数据用于土壤缓效钾含量统计分析、与环境协变量之间相关性分析以及研究区土壤缓效

基金项目:国家自然科学基金项目(40971128)资助。

* 通讯作者(jchen@zzu.edu.cn)

作者简介:刘晓冰(1989—),女,河南平顶山人,硕士研究生,主要从事土壤变异及空间预测研究。E-mail: lxbhl2006@126.com

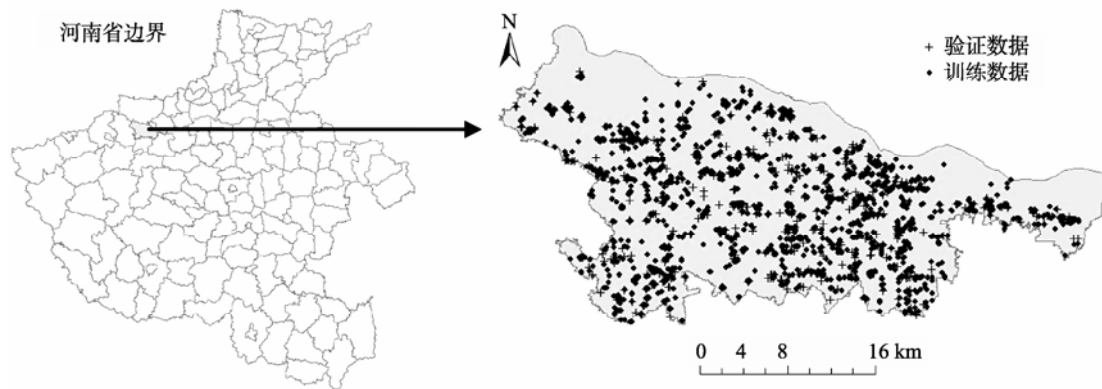


图 1 研究区土壤样点分布图

Fig. 1 Distribution of the sampled sites in study area

钾预测制图,验证数据用于对预测制图输出结果的精度检验。基于先期构建的研究区 1:2.5 万 DEM,利用 ArcGIS9.3 计算坡度、地面曲率和复合地形指数 (CTI) 等环境协变量因子,并通过 Spatial Analysis Tools 中的 Extract Values to Points 工具提取训练样点对应空间位置的环境协变量数据^[7-8]。然后,运用 SPSS 15.0 相关功能模块对表层土壤缓效钾含量与坡度、高程、地面曲率以及复合地形指数等之间的相关性进行分析,进而选取相关性显著的协变量进行多元回归分析。在空间数据分析软件 GeoDa095i 中计算不同阈值下的空间权重矩阵,分析不同空间权重矩阵下数据的空间自相关性,并用 P -value 和 Z -score 检验空间自相关性的显著性,根据相关性及显著性检验结果,选取最佳的空间回归模型进行回归分析,拟合缓效钾与各环境协变量间的回归方程。

2 结果与分析

2.1 缓效钾的描述性统计分析

2 145 个训练样点数据的统计分析结果显示,研究区表层土壤缓效钾平均含量为 865.8 mg/kg,最大值、最小值分别为 1 371.0 mg/kg 和 403.0 mg/kg,变异系数 19.1%,属中等变异^[9]。529 个验证样点的缓效钾平均含量、最大最小值和变异系数,均与训练样点没有明显差异,表明训练数据和验证数据,对研究区表层土壤缓效钾含量均具有很好的代表性(表 1)。

表 1 研究区表层土壤缓效钾含量训练数据和验证数据的描述性统计分析(mg/kg)
Table 1 Descriptive statistics of slowly available K contained in topsoil of study area

样点	样点数	最大值	最小值	平均值	变异系数(%)
训练样点	2145	1371.0	403.0	865.8	19.1
验证样点	529	1371.0	382.0	870.8	19.8
所有样点	2674	1371.0	382.0	866.8	6.2

