

地理加权回归及其在土壤和环境科学上的应用前景^①

瞿明凯¹, 李卫东^{2*}, 张传荣², 黄 标¹

(1 中国科学院土壤环境与污染修复重点实验室(南京土壤研究所), 南京 210008;

2 康涅狄格大学地理系, 美国康涅狄格州斯托斯市 06269)

摘 要: 地理加权回归(GWR)是近些年来出现的一种新的空间局部回归技术。它是将数据的空间位置嵌入线性回归模型中, 以此来探测空间关系的非平稳性。在运用于空间数据分析方面, 相对于传统的普通最小二乘回归法, 具有明显的优势。本文首先介绍了 GWR 的理论起源并描述了该方法的基本原理、存在的不足以及后来的扩展; 然后为了更准确地了解 GWR 的应用状况和研究进展, 进行了一个文献调查; 接着回顾了 GWR 在土壤和环境科学上的初步应用情况; 最后对该方法在土壤和环境科学上的应用前景作了展望。目的是为我国土壤和环境科学领域的同行了解和应用 GWR 提供参考。经过国内外研究者多年的研究和实践, GWR 方法已被证明是一个理论上较为成熟, 能够应用到实际研究中的优秀空间统计学方法。因此, GWR 在土壤和环境科学上将会有着广泛的应用前景。

关键词: 地理加权回归; 非稳态; 普通最小二乘法; 土壤和环境科学; 空间数据分析

中图分类号: S15; X5

土壤和环境属性的预测性制图对精准农业生产和环境保护治理至关重要。但为提高制图质量而在大尺度上进行大量高密度田间取样在经济和劳力花费等方面都是不现实的, 对于地形复杂和偏远地区尤其如此。所幸许多土壤和环境属性在空间上往往与一些环境因子(如地形、景观、气候、地质和人类活动等)密切相关, 其中一些环境因子的数据较为容易获得。那么利用目标变量和环境影响因子变量之间的相关性建立定量回归关系, 则可以达到以有限稀疏样本数据进行土壤和环境属性制图的目的。比如土壤有机质含量, 常常与如成土母质、地面高程和土地利用方式等自然和人文环境因子有关, 因此可以利用易获取的环境因子数据推导土壤有机质含量的空间分布。回归分析是处理变量间相关关系的一种经典统计方法, 由于其简单、易用、计算效率高和易于理解的特点, 已经被广泛应用于土壤及环境的预测、监测、评价等多个领域。

基于普通最小二乘法(OLS)的多元线性回归模型(MLR)是一种总体回归技术, 因其具有完备的理论体系和统计推断方法, 常用来确定和分析目标变量和影响因子变量之间的关系, 在土壤和环境领域有着非常

广泛的应用。该模型成立的前提假设是观测值之间彼此独立。然而在空间上, 事物普遍与周围的其他事物是相互关联的, 只是距离近的事物通常比距离远的关联程度更大, 这一现象也被称为地理学第一法则^[1]。故经典的多元线性回归模型并不非常适用于分析非平稳性的空间数据。于是局部回归技术应运而生。局部回归技术将数据的空间结构嵌入回归模型中, 使回归参数变成观测点地理位置的函数。Pelto 等^[2]于 1968 年最早将此类局部回归思想用于处理非等间隔分布的高程数据, 得到了满意结论。为了寻找一种可以强化从散点图抽取的可视信息的工具, 该类方法于 1979 年被 Cleveland^[3]引入统计学文献中。由于 Ripley^[4]和 Fedorov^[5]的支持, 该方法被广泛运用于空间统计学领域。近年来, 由 Fotheringham 和 Brunson 等人^[6-8]在局部回归的基础上利用局部光滑技术, 提出了地理加权回归(GWR)。

目前, GWR 已经发展成为一门完善成熟的空间统计学技术, 同时也有相对成熟的软件, 如: GWR3.0 和 ArcGIS9.3 地理加权回归模块等。国内关于 GWR 的理论和应用研究还比较少, 且主要在人文地理和统计学领域中。考虑到 GWR 也很适用于制图土壤和环

基金项目: 国家自然科学基金面上项目(40971269, 41371227)、中国博士后科学基金项目(2013M530273)、中国科学院知识创新工程重大项目(KSCX1-YW-09-02)和公益性行业(农业)科研专项经费项目(200903001-01)资助。

* 通讯作者(weidongwoody@gmail.com)

作者简介: 瞿明凯(1983—), 男, 湖北荆州人, 博士后, 主要从事资源环境信息学方面的研究。E-mail: qumingkai@issas.ac.cn

境属性,但目前国内土壤和环境领域研究人员对该方法还比较陌生,本文后面部分将对该方法作一系统介绍。同时为了更准确地了解GWR的应用和进展状况,本文也作了一个文献调查。随后本文回顾了GWR在土壤和环境科学上的初步应用情况,并对该方法在土壤和环境科学上的应用前景作了展望。

