引文格式:赵阳阳,张小璐,张福浩,等.一种局部多项式时空地理加权回归方法[J].测绘学报,2018,47(5):663-671. DOI:10.11947/j.AGCS.2018.20170674.

ZHAO Yangyang, ZHANG Xiaolu, ZHANG Fuhao, et al. A Local Polynomial Geographically and Temporally Weight Regression[J]. Acta Geodaetica et Cartographica Sinica,2018,47(5):663-671. DOI:10.11947/j.AGCS.2018.20170674.

一种局部多项式时空地理加权回归方法

赵阳阳[1],张小璐[1],张福浩[1],仇阿根[1],杨 毅[2],石丽红[1],刘晓东[1]

1. 中国测绘科学研究院政府地理信息系统研究中心,北京 海淀 100830; 2. 淮海工学院测绘与海洋信息学院,江苏 连云港 222000

A Local Polynomial Geographically and Temporally Weight Regression

ZHAO Yangyang[1], ZHANG Xiaolu[1], ZHANG Fuhao[1], QIU Agen[1], YANG Yi[2], SHI Lihong[1], LIU Xiaodong[1]

1. Chinese Academy of Surveying and Mapping, Research Center of Government Geographic Information System, Beijing 100830, China; 2. School of Geomatics and Marine Information, Huaihai Institute of Technology, Lianyungang 222000, China

Abstract: Geographically and temporally weight regression (GTWR) estimates regression coefficients and fitted value by weighted least squares (WLS), which under the assumption of the same minimum random variance. As without considering the spatio-temporal heteroscedasticity, it may reduce the accuracy of estimation. Local polynomial estimation is a nonparametric estimation method to eliminate heteroscedasticity in statistics. On the basis of the local polynomial estimation, the local polynomial geographically and weight regression temporally (LPGTWR) approach is proposed in this paper. It reconstructs the spatio-temporal coefficients using three-dimensional Taylor Series in order to satisfy the Gauss-Markov assumption of independent identical distribution. Then estimate the regression coefficients and fitting value using weighted least squares. The experiments use both simulated data and real data to compare LPGTWR, GTWR and local linear-fitting-based geographically weight regression (LGWR). Experiments using simulated data showed that LPGTWR can significantly improve the accuracy of estimation not only in goodness-of-fit of the fitted value, but also in reducing bias of the coefficient estimation and the estimation. It is useful by adopting LPGTWR to eliminate heteroscedasticity effect and improve estimation accuracy.

Key words: geographically and temporally weighted regression; weighted least squares; local polynomial; Taylor series

Foundation support: The National Key Research and Development Program of China (No. 2016YFC0803101); The Basic Scientific Research of Chinese Academy of Surveying and Mapping(No. 7771812)

摘 要:基于加权最小二乘估计的时空地理加权回归方法,在随机项方差相同且最小的假设条件下估计回归参数和拟合值,由于没有考虑时空分析中异方差影响而导致估计结果存在一定偏差。局部多项式估计是一种消除异方差影响的非参数估计方法。本文在局部多项式估计原理基础上,提出了局部多项式时空地理加权回归方法。它是采用三元一阶泰勒级数展开式重构时空回归系数和自变量矩阵,进而建立满足高斯-马尔可夫独立同分布假定要求的新模型,利用新模型回归系数估计值、拟合值以及新模型与原模型的关系,可得到原模型回归系数估计值和拟合值。本文采用模拟数据和真实数据进行试验,以 GTWR 与局部线性地理加权回归作为对比方法,从方法适用性、整体估计效果、回归系数估计偏差和拟合优度、整体估计偏差等方面分析了 LPGTWR 方法性能,有效证明了 LPGTWR 方法能消除异方差影响提升估计精度。

关键词:时空地理加权回归;加权最小二乘估计;局部多项式;泰勒级数

中图分类号:P28 文献标识码:A 文章编号:1001-1595(2018)05-0663-09
基金项目:国家重点研发计划(2016YFC0803101);中国测绘科学研究院基本科研业务费(7771812)

在时空回归分析中,回归点与一定时空范围内的观测点有关,因此可以利用这些观测点计算回归点的拟合值[1]。时空地理加权回归是一个典型的时空回归分析方法,其步骤是先利用固定型带宽或调整型带宽准则[2]确定对回归点产生影响的观测点,再通过观测点与回归点之间的时空距离和权函数计算权重矩阵[3],最后采用加权最小二乘估计方法估算回归系数值和拟合值[4-5]。实践证明时空地理加权回归是探测时空非平稳特征的有效方法,应用广泛。文献[6]采用时空地理加权回归方法,在考虑了房价自身影响的情况下,研究了深圳市房价的时空非平稳变化情况;文献[7]利用时空地理加权回归方法建立了美国马里兰州巴尔的摩县的土地利用时空变化模型;文献[8]和[9]利用时空地理加权回归方法,研究了 PM2.5、PM10 的时空非平稳特征。