2.2 空间回归分析

2.2.1 土壤缓效钾与环境协变量相关性 影响土壤形成的环境因素主要有母质、气候、生物、地形、时间和人类活动,而在较小的空间范围内,气候、母质等成土因素的差异性一般通过高程等地形因子体现。地形要素是较小空间内最能反映土壤空间差异的环境因素,也是最为重要的要素。本研究选取高程、坡度、坡向和 CTI 等主要地形要素作为环境协变量对缓效钾进行空间预测。高程是最基本的地形属性,它和气候条件、水文状况、植被分布等一些与土壤形成发育过程密切相关的环境要素均有密切的关系;坡度反映了局部地形表面的倾斜程度,与土壤的稳定程度、地表水流的排聚能力都有直接关系;不同坡向具有不同的日照和通风条件,使土壤质地、土层厚度等出现地域分异;曲率可用于刻画地表物质运动,也影响土壤有机物含量等一些土壤属性的分布。

研究区表层土壤缓效钾含量与环境协变量之间的相关性分析结果列入表 2:2 145 个训练样点土壤缓效钾含量与 CTI 相关系数为 0.013,显著性检验值 $0.533 > 0.05$,二者相关性不显著;土壤缓效钾含量与坡度、高程和地面曲率的相关系数分别为 -0.064 、 -0.076 和 -0.065 ,显著性检验值分别为 0.003、0.000 和 0.003,均小于 0.01,表明高程、坡度和地面曲率 3 个地形因子与土壤缓效钾含量呈极显著相关关系,对土壤缓效钾含量具有重要影响,体现了与之关系密切的气候条件、地表水流和地表物质运动等环境要素对土壤形成和发育的作用。土壤缓效钾含量受自然成土因素影响的同时,耕作施肥等非自然因素也对其起影响作用,赵彦锋等^[10]研究了城市化对区域土壤肥料演变的影响得出施肥过量和施肥不平衡与区域土壤肥料演化有直接关系,稻田转化为菜地加速了土壤属性演化。各种非自然因素的影响干扰了自然成土因

素对土壤的作用,因此高程、坡度和地面曲率等地形因子与土壤缓效钾含量虽然呈极显著相关,但相关系数小。以上分析说明,虽然高程、坡度和地面曲率与土壤缓效钾含量的相关系数小,但相关性极显著,高程、坡度和地面曲率对土壤缓效钾含量具有重要影响,可用于缓效钾的含量预测。基于上述结果,选取坡度、高程和地面曲率等 3 个地形因子作为研究区表层土壤缓效钾含量空间预测模型的环境协变量^[9]。

表 2 缓效钾含量与环境协变量之间的相关性
Table 2 Correlation between slowly available K contained in topsoil and environmental covariates

项目	坡度	高程	地面曲率	CTI
相关系数	-0.064	-0.076	-0.065	0.013
显著性	0.003	0.000	0.003	0.533

2.2.2 空间权重矩阵 空间自相关分析是空间回归分析的前提,而表达数据间空间相邻关系的空间权重矩阵则是空间自相关分析的基础^[11-12]。空间权重矩阵的构建方式主要有两种:基于邻接关系的空间权重矩阵和基于距离的空间权重矩阵^[13]。鉴于土壤样点为离散点这一客观事实,本研究基于距离进行空间权重矩阵的构建。选取的距离阈值最小值为 5 000 m,最大值为 50 000 m,在 5 000 ~ 25 000 m 间以 5 00 m 为间隔递增,25 000 ~ 50 000 m 间以 5 000 m 为间隔递增。

2.2.3 自相关性判断与分析 空间自相关是指同一个研究区内的观测数据之间潜在的相互依赖关系。首先在 Geoda095i 中利用基于距离的空间权重矩阵计算不同空间权重矩阵下的 Moran'I 指数,然后采用蒙特卡罗模拟检验空间相关性,即通过计算 P-value 和 Z-score 判断 Moran'I 指数的显著性。P-value<0.05 表

示空间相关性显著、<0.01 表示极显著^[14];在显著性水平 $\alpha = 0.05$ 条件下,Z-score>1.96 表示数据间存在显著空间正相关性、<-1.96 表示存在显著空间负相关性、-1.96 ~ 1.96 之间说明数据的空间自相关性不显著^[15]。