1 GWR 理论

1.1 地理加权回归模型及其估计

在GWR中,对于每个模型校准位 u ,GWR模型为:

$$y(u) = \beta_0(u) + \sum_{k=1}^p \beta_k(u) \cdot x_k(u) + \varepsilon(u) \quad (1)$$

其中 $y(u)$ 为在位置 u 的因变量值, $x_k(u)$ 为位置 u 的第 k 个协变量的值, $\beta_0(u)$ 是截距项, $\beta_k(u)$ 是第 k 个协变量的回归系数, p 是回归项的个数, $\varepsilon(u)$ 是位置 u 的随机误差项。这个模型与普通线性回归模型最明显的区别是该模型的回归系数在每个空间数据位置都要分别被估计,而普通线性回归模型对回归系数的估计是基于整个研究区域的。为了便于论述,GWR模型通常被表示为如下的矩阵形式:

$$y(u) = \mathbf{X}(u)\boldsymbol{\beta}(u) + \varepsilon(u) \quad (2)$$

其中 $\boldsymbol{\beta}(u)$ 为位置 u 的回归系数列矢量, $\mathbf{X}(u)$ 为位置 u 的自变量行矢量。位置 u 的回归系数的估计值为:

$$\hat{\boldsymbol{\beta}}(u) = [\mathbf{X}^T \mathbf{W}(u) \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{W}(u) \mathbf{Y} \quad (3)$$

其中 \mathbf{Y} 为因变量的 $n \times 1$ 矢量; $\mathbf{X} = [\mathbf{X}_1^T, \mathbf{X}_2^T, \dots, \mathbf{X}_n^T]^T$ 是自变量的设计矩阵,它包含了一列1截距项; $\mathbf{W}(u) = \text{diag}[W_1(u), K, W_n(u)]$ 是位置 u 的 $n \times n$ 对角权重矩阵; $\hat{\boldsymbol{\beta}}(u) = (\hat{\beta}_0(u), \hat{\beta}_1(u), K, \hat{\beta}_p(u))^T$ 是位置 u 的回归系数矢量,包含 p 个自变量的回归系数和1个截距项的回归系数。这里 n 为局部回归所用的观测数据点数, p 为自变量个数。根据等式(3),GWR可被视为局部加权最小二乘回归模型。要计算模型在位置 u 的回归系数,必须先计算周围数据位置相对于该待估位置的权重矩阵。

1.2 空间权重函数

GWR模型的核心是空间权重矩阵^[9],它是通过选取不同的空间权重函数来表达对数据间空间关系的不同认识。空间权重函数的正确选取对GWR模型参数的正确估计至关重要。GWR的权重函数主要分为两类:一类是固定权重函数,它在整个研究区域内是固定的;另一类是自适应权重函数,它能根据校准位置周围的数据密度进行相应调整^[8]。距离阈值法是

最简单的权重函数选取方法,它的关键是选取合适的距离阈值。由于该方法并不符合随着距离的增加而权重不断衰减的思想,尽管其简单,但没有被广泛采用。在GWR模型的实际应用中,倾向于选择随着距离增加权重单调递减且连续的权重函数。如高斯权重函数^[8]

$$W_{ij} = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{r}\right)^2\right] \quad (4)$$

其中的权重带宽参数 r 控制观测点空间相关性的范围和衰减模式。在这个函数中,观测点 j 相对于待估点 i 的权重是距离 d_{ij} 和带宽参数 r 的函数。

1.3 空间权重函数带宽参数的选择

在校准GWR模型之前,研究者必须确定一种权重函数。Wheeler等^[10]指出,相对于权重函数的具体函数形式,GWR模型中带宽参数对预测结果的影响更大。而且,在以上所有权重函数中,必须首先选择或从数据中估计未知的带宽参数。此时,迫切需要一种能够选择带宽的标准。在GWR模型中,目前主要有3种估计权重函数带宽参数的方法:直接分配最近邻居数带宽法^[11],交叉验证法^[3]和校正的Akaike信息标准(corrected Akaike information criterion, AIC_c)法^[12]。其中运用最广泛的是Cleveland^[3]于1979年提出的交叉验证(CV)法。当估计了最优带宽后,即可确定权重函数。每个模型校准位的权重矩阵可由该被估计的权重函数得到,进而得到被估计的局部回归系数。

1.4 存在的不足

GWR模型主要存在以下4个方面的不足之处:

(1) 从根本上来说,GWR模型没有为变异源提供一个基本模型,因此将其当作一种启发式方法更合适。也可认为GWR模型缺乏统一的统计框架。它本质上是局部地理回归的结合,不同数据位置的回归系数之间的依赖关系没有被包含到该局部地理回归模型中^[10]。

(2) 在不同的模型校准位,反复使用数据来估计模型参数,会产生多重比较的情况。这样,随着被估计的局部模型越来越多,显著性水平测试有意义的概率会越来越来大^[10]。而这未必是实际情况的反映。

(3) 不恰当的带宽参数会引起回归系数出现强烈的空间变异或平滑。如果带宽很大,包含了大量的观测,那么被估计的回归系数的空间变异性将很小或几乎没有;如果带宽很小,被估计的回归系数的空间变异性将很大^[10]。这样,被估计的回归系数的空间模式被人为地引入,而这未必代表真实的回归效果。