GTWR 的加权最小二乘估计方法是在随机项方差相同且最小的假设条件下估计回归参数和拟合值,而现实中自变量的随机项方差是不相同的,因此基于加权最小二乘的 GTWR 估计结果会产生偏差。随机项方差不同又称异方差,它普遍存在时空分析中,例如房价变化,一线城市城区和郊区的房价差异,比二三线城市城区和郊区的波动性大,在考虑时间因素后,时空因素的变化对房价的波动影响也不相同。又例如人口变化,经济发展好的城市对人口的吸引力远大于经济条件一般的城市,随着时间的变化,这种人口增长差距更大。对于上述现象,基于加权最小二乘估计的时空地理加权回归方法分析结果会出现偏差。局部多项式估计是一种良好的非参数估计方法,可以消除异方差的影响,减小回归系数估计值偏差,提升拟合精度。文献[10]将局部多项式回归与地理加权回归(geographically weight regression,GWR)方法相结合,提出了局部线性地理加权回归方法;文献[11]在局部线性地理加权回归方法的基础上研究了回归方法的稳健性,并证明在采用局部多项式改进后,能有效消除异方差,提升估计精度。由于地理加权回归方法只涉及空间非平稳性[12],因此 LGWR 只能解决二维空间的异方差,而无法消除时间维度的影响,因此不能直接应用到 GTWR 方法估计。

为了消除时空地理加权回归中同方差假设带来的拟合偏差,本文借鉴局部线性地理加权回归方法原理,在充分考虑时间和空间的双重影响基础上,提出了局部多项式时空地理加权回归方法。重点介绍了利用三元一阶泰勒级数重构满足高斯—马尔可夫独立同分布假定要求的新时空地理加权回归方程原理,推导了新方程回归系数、拟合值与原方程回归系数、拟合值之间的关系,并给出了局部多项式时空地理加权回归方法的算法流程。此外,本文以 LGWR 和 GTWR 为对比方法,通过模拟数据和真实数据试验验证了 LPGTWR 在消除异方差方面的有效性。

1 GTWR 原理

GTWR 假定回归系数是地理位置和观测时刻的任意函数[1,13],公式如下

$$y_i=\beta_0(u_i,v_i,t_i)+\sum_{k=1}^{d}\beta_k(u_i,v_i,t_i)x_{ik}+\varepsilon_i \quad i=1,2,\cdots,n \tag{1}$$

式中,$(x_{i1},x_{i2},,\cdots,x_{id};y_i)$表示观测点$(u_i,v_i,t_i)$,$(i=1,2,\cdots,n)$处的因变量 $\boldsymbol{y}$ 和自变量 x_1,$x_2,\cdots,x_d$ 的 n 组观测值;$\beta_k(u_i,v_i,t_i)$,$(k=0,1,\cdots,d)$是第 i 个数据点(u_i,v_i,t_i)处的未知回归系数;各系数是观测点(u_i,v_i,t_i)处的任意函数;$(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)$为独立同分布的误差项,通常假定均值为零,方差为 σ^2。

根据加权最小二乘方法,第 i 个观测点的回归系数估计值 $\hat{\boldsymbol{\beta}}_i$ 为

$$\hat{\boldsymbol{\beta}}_i=(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{y} \tag{2}$$

第 i 个观测点因变量的拟合值 $\hat{y}_i$ 为

$$\hat{y}_i=\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i=\boldsymbol{X}_i(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{y} \tag{3}$$

式中,$\boldsymbol{X}_i$ 表示自变量 $\boldsymbol{X}$ 矩阵中的第 i 行向量$\boldsymbol{X}_i=(1,x_{i1},x_{i2},\cdots,x_{ip})$;$\boldsymbol{W}_i$ 表示空间权重矩阵 $\boldsymbol{W}_i=\begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix}$;$w_{ij}$ 表示观测点 j 对观测点 i 影响的权重值。常用的计算方法有距离阈值法、距离反比法、高斯(Gauss)核函数和截尾型(Bi-square)核函数等方法[14,15]。