图 2 显示,随着阈值的增大,Moran'I 指数呈减小趋势,表明数据间的正相关性逐渐减小,当阈值为 45 000 ~ 50 000 m 时呈现负相关性。而 Z-score 随着计算范围的增大,呈先减小后增大趋势,当阈值为 20 500 ~ 22 000 m 时,Z-score 介于-1.96 ~ 1.96 之间,相关性不显著;当阈值>22 000 m 时,Z-score 变化波动较大,说明数据空间相关的显著性不稳定(图 3)。P-value 随阈值的变化表明 P-value 在阈值<16 000 m 时均<0.05,空间自相关性较显著;当阈值>16 000 m 时,P-value 变化波动较大,数据空间相关的显著性不稳定(图 4)。综合考虑 Moran'I 指数、P-value 和 Z-score 随不同阈值的变化情况,选择 10 000 m 作为空间自相关分析的距离阈值,P-value 和 Z-score 分别为 0.000 000 3 和 6.764 804,土壤样点数据具有显著空间自相关性。

2.2.4 模型选择 在 Geoda095i 的空间回归决策过程中,首先考虑阈值为 10 000 m 时,空间滞后模型和空间误差模型的拉格朗日乘子,即标准 LM-lag 和 LM-error 检验统计量。如果 LM-lag 和 LM-error 检验统计量结果均不显著,选取普通回归模型实施土壤缓效钾空间预测制图即可;如果其中一个检验结果显著,需选用相应的空间回归模型。如果上述两个检验结果均显著,则比照分析稳健拉格朗日算子 Robust LM-lag 和 Robust LM-error 的显著性,选取稳健拉格朗日算子显著的空间回归模型^[15]。