(4) 回归系数的标准误差。由于在多个位置重复

使用数据来估计参数^[13-14]，同时由于使用数据在交叉验证的基础上估计带宽进而得到回归系数^[15]，这些都会导致 GWR 模型回归系数的标准误差只是实际误差值的一个近似值^[10]。

(5) GWR 考虑的主要是因变量和自变量间的交叉相关问题，而假设了自变量间独立，也就是说，自变量间的交叉相关没有考虑进去。尽管在回归系数估算时考虑了观测数据的地理权重，这在一定程度上隐含地包括了数据的空间自相关效果，但它并没有明确地考虑自变量和因变量的自相关问题。由于参数的局部化，明确地考虑自相关相对不易。

随着研究实践的不断深入，对 GWR 不足的认识不断加强。尽管如此，但由于该模型考虑了周围空间数据关系，在空间数据分析方面较常用的多元线性回归模型具有明显优势，依然是目前最有效的空间局部回归方法。同时 GWR 模型还在不断发展之中，自身的一些缺陷也将不断得到修正或解决，在空间插值和变量间关系空间分布变化分析上具有广泛应用前景。

1.5 诊断工具

对于普通 MLR 模型，有很多比较著名的诊断工具可以用于诊断如自相关、影响点、共线性等问题。更复杂的线性回归模型 GWR 也有相应的诊断工具。

Leung 等^[16]于 2000 年在比较完善的 Moran's I 和 Geary's c 这两个空间自相关统计量基础上提出了确定 GWR 模型空间残差自相关的方法。该方法将每个因变量的局部预测值与实际观测值对比以得到残差的空间分布模式。由于该统计量不是基于模型的，因此不能确定自相关是否被人为引入。另一个空间依赖测试方法由 Páez 等^[9]于 2002 年提出。该方法基于方差异性模型，但并没有一个将所有局部模型联系在一起的统一框架。由于无法确定 GWR 系数强烈的相关性是否来自模型的局部共线性，故研究者在估计 GWR 系数之前应先进行共线性检验。目前有几种诊断工具可以确定 GWR 模型中是否有严重的共线性效应。除了局部回归系数散点图、近似局部回归系数相关图^[17]和局部方差膨胀因子(VIFs)^[18]外，还有方差分解比和相应的条件指数法^[18-19]。在量化由于共线性引起的回归系数的方差方面，方差分解法较 VIFs 的优点是前者能够同时度量和表达所有回归项(包括截距项)在共线性方面的特点。方差分解比和条件指数诊断工具由 Belsley^[18]于 1991 年提出，Wheeler^[19]于 2007 年加以修改从而运用于 GWR 模型中。该方法对 GWR 权重设计矩阵进行单值分解来得到系数协方差矩阵的条件指数和方差分解比。方差分解比是回

归系数方差的比率，它可由一部分方差矩阵分解来解释。同时，该方法还有一个附属的条件指数，该指数为分解的最大奇异值和最小奇异值的比率。

Belsley^[18]于 1991 年提出了在 OLS 回归环境下运用方差分解比和条件指数法的一些相关准则。一般情况下，强烈的共线性对应于大的条件指数。另一个准则是对于同一个方差分量，当两个或多个方差分解比大于 0.5 时，即说明存在共线性。这些准则同样适用于诊断 GWR 模型中的共线性。这样，方差分解比和条件指数诊断工具能够揭示 GWR 模型校准位的局部共线性。进一步，可以做出与局部回归系数对应的诊断值的空间分布图。

1.6 扩展

为了扩展 GWR 分析的适用性，一些不同的模型被提出并被结合到 GWR 模型中。其中 3 个扩展如下：

(1) 自回归 GWR。第一个对 GWR 模型的扩展是为了在该模型中结合空间相关性^[20]。当在包含空间自回归成分的 GWR 模型中使用互验证来估计回归系数时在计算方面遇到很大挑战，因为该过程需要计算 $(n-1) \times (n-1)$ 矩阵的行列式 n 次。Brunsdon 等^[20]于 1998 年提出在随机选取的子样本点基础上进行互验证。但 Farber 和 Páez^[21]于 2007 年证明，由于互验证中影响点的存在，这个方法是不适合的。第二个基于 GWR 的空间自回归局部模型于 2002 年被 Páez 等^[22]提出。该方法采用非恒定方差模型，能够参数化模型系数(包括权重函数的带宽参数)估计。第三个空间自回归局部估计模型是 Pace 和 LeSage^[23]于 2004 年提出的，该方法是基于估计矩阵的分解。Mur 等^[24]于 2008 年又提出了 ZOOM 模型。

(2) 约束 GWR。通过对回归系数变异性的程度加以限制可以解决由于共线性而产生的问题。在 GWR 模型中，目前有两种方法可以解决该问题。这两种方法为 Wheeler^[19,25]分别于 2007 年提出的地理加权岭回归(geographically weighted ridge regression-GWRR)和 2009 年提出的地理加权 lasso(GWL)。顾名思义，这两方法分别是基于岭回归^[26]和 lasso 技术^[27]。这两种方法均引入了对回归系数规模的限制。在 GWL 模型中，由于回归系数使用了绝对值，使得估计为非线性。幸运的是，目前有足够多的算法来估计其参数^[25]。