2　局部多项式时空地理加权回归方法

2.1　LPGTWR 原理

设定局部多项式时空地理加权回归模型的回归系数分别对横坐标 u、纵坐标 v 和时间 t 存在连续的二阶偏导数。根据泰勒级数公式，回归系数$\beta_j(u,v,t)$可以表示为在点(u_0,v_0,t_0)的某邻域范围内的泰勒级数展开，表达式如下

$$\beta_j(u,v,t)\approx\beta_j(u_0,v_0,t_0)+\beta_j^{(u)}(u_0,v_0,t_0)(u-u_0)+\beta_j^{(v)}(u_0,v_0,t_0)(v-v_0)+\beta_j^{(t)}(u_0,v_0,t_0)(t-t_0)\quad j=1,2,\cdots,p \tag{4}$$

式中，$\beta_j^{(u)}(u_0,v_0,t_0)$为回归系数$\beta_j(u_0,v_0,t_0)$在横坐标 u 方向的一阶偏导数；$\beta_j^{(v)}(u_0,v_0,t_0)$为回归系数 $\beta_j(u_0,v_0,t_0)$在纵坐标 v 方向的一阶偏导数；$\beta_j^{(t)}(u_0,v_0,t_0)$为回归系数 $\beta_j(u_0,v_0,t_0)$在时间 t 方向的一阶偏导数。

重新构建点(u_0,v_0,t_0)的自变量矩阵 $\boldsymbol{X}(u_0,v_0,t_0)$为 $n\times4p$ 阶矩阵

$$\boldsymbol{X}(u_0,v_0,t_0)=\begin{bmatrix} x_{11} & x_{11}(u_1-u_0) & x_{11}(v_1-v_0) & x_{11}(t_1-t_0) & \cdots & x_{1p} & x_{1p}(u_1-u_0) & x_{1p}(v_1-v_0) & x_{1p}(t_1-t_0)\\ x_{21} & x_{21}(u_2-u_0) & x_{21}(v_2-v_0) & x_{21}(t_2-t_0) & \cdots & x_{2p} & x_{2p}(u_2-u_0) & x_{2p}(v_2-v_0) & x_{2p}(t_2-t_0)\\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots\\ x_{n1} & x_{n1}(u_n-u_0) & x_{n1}(v_n-v_0) & x_{n1}(t_n-t_0) & \cdots & x_{np} & x_{np}(u_n-u_0) & x_{np}(v_n-v_0) & x_{np}(t_n-t_0)\end{bmatrix} \tag{5}$$

那么，回归系数列向量 $\boldsymbol{B}(u_0,v_0,t_0)$可以表示为

$$\boldsymbol{B}(u_0,v_0,t_0)=\begin{bmatrix}\beta_1(u_0,v_0,t_0) & \beta_1^{(u)}(u_0,v_0,t_0) & \beta_1^{(v)}(u_0,v_0,t_0) & \beta_1^{(t)}(u_0,v_0,t_0)\\ \cdots & \beta_p(u_0,v_0,t_0) & \beta_p^{(u)}(u_0,v_0,t_0) & \beta_p^{(v)}(u_0,v_0,t_0) & \beta_p^{(t)}(u_0,v_0,t_0)\end{bmatrix}^{\mathrm{T}} \tag{6}$$

第 j 个回归系数、一阶偏导数和回归系数列向量 $\boldsymbol{B}(u_0,v_0,t_0)$的关系可以表示为

$$\left.\begin{aligned}\beta_j(u_0,v_0,t_0)&=\boldsymbol{l}_{4j-3,4p}^{\mathrm{T}}\boldsymbol{B}(u_0,v_0,t_0)\\ \beta_j^{(u)}(u_0,v_0,t_0)&=\boldsymbol{l}_{4j-2,4p}^{\mathrm{T}}\boldsymbol{B}(u_0,v_0,t_0)\\ \beta_j^{(v)}(u_0,v_0,t_0)&=\boldsymbol{l}_{4j-1,4p}^{\mathrm{T}}\boldsymbol{B}(u_0,v_0,t_0)\\ \beta_j^{(t)}(u_0,v_0,t_0)&=\boldsymbol{l}_{4j,4p}^{\mathrm{T}}\boldsymbol{B}(u_0,v_0,t_0)\end{aligned}\right\} \tag{7}$$

式中，$\boldsymbol{l}_{4j-3,4p}^{\mathrm{T}}$为 $4p$ 行列向量，其中第$(4j-3)$个元素为 1，其他元素为 0；$\boldsymbol{l}_{4j-2,4p}^{\mathrm{T}}$为 $4p$ 行列向量，其中第$(4j-2)$个元素为 1，其他元素为 0；$\boldsymbol{l}_{4j-1,4p}^{\mathrm{T}}$为 $4p$ 阶列向量，其中第$(4j-1)$个元素为 1，其他元素为 0；$\boldsymbol{l}_{4j,4p}^{\mathrm{T}}$为 $4p$ 行列向量，其中第 $4j$ 个元素为 1，其他元素为 0。