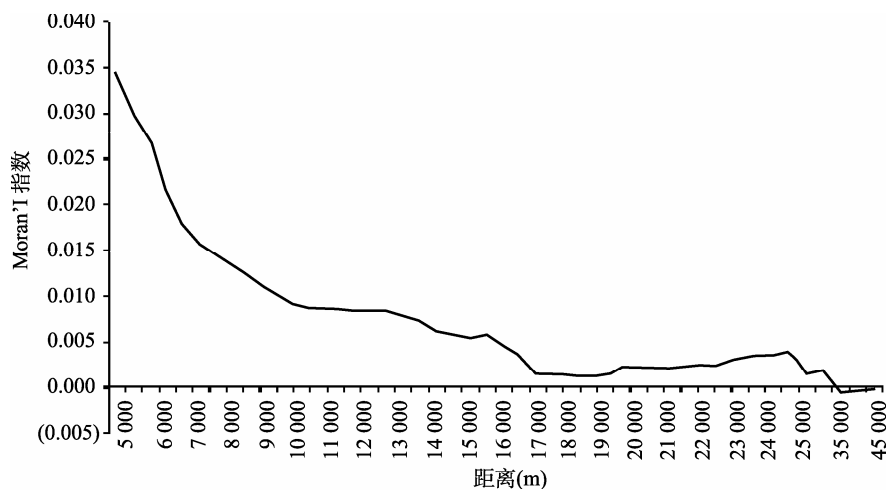


图 2 Moran'I 指数随阈值的变化
Fig. 2 Variation of Moran'I index with threshold

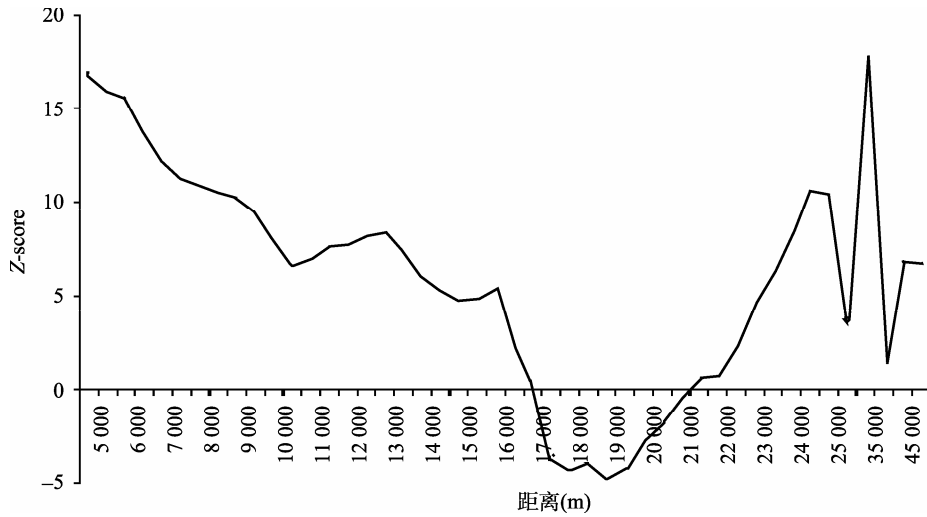


图 3 Z-score 随阈值的变化
Fig. 3 Variation of Z-score with threshold

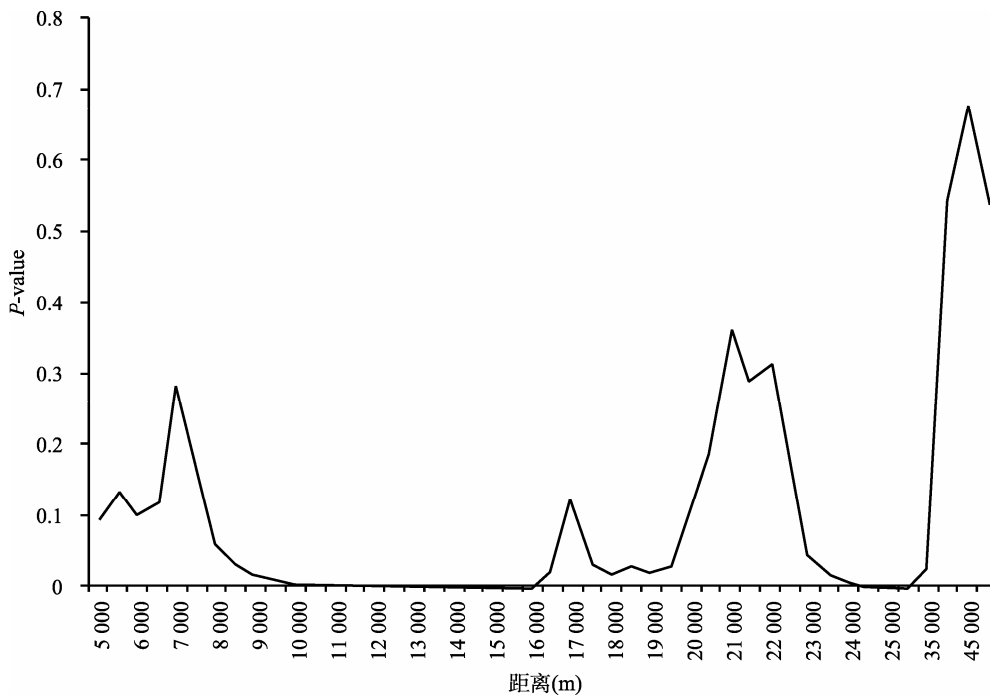


图 4 P-value 随阈值的变化
Fig. 4 Variation of P-value with threshold

表 3 中的检验结果显示 标准 LM-lag 和 LM-error 检验统计量均显著, Robust LM-lag 和 Robust LM-error 检验统计量的 Z-score 均 > 1.96, 表明本研究土壤样点数据空间自相关性显著。然而 Robust LM-error 检验统计量的 P-value > 0.05, 表明 P-value 不符合要求, 因此需选用空间滞后模型, 此模型的分析方法为强行进入法, 拟合的空间回归方程为:

$$y = 0.70897y - 0.038185x_1 + 3.2649425x_2 - 3.274584x_3 + 245.362 \quad (1)$$

式中, y 为表层土壤缓效钾含量, x₁ 为高程, x₂ 为坡

度, x₃ 为地面曲率。

表 3 空间回归模型选择检验
Table 3 Verification of selecting spatial regression model

指数	LM-lag	LM-error	Robust LM-lag	Robust LM-error
Z-score	35.377	27.743	10.424	2.789
P-value	0.000	0.000	0.001	0.095