(3) 地理加权的 Logistic 和 probit 模型。除了以上介绍的线性框架，地理加权的思想已经被运用于对名义变量进行建模。Atkinson 等^[28]于 2003 年提出了地理加权 logistic 模型。Páez^[29]于 2006 年提出了地理加权 probit 模型。

2 GWR 文献调查

本文上面部分对 GWR 相关的理论方法进行了介绍,为了更准确地了解 GWR 的应用和研究状况,有必要对涉及 GWR 的文献进行调查。在关于 GWR 应用和研究的广度方面,为了给研究者提供有价值的信息,本文选择 Web of Science 数据库的科学引文索引(SCIE)和社会科学引文索引(SSCI)中关于 GWR 的文献进行了一个调查。对这两个索引检索关键词“Geographically Weighted Regression”同时通过作者检查,共 324 篇文章涉及 GWR 模型或对 GWR 模型进行了理论方面的研究。其中 SCIE 收录 216 篇,SSCI 收录 185 篇(77 篇文章同时被两索引检索)。这说明该方法目前已经被广泛运用于社会科学和自然科学领域。下面文献分析以这 324 篇文献为样本进行。

2.1 文献发表年度分布

表 1 给出了关于 GWR 的论文的发表年度分布情况。从表 1 可以看出,运用 GWR 模型的理论或应用论文呈逐年增多的趋势。特别是最近几年,发表论文数增长显著,77.16% 的文章(250 篇)为 2006 到 2011 年发表。

2.2 文献出版期刊分布

324 篇文献分布在 181 个学术期刊。总共有 10 种期刊收录超过 6 篇与 GWR 有关的文章,占总文章的四分之一(82 篇文章)。这 10 种期刊发表 GWR 有

表 1 GWR 有关论文发表年度频率分布
Table 1 Frequency of GWR-related papers by year

年	篇数	比例(%)
2010—2011	120	37.04
2008—2009	81	25.00
2006—2007	49	15.12
2004—2005	34	10.49
2002—2003	18	5.56
2000—2001	9	2.78
1998—1999	11	3.40
1996—1997	2	0.62
共计	324	100.00

注:比例 = 篇数/324 × 100%,下同。

关论文情况见表 2。其中 期刊 ENVIRONMENT AND PLANNING A 收录的关于 GWR 的文章最多(16 篇,4.94%)。在这些收录关于 GWR 文章最多的 10 个期刊中,人文地理方面的占 6 个(即 ENVIRONMENT AND PLANNING A, ANNALS OF THE ASSOCIATION OF AMERICAN GEOGRAPHERS, GEOGRAPHICAL ANALYSIS, INTERNATIONAL JOURNAL OF HEALTH GEOGRAPHICS, JOURNAL OF REGIONAL SCIENCE, JOURNAL OF GEOGRAPHICAL SYSTEMS),其他为生态、环境和遥感方面的,这可能是由于该方法最先出现在人文地理领域,也最先为该领域的研究者所熟悉利用的原因。

表 2 GWR 论文在主要发表刊物上的分布
Table 2 Frequency of papers by journals

期刊名	篇数	比例(%)
ENVIRONMENT AND PLANNING A	16	4.94
GISCIENCE & REMOTE SENSING	9	2.78
ANNALS OF THE ASSOCIATION OF AMERICAN GEOGRAPHERS	8	2.47
GEOGRAPHICAL ANALYSIS	8	2.47
INTERNATIONAL JOURNAL OF HEALTH GEOGRAPHICS	8	2.47
GLOBAL ECOLOGY AND BIOGEOGRAPHY	7	2.16
INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE	7	2.16
JOURNAL OF REGIONAL SCIENCE	7	2.16
ECOLOGICAL MODELLING	6	1.85
JOURNAL OF GEOGRAPHICAL SYSTEMS	6	1.85

3 GWR 在土壤和环境科学中的应用

经典的多元线性回归模型利用全局参数,即假设变量间的相互关系在空间上是平稳的。因此经典的多元线性回归模型并非分析空间数据的有效方法。而 GWR 由于考虑了变量间相互关系的空间非平稳性,因此在应用于空间数据分析和制图方面较经典线性

回归具有明显的优势,已引起越来越多的研究者的关注。在土壤和环境领域,近年来也出现了 GWR 技术的应用研究。

Foody^[30]于 2003 年通过一个实例研究,将 GWR 模型引入环境领域。作者研究了北美和中东的归一化植被指数(NDVI)和降雨量之间的关系。结果显示 GWR 模型能够描述两研究对象之间的空间非稳态