2.2.5 空间回归模型与普通回归模型对比分析 假若忽略数据的空间自相关性, 基于研究区环境协变量的表层土壤缓效钾含量的普通回归方程拟合如下:

$$y = -0.094\ 634\ 96\ x_1 + 3.049\ 4285\ x_2 - 3.003\ 159\ x_3 + 876.203\ 4 \quad (2)$$

式中, y 为缓效钾含量, x_1 为高程, x_2 为坡度, x_3 为地面曲率。

空间位置土壤属性预测输出结果包括确定性趋势值和非确定性随机误差,其中确定性趋势值通过以环境协变量为因变量的多项式表示,即上文的式(1)和(2);非确定性随机误差通过样点土壤属性实测值减去相应空间位置的确定性趋势值。本研究通过空间

回归模型和普通回归模型获得的研究区表层土壤缓效钾含量确定性趋势输出结果分别作图 5a 和图 5b。提取训练样点的土壤缓效钾含量确定性趋势值两种输出结果,分别计算各样点土壤缓效钾含量非确定性随机误差,利用普通 Kriging 插值手段获取研究区表层土壤缓效钾含量非确定性残差空间分布图(图 6a、6b)。将图 5a 和图 6a、图 5b 和图 6b 叠加,分别获得基于空间回归模型和普通回归模型的研究区表层土壤缓效钾含量空间预测图(图 7a、7b)。

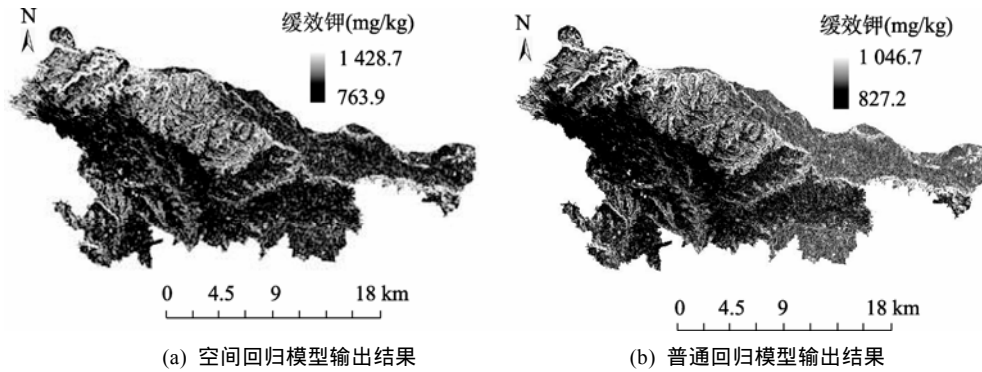


图 5 研究区表层土壤缓效钾含量确定性趋势值空间分布图
Fig. 5 Distribution of spatial trend of slowly available K contained in topsoil of the study area