性,而传统的 OLS 模型只能描述两研究对象之间大致的总体关系。

Tu 和 Xia^[31]于 2008 年利用 GWR 研究了美国马萨诸塞州东部的土地利用和水质的关系,同时也使用了普通 MLR 模型作为参照。GWR 模型在 F 检验, R^2 比较和 AIC_c 方面均较普通 MLR 模型优越。研究显示,土地使用和水质之间的关系并不是恒定不变的,而是存在空间非平稳性。另外,GWR 模型能够揭示以前被普通 MLR 模型所掩盖的信息,因此提高了解释局部水质情况的能力。

Zhang 等^[32]于 2009 年使用 GWR 等多种不同方法(局部 Moran 统计量,克里格交叉验证和 GWR)研究了爱尔兰高威市城市土壤铅的离群值。研究表明这 3 种方法一般能识别出高含量的离群值,但是在识别空间离群值方面也存在一定差异:Moran 方法在识别低值的离群值方面更有效,克里格交叉验证方法能有效识别高值离群值,而 GWR 识别的空间离群值在大小上是随机分布的。这一研究结果为准确划定城市污染土地提供了有价值的信息。

Li 等^[33]于 2010 年调查了中国深圳市地表温度(LST)和环境因子之间的尺度依赖和空间非稳态性关系。使用的环境因子包括林地、湿地、建设用地和未利用地的地块丰富密度、城市和高速公路之间的距离、和地形因子(高程和坡度)。结果显示:相对于传统的 MLR 模型,GWR 模型不但提供了更好的拟合,而且提供关于局部 LST 的详细信息。使用 GWR 模型,回归关系的强度显著增加,平均 59% 的 LST 改变量可以被预测变量解释,而使用普通 MLR 模型该值为 43%。通过计算平稳性指标,可以发现不同的预测模型可以得到不同的变异趋势,而且变异趋势随着空间尺度的扩大趋向静态。该研究还显示,潜在的空间过程影响地表温度而且在不同的空间尺度上展现不同的空间模式。最后该作者认为在地理和生态问题上,采用 GWR 模型解决空间非稳态和尺度依赖问题是个很好的选择。

Mishra 等人^[34]2010 年使用 GWR 方法对美国中西部地区的土壤有机碳含量进行了预测。该研究用地形属性、气候数据、土地利用数据、基岩地质数据和 NDVI 数据来预测美国中西部 7 个州土壤有机碳的含量。同时作者将该方法与多元线性回归(MLR)和回归克里格(RK)方法相比较,结果显示 GWR 模型取得明显优势:相对于 MLR 模型和 RK 模型,GWR 模型的预测结果分别提高了 22% 和 2%。

Zhang 等^[35]于 2011 年利用降雨量、土地覆盖类型和土壤类型等环境因子作为因变量来构建 GWR 模

型,对爱尔兰土壤有机碳的空间分布进行了预测。同时普通克里格(OK)、反距离权重(IDW)模型和多元线性回归(MLR)模型被用于参照方法,以验证 GWR 模型的预测效果。结果显示 GWR 取得了更为理想的预测结果。

Wang 等^[36]于 2012 年比较了 GWR 模型和 RK 模型对土壤有机碳的预测效果。采用的数据有土壤有机碳田间采样数据和相关环境因子(高程、坡度、有色金属矿产指数和归一化植被指数)。结果显示相对于 RK,由 GWR 得到的土壤有机质插值结果与环境因子的相关程度更高,且不易受样本极值影响,具有更高的插值精度。

Kumar 等^[37]于 2012 年利用 GWR 的扩展模型地理加权回归克里格(GWRK)预测了美国宾夕法尼亚州 1 m 深内的土壤有机碳储量。所用的环境因子包括温度、降水量、海拔、坡度、地质、土地使用、归一化植被指数等。结果显示 GWRK 不但结合了样本观测数据的空间非平稳性,而且考虑了残差数据的空间自相关性,因而得到的结果较 RK 模型结果更精确。

Mishra 等^[38]于 2012 年利用 GWR 模型估算了美国阿拉斯加州土壤有机碳储量。该研究对该地区的活动层和永久冻土层的储量分别予以估计。得到了较以前研究更精准的结果。

郭龙等^[39]2012 年利用 GWR 模型预测了宜昌市红花套镇土壤 pH、有机质、有效磷、速效钾、碱解氮的空间分布。同时,协同克里格模型被用于参照方法。结果显示,相对于协同克里格模型,GWR 模型能够更方便地结合多个辅助变量。同时,GWR 模型的预测误差及跨度更小,因而该模型更适用于对土壤属性空间变异性的分析和预测。

总之,GWR 在土壤和环境科学上的应用还十分有限,尚需要土壤和环境学家的进一步尝试。

4 讨论与结论

GWR 将数据的空间位置嵌入线性回归模型中,以此来探测空间关系的非平稳性。而经典多元线性回归模型则假定环境变量和因变量之间的关系是恒定不变的。大量研究表明,不同采样尺度下的很多土壤和环境属性具有中等到强烈的空间依赖性^[40-43],因此在运用于土壤和环境空间数据分析和制图方面,GWR 模型相对于传统的线性回归模型无疑具有优势。GWR 既可以考虑多个环境影响因子又可考虑空间局部变化,而经典的地统计学方法则较难(如:协同克里格通常只考虑一个辅助变量,多了则难以操作,需要施加很多限制)。当所要预测的变量无辅助