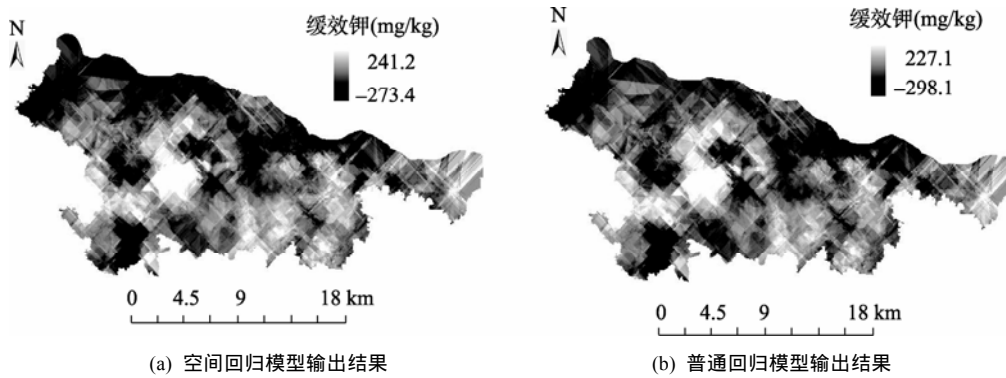


图 6 研究区表层土壤缓效钾含量非确定性残差空间分布图
Fig. 6 Distribution of residual error of slowly available K contained in topsoil of study area

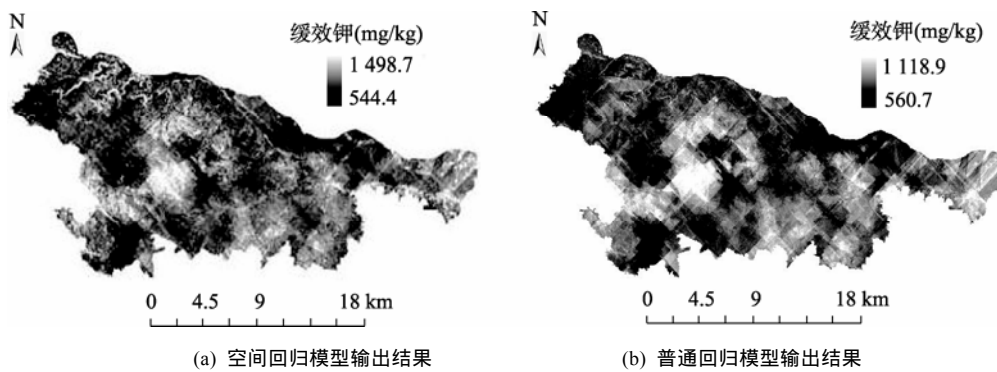


图 7 研究区表层土壤缓效钾含量空间预测图
Fig. 7 Predicated maps of slowly available K contained in topsoil of study area

利用 529 个验证数据,选取均方根误差及实测值和预测值的相关系数分别检验基于空间回归和普通回归模型获取的研究区表层土壤缓效钾含量预测输出结果的精度(表 4)。其中,均方根误差(root mean square error, RMSE)表示样本的离散程度,其计算公式为:

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

式中, n 为测量次数, y_i 为测量值, \hat{y}_i 为平均值。

表 4 空间回归模型和普通回归模型输出结果精度对比
Table 4 Precision comparison between predictive output of the spatial regression model and that of the classic regression one

	普通回归	空间回归
均方根误差	160.3	163.8
相关系数	0.264	0.331

表 4 中结果对比显示,普通回归模型输出结果的均方根误差略小于空间回归分析,而就模型输出预测值与土壤样点实测值之间的相关性而言,空间回归模型优于普通回归模型。因此,仅依据均方根误差和预测值与实测值之间相关性等统计学指标,难以判别两类模型的相对优劣。然而,对比图 7a 和图 7b 可以发

现,空间回归模型与普通回归模型输出的研究区表层土壤缓效钾含量空间分布总体格局相似:中部及东南部高、北部及西北部低。基于普通回归模型获得的研究区表层土壤缓效钾含量预测值范围在 560.7~1 118.9 mg/kg 之间,预测值最小值大于验证数据最小值,二者相差 178.7 mg/kg;预测值最大值小于验证数据最大值,二者相差-252.1 mg/kg;表明普通回归模型输出结果具有明显的平滑压缩现象。基于空间回归模型获取的研究区表层土壤缓效钾含量预测值最小、最大值分别为 544.4、1 498.7 mg/kg,与验证数据缓效钾的最小、最大值分别相差 162.4、127.7 mg/kg,平滑压缩现象明显较弱。