变量时,传统的插值方法如克里格及一些新出现的插值方法更适合。但克里格模型使用的方差函数通常是基于整个研究区域的,即在整個克里格过程中一直不变。这样,部分局部信息会被忽视。Harris^[44]于 2010 年将 GWR 的思想结合进方差函数,并运用于克里格模型的预测,取得了较好的效果。这样,GWR 也为传统的地统计学注入了新的活力。

与传统多元线性回归模型相似,GWR 模型也受共线性影响很大。所以,在运用 GWR 模型之前,先要检查数据质量,以探测数据是否具有强烈共线性。Wheeler 等^[10]认为,相对于权重函数的具体函数形式,GWR 模型中带宽参数对预测结果的影响更大。一般来说,规则布点倾向于选择固定权重函数,随机布点倾向于选择自适应权重函数。随着研究的不断深入,GWR 模型不断得到扩展(见 1.6 节)。这些扩展增加了 GWR 模型的运用范围和预测的能力。同时发现的问题也越来越多(见 1.4 节)。该模型还在发展之中,随着研究的不断深入,GWR 模型的适应面越来越广,分析能力越来越强。但是如果没有适合的软件,该方法也只能停留在理论上,不可能广泛地应用。界面友好,功能强大的软件对于 GWR 技术能否广泛运用于实践至关重要。因此,开发功能更强大的软件,使其集成最新功能,也是 GWR 的一个重点领域。

GWR 的出现并不意味着传统的线性回归模型不再具有价值。一方面,并不是在所有情况下参数都随地理空间的变化而变化,有些情况下参数在空间上是稳定的,或者变化非常小可以忽略不计^[7,45-46]。对于这些在空间上基本不变的参数,传统的线性回归模型依然适用。另一方面,传统线性回归主要是从总体上探测一个因变量与一到多个自变量之间的相关关系,其应用目的并非都是空间制图;而 GWR 则主要是一种预测性制图方法,既可以对因变量的空间变化进行预测制图,也可以对因变量与自变量间的相关关系的空间变化分布进行预测制图,因此具有其特有的应用价值。

5 GWR 在土壤和环境科学中的应用前景

地学和环境信息具有明显的空间非均质特性,除了自相关外,不同变量间的相关特性在空间上也往往是非稳态的,这就需要能够有效处理非稳态相关性的分析手段来辅助分析。GWR 作为在 OLS 回归基础上发展起来的一门局部空间统计技术,正可以满足土壤和环境科学领域有关方面的需要。GWR 既可以考虑多个环境影响因子又可考虑空间局部变化,而传统的

地统计学方法则在这些方面有明显欠缺(即操作上难以考虑多个环境因子的影响),这样 GWR 则可以弥补地统计学的不足。许多土壤和环境属性的空间分布是决定于多个环境因子的(也叫预测或解释变量,如地形要素,土地覆被利用状况,地质要素,和其他环境土壤属性等),这样在 GWR 模型下由丰富的环境因子数据加一定量的样本点即可进行较高精度预测性制图。

由于 GWR 方法出现的时间比较晚,实际的应用只有十来年,所以该方法在理论和实际应用上都还有许多有待探索和解决的问题。同时 GWR 模型不是单一的一种分析技术,它是一类分析技术的集合,仍需在应用中完善创新,才能不断满足实际应用的需要。如 Huang 等^[47]将 GWR 模型扩展到时间维,这样 GWR 模型就具备了分析时空数据的能力。这个扩展功能同样适用于土壤和环境科学领域。Harris 等^[44]将 GWR 和克里格结合,提高了模型的预测能力。如本文的 1.6 节介绍的自回归 GWR,约束 GWR 和地理加权的 Logistic 和 probit 模型均为近年来出现的对基本 GWR 理论的发展。由于 GWR 理论的不完善,使得该模型的适用面越来越广,处理问题的能力越来越强,这也为其在土壤和环境科学上的应用提供了更好的前景。

另一方面,一门新的统计技术要想得到广泛应用,离不开有效的计算机程序的支持。目前实现 GWR 分析的软件有 GWR 3.0 和 ArcGIS 9.3 的 GWR 分析模块。这两个软件集成了 GWR 模型的一些基本功能,同时 GWR 3.0 处理的结果可以很方便地导入 ArcGIS 软件。这样,GWR 分析结果可以利用 ArcGIS 软件强大的 GIS 空间分析能力作进一步处理,从而也将 GWR 方法运用于土壤和环境科学提供了方便。由于 GWR 方法的特点与一些土壤和环境属性的特点相适应,随着 GWR 方法的发展,可以预见它在土壤和环境科学上会有广阔的应用前景。

参考文献:

- [1] Tobler WR. A computer movie simulating urban growth in the Detroit region[J]. *Economic Geography*, 1970, 46(2): 234-240
- [2] Peltó CR, Elkins TA, Boyd HA. Automatic contouring of irregularly spaced data[J]. *Geophysics*, 1968, 33(3): 424-430
- [3] Cleveland WS. Robust locally weighted regression and smoothing scatterplots[J]. *Journal of the American Statistical Association*, 1979, 74(368): 829-836
- [4] Ripley B D. *Spatial Statistics*[M]. New York: Wiley, 1981

- [5] Fedorov VV. Kriging and other estimators of spatial field characteristics[J]. *Atmospheric Environment*, 1989, 23(1): 175–184
- [6] Fotheringham AS, Brunson C, Charlton M. The geography of parameter space: an investigation of spatial non-stationarity[J]. *International Journal of Geographical Information Systems*, 1996, 10: 605–627
- [7] Brunson C, Fotheringham AS, Charlton M. Geographically weighted regression: A method for exploring spatial non-stationarity[J]. *Geographical Analysis*, 1996, 28(4): 281–298
- [8] Fotheringham AS, Brunson C, Charlton M. *Geographically Weighted Regression: the analysis of spatially varying relationships*[J]. Chichester: Wiley, 2002
- [9] Páez A, Uchida T, Miyamoto K. A general framework for estimation and inference of geographically weighted regression models: 1. location-specific kernel bandwidths and a test for locational heterogeneity[J]. *Environment and Planning A*, 2002, 34(4): 733–754
- [10] Wheeler DC, Páez A. Geographically weighted regression[A] // Fischer MM, Getis A. *Handbook of Applied Spatial Analysis*[M]. Heidelberg, Dordrecht, London and New York: Springer, 2010: 461–486
- [11] McMillen DP. One hundred fifty years of land values in Chicago: A nonparametric approach[J]. *Journal of Urban Economics*, 1996, 40(1): 100–124
- [12] Hurvich CM, Simonoff JS, Tsai CL. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion[J]. *Journal of Royal Statistical Society, Series B*, 1998, 60: 271–293
- [13] Congdon P. Modelling spatially varying impacts of socioeconomic predictors on mortality outcomes[J]. *Journal of Geographical Systems*, 2003, 5(2): 161–184
- [14] LeSage JP. A family of geographically weighted regression models[A] // Anselin L, Florax RJGM, Rey SJ. *Advances in spatial econometrics: Methodology, tools and applications*[M]. Berlin, Heidelberg and New York: Springer, 2004: 241–264
- [15] Wheeler DC, Calder C. An assessment of coefficient accuracy in linear regression models with spatially varying coefficients[J]. *Journal of Geographical Systems*, 2007, 9(2): 145–166
- [16] Leung Y, Mei CL, Zhang WX. Testing for spatial autocorrelation among the residuals of the geographically weighted regression[J]. *Environment and Planning A*, 2000, 32(5): 871–890
- [17] Wheeler DC, Tiefelsdorf M. Multicollinearity and correlation among local regression coefficients in geographically weighted regression[J]. *Journal of Geographical Systems*, 2005, 7(2): 161–187
- [18] Belsley DA. *Conditioning diagnostics: Collinearity and weak data in regression*[J]. New York, Chichester, Toronto and Brisbane: Wiley, 1991
- [19] Wheeler DC. Diagnostic tools and a remedial method for collinearity in geographically weighted regression[J]. *Environment and Planning A*, 2007, 39(10): 2 464–2 481
- [20] Brunson C, Fotheringham AS, Charlton M. Spatial non-stationarity and autoregressive models[J]. *Environment and Planning A*, 1998, 30(6): 957–973
- [21] Farber S, Páez AA. Systematic investigation of cross-validation in GWR model estimation: Empirical analysis and Monte Carlo simulations[J]. *Journal of Geographical Systems*, 2007, 9(4): 371–396
- [22] Páez A, Uchida T, Miyamoto KA. A general framework for estimation and inference of geographically weighted regression models: 2. spatial association and model specification tests[J]. *Environment and Planning A*, 2002, 34(5): 883–894
- [23] Pace RK, LeSage JP. Spatial autoregressive local estimation[A] // Getis A, Mur J, Zoller H. *Recent Advances in Spatial Econometrics*[M]. New York: Palgrave Macmillan, 2004: 31–51
- [24] Mur J, López F, Angulo A. Symptoms of instability in models of spatial dependence[J]. *Geogr Anal*, 2008, 40(2): 189–211
- [25] Wheeler DC. Simultaneous coefficient penalization and model selection in geographically weighted regression: The geographically weighted lasso[J]. *Environment and Planning A*, 2009, 41(3): 722–742
- [26] Frank IE, Friedman JH. A statistical view of some chemometrics regression tools[J]. *Technometrics*, 1993, 35: 109–148
- [27] Tibshirani R. Regression shrinkage and selection via the Lasso[J]. *Journal of the Royal Statistical Society, Series B*, 1996, 58: 267–288
- [28] Atkinson PM, German SE, Sear DA. Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression[J]. *Geographical Analysis*, 2003, 35(1): 58–82
- [29] Páez A. Exploring contextual variations in land use and transport analysis using a probit model with geographical weights[J]. *Journal of Transport Geography*, 2006, 14(3): 167–176
- [30] Foody GM. Geographical weighting as a further refinement to regression modelling: An example focused on the NDVI–rainfall relationship[J]. *Remote Sensing of Environment*, 2003, 88: 283–293
- [31] Tu J, Xia ZG. Examining spatially varying relationships between land use and water quality using geographically weighted regression I: Model design and evaluation[J]. *Science of the Total Environment*, 2008, 407: 358–378
- [32] Zhang CS, Tang Y, Luo L. Outlier identification and visualization for Pb concentrations in urban soils and its implications for identification of potential contaminated land[J]. *Environmental Pollution*, 2009, 157: 3 083–3 090
- [33] Li S, Zhao Z, Xie M. Investigating spatial non-stationary and scale-dependent relationships between urban surface temperature and environmental factors using geographically weighted regression[J]. *Environmental Modelling & Software*, 2010, 25: 1 789–1 800
- [34] Mishra U, Lal R, Liu D. Predicting the spatial variation of the soil organic carbon pool at a regional scale[J]. *Soil Science Society of America Journal*, 2010, 3: 906–914