采用分位数法将研究区 529 个验证样点的缓效钾含量数据分为 5 级(根据土壤缓效钾含量从低到高依次为第 1、2、3、4、5 类),每级占验证样点总数的 20%,各级样点的缓效钾含量范围分别为 382.0 ~ 725.0、725.0 ~ 822.0、822.0 ~ 909.0、909.0 ~ 1 009.0 和 1 009.0~1 371.0 mg/kg。按照上述分级标准,将基于两种回归模型的输出结果在空间上进行可视化表达(图 8),不同缓效钾含量级别的土壤分布面积见表 5。

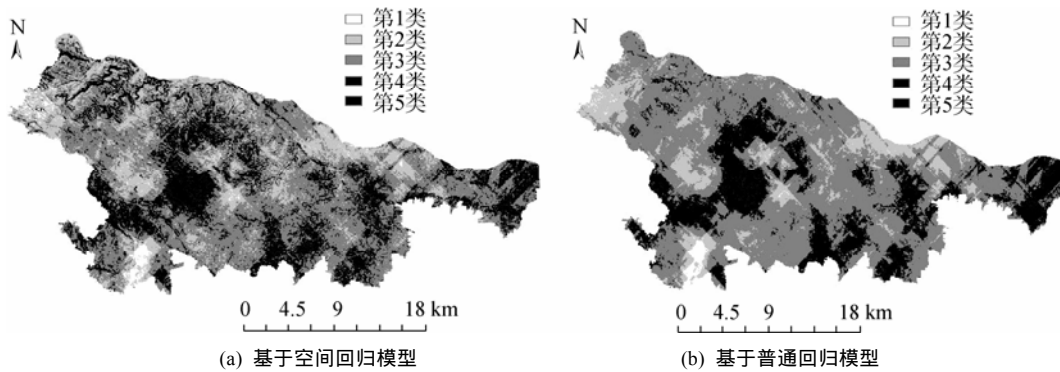


图 8 研究区不同缓效钾含量的土壤分布面积预测图

Fig. 8 Predicated maps showing distribution of topsoil with different levels of slowly available K content

表 5 研究区不同缓效钾含量水平的土壤分布面积百分比(%)

Table 5 Predicated coverage of soils with different levels of slowly available K contained in topsoil

	第 1 类	第 2 类	第 3 类	第 4 类	第 5 类
空间回归	5.2	21.9	38.7	26.9	7.3
普通回归	0.9	12.8	59.5	26.1	0.7

表 5 中计算结果显示,两种回归预测模型的输出结果中,研究区表层缓效钾含量属于第 3 级的土壤分布面积均大于实测样点分布的实际面积,第 2、5 级的分布面积则小于实际面积,第 4 级与实际分布面积最为接近。比较而言,基于普通回归模型输出结果获得的不同缓效钾含量水平分布格局与实际

情况的偏离尤为明显,预测结果在属性空间上过度集中于中值区域分布。以上分析表明,尽管空间回归模型与普通回归模型输出结果均具有平滑压缩效应,但相比之下,空间回归模型的预测输出结果能相对准确地表征和揭示研究区表层土壤缓效钾含量的分布格局和空间变异的细节特征,主要原因可能是空间回归模型考虑了土壤属性的空间自相关性,有效地降低了对数据的平滑压缩效应。

3 结论

(1) 研究区表层土壤缓效钾含量数据具有显著的空间自相关特征,空间自相关距离阈值约为

10 000 m。

(2)表层土壤缓效钾含量与研究区坡度、高程和地面曲率3个环境协变量存在显著相关性,而与复合地形指数(CTI)的相关性不显著。

(3)基于普通回归模型与空间回归模型的表层土壤缓效钾含量预测输出结果均具有平滑压缩效应,前者这方面的表现尤为明显。

(4)相比而言,空间回归模型由于考虑了土壤属性的自相关性,能够有效降低预测输出结果的平滑压缩作用;尽管预测精度与普通回归模型相近,但空间回归模型能相对准确地表征研究区表层土壤缓效钾含量的空间分布格局和变异特征。

参考文献:

- [1] Chien YJ, Lee DY, Guo HY. Geostatistical analysis of soil properties of mid-west Taiwan soils[J]. *Soil Science*, 1997, 162(4): 291-298
- [2] 霍霄妮, 李红, 张微微, 孙丹峰, 周连第, 李保国. 北京耕作土壤重金属多尺度空间结构[J]. *农业工程学报*, 2009, 25(3): 223-229
- [3] 于洋, 刘吉平, 徐艳艳. 东北典型黑土区土壤养分空间分异影响因素分析[J]. *水土保持研究*, 2009, 16(45): 66-69
- [4] 连纲, 郭旭东, 傅伯杰, 虎陈霞. 黄土高原县域土壤养分空间变异特征及预测——以陕西省衡山县为例[J]. *土壤学报*, 2008, 45(4): 577-584
- [5] 孙孝林, 赵玉国, 张甘霖, 李德成, 张甘霖. 应用土壤-景观定量模型预测土壤属性空间分布及制图[J]. *土壤*, 2008, 40(5): 837-842
- [6] Shawna JD. The biogeography of invasive alien plants in California: An application of GIS and spatial regression analysis[J]. *Diversity and Distributions*, 2004, 10: 1-9
- [7] 连纲, 郭旭东, 傅伯杰, 虎陈霞. 基于环境相关法和地统计学的土壤属性空间分布预测[J]. *农业工程学报*, 2009, 25(7): 237-242
- [8] Neil JM, Philip JR. Spatial prediction of soil properties using environment correlation[J]. *Geoderma*, 1999, 89: 67-94
- [9] 张玉铭, 毛任钊, 胡春胜, 张佳宝, 朱安宁. 华北太行山前平原农田土壤养分的空间变异性研究[J]. *应用生态学报*, 2004, 15(11): 2 049-2 054
- [10] 赵彦锋, 史学正, 黄标, 张学雷. 太湖流域 20 年来的施肥变化与土壤属性演变——以无锡市为例[J]. *土壤通报*, 2009, 42(2): 297-301
- [11] 王红亮, 胡伟平, 吴驰. 空间权重矩阵对空间自相关的影响分析[J]. *华北师范大学学报(自然科学版)*, 2010(1): 110-115
- [12] Dayton ML, James LD. A comparison of four spatial regression models for yield monitor data: A case study from Argentina[J]. *Precision Agriculture*, 2004, 5: 579-600
- [13] Chi GQ, Zhu J. Spatial regression models for demographic analysis[J]. *Population Research and Policy Review*, 2008, 27: 17-42
- [14] 梁二, 王小彬, 蔡典雅, 刘爽, 王燕. 河南省土壤有机碳分布空间自相关分析[J]. *应用生态学报*, 2007, 18(6): 1 305-1 310
- [15] 霍霄妮, 李红, 张微微, 周连第, 李保国. 北京耕作土壤重金属含量的空间自相关分析[J]. *环境科学学报*, 2009, 29(6): 1 339-1 344

Study on Predicated Mapping of Soil Property Based on Spatial Regression Analysis

LIU Xiao-bing¹, CHENG Dao-quan², LIU Peng-fei¹, SONG Xuan¹, CHEN Jie^{1*}

(1 School of Environment and Water Conservancy, Zhengzhou University, Zhengzhou 450001, China;

2 Soil and Fertilizer Station of Henan Province, Zhengzhou 450002, China)

Abstract: In this paper Mengjin County of Henan Province we selected as the study area, slowly available K contained in topsoil was spatially predicated by the method of spatial regression analysis with the terrain characteristics, included slope, elevation, curvature and CTI. The results indicated that the slowly available K had a distance threshold of spatial autocorrelation around 10 000 m and a significant correlation between slope, elevation, curvature and soil available K. Although output precision of the spatial regression model was quite similar to that of the classic regression ones, the spatial regression model displayed an evident advantage to reveal further the spatial distribution pattern and variation characteristics of soil property.

Key words: Soil property, Slowly available potassium, Environmental covariate, Spatial regression model, Predictive soil mapping