- [35] Zhang CS, Tang Y, Xu XL. Towards spatial geochemical modelling: use of geographically weighted regression for mapping soil organic carbon contents in Ireland[J]. *Applied Geochemistry*, 2011, 26: 1 239–1 248
- [36] Wang K, Zhang C, Li W. Comparison of geographically weighted regression and regression kriging for estimating the spatial distribution of soil organic matter[J]. *GIScience & Remote Sensing*, 2012, 49(6): 915–932
- [37] Kumar S, Lal R, Liu D. A geographically weighted regression kriging approach for mapping soil organic carbon stock[J]. *Geoderma*, 2012, 189: 627–634
- [38] Mishra U, Riley WJ. Alaskan soil carbon stocks: Spatial variability and dependence on environmental factors[J]. *Biogeosciences*, 2012, 9: 3 637–3 645
- [39] 郭龙, 张海涛, 陈家赢, 李锐娟, 秦聪. 基于协同克里格插值和地理加权回归模型的土壤属性空间预测比较[J]. *土壤学报*, 2012, 49(5): 1 037–1 042
- [40] Solie J B, Raun WR, Stone ML. Submeter spatial variability of selected soil and bermudagrass production variables[J]. *Soil Science Society of America Journal*, 1999, 63: 1 724–1 733
- [41] Cambardella CA, Karlen DL. Spatial analysis of soil fertility parameters[J]. *Precision Agriculture*, 1999, 1: 5–14
- [42] Geypens M, Vanongeval L, Vogels N. Spatial variability of agricultural soil fertility parameters in a gleyic podzol of Belgium[J]. *Precision Agriculture*, 1999, 1: 319–326
- [43] Iqbal J, Thomasson JA, Jenkins JN. Spatial variability analysis of soil physical properties of alluvial soils[J]. *Soil Science Society of America Journal*, 2005, 69: 1 338–1 350
- [44] Harris P, Charlton M, Fotheringham AS. Moving window kriging with geographically weighted variograms[J]. *Stochastic Environmental Research and Risk Assessment*, 2010, 24: 1 193–1 209
- [45] 覃文忠, 王建梅, 刘妙龙. 地理加权回归分析空间数据的空间非平稳性[J]. *辽宁师范大学学报(自然科学版)*, 2005, 28(4): 476–479
- [46] 覃文忠, 王建梅, 刘妙龙. 混合地理加权回归模型算法研究[J]. *武汉大学学报·信息科学版*, 2007, 32(2): 115–119
- [47] Huang B, Wu B, Barry M. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices[J]. *International Journal of Geographical Information Science*, 2010, 24: 383–401

Geographically Weighted Regression and Its Application Prospect in Soil and Environmental Sciences

QU Ming-kai¹, LI Wei-dong^{2*}, ZHANG Chuan-rong², HUANG Biao¹

(1 *Key Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China*; 2 *Department of Geography, University of Connecticut, Storrs, Connecticut 06269, USA*)

Abstract: Geographically Weighted Regression (GWR) is a new spatially local regression technology, which emerged in last decade and is attracting more and more attention in recent years. The technology is used for exploring spatial non-stationarity through embedding spatial locations in linear regression models. Compared with the traditional ordinary least squares regression method which uses global parameters, the GWR method uses local correlation coefficients to incorporate spatial heterogeneity which is usually non-stationary, thus more effectively dealing with spatial data. This paper first introduced the theoretical origin, principle, existing deficiencies and further expansion of GWR. At the meantime, in order to understand the research and application status of GWR more accurately, a literature survey was conducted. Then, the application situation of GWR in the soil and environmental sciences were reviewed and a look into the future was made. After years of development and practice, GWR has been proved to be a mature outstanding approach and should have a broad prospect of application in evaluation of resources and environment.

Key words: Geographically weighted regression, Non-stationarity, Ordinary least squares, Environmental and soil sciences, Spatial data analysis