因此，第 i 个点(u_i,v_i,t_i)自变量和因变量的关系可以表示为

$$y_i=\sum_{j=1}^{p}\beta_j(u_i,v_i,t_i)x_{ij}=\boldsymbol{x}_i\boldsymbol{\beta}(u_i,v_i,t_i)\quad i=1,2,\cdots,n \tag{8}$$

可得，

$$\boldsymbol{\beta}(u_i,v_i,t_i)=\begin{bmatrix}\beta_1(u_i,v_i,t_i)\\ \beta_2(u_i,v_i,t_i)\\ \vdots\\ \beta_p(u_i,v_i,t_i)\end{bmatrix}=\begin{bmatrix}\boldsymbol{l}_{1,4p}^{\mathrm{T}}\boldsymbol{B}(u_i,v_i,t_i)\\ \boldsymbol{l}_{5,4p}^{\mathrm{T}}\boldsymbol{B}(u_i,v_i,t_i)\\ \vdots\\ \boldsymbol{l}_{4p-3,4p}^{\mathrm{T}}\boldsymbol{B}(u_i,v_i,t_i)\end{bmatrix}=\boldsymbol{Q}\boldsymbol{B}(u_i,v_i,t_i) \tag{9}$$

$$\boldsymbol{Q}=\begin{bmatrix}\boldsymbol{l}_{1,4p}^{\mathrm{T}}\\ \boldsymbol{l}_{5,4p}^{\mathrm{T}}\\ \vdots\\ \boldsymbol{l}_{4p-3,4p}^{\mathrm{T}}\end{bmatrix} \tag{10}$$

局部多项式时空地理加权回归的拟合值可以表示如下

$$\hat{\boldsymbol{Y}}=\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \vdots\\ \hat{y}_n\end{bmatrix}=\begin{bmatrix}\boldsymbol{x}_1\boldsymbol{Q}\hat{\boldsymbol{B}}(u_1,v_1,t_1)\\ \boldsymbol{x}_2\boldsymbol{Q}\hat{\boldsymbol{B}}(u_2,v_2,t_2)\\ \vdots\\ \boldsymbol{x}_n\boldsymbol{Q}\hat{\boldsymbol{B}}(u_n,v_n,t_n)\end{bmatrix} \tag{11}$$

局部多项式时空地理加权回归模型基于泰勒级数展开重构自变量矩阵，利用重构的自变量矩阵和因变量建立了满足高斯-马尔可夫假定独立同分布要求的时空地理加权回归模型，新时空地理加权回归模型的拟合值可以利用加权最小二乘方法进行估计。因此，回归系数估计值 $\hat{\boldsymbol{B}}(u_0,v_0,t_0)$可以通过式(12)得到

$$\hat{\boldsymbol{B}}(u_0,v_0,t_0)=\begin{bmatrix}\hat{\beta}_1(u_0,v_0,t_0) & \hat{\beta}_1^{(u)}(u_0,v_0,t_0) & \hat{\beta}_1^{(v)}(u_0,v_0,t_0) & \hat{\beta}_1^{(t)}(u_0,v_0,t_0)\\ \cdots & \hat{\beta}_p(u_0,v_0,t_0), & \hat{\beta}_p^{(u)}(u_0,v_0,t_0) & \hat{\beta}_p^{(v)}(u_0,v_0,t_0) & \hat{\beta}_p^{(t)}(u_0,v_0,t_0)\end{bmatrix}^{\mathrm{T}}=[\boldsymbol{X}^{\mathrm{T}}(u_0,v_0,t_0)\boldsymbol{W}_0\boldsymbol{X}(u_0,v_0,t_0)]^{-1}\boldsymbol{X}^{\mathrm{T}}(u_0,v_0,t_0)\boldsymbol{W}_0\boldsymbol{Y} \tag{12}$$

那么,第 j 项回归系数估计值可表示为

$$\left.\begin{aligned}\hat{\beta}_j(u_0,v_0,t_0)&=\boldsymbol{l}_{4j-3,4p}^{\mathrm{T}}\hat{\boldsymbol{B}}(u_0,v_0,t_0)\\ \hat{\beta}_j^{(u)}(u_0,v_0,t_0)&=\boldsymbol{l}_{4j-2,4p}^{\mathrm{T}}\hat{\boldsymbol{B}}(u_0,v_0,t_0)\\ \hat{\beta}_j^{(v)}(u_0,v_0,t_0)&=\boldsymbol{l}_{4j-1,4p}^{\mathrm{T}}\hat{\boldsymbol{B}}(u_0,v_0,t_0)\\ \hat{\beta}_j^{(t)}(u_0,v_0,t_0)&=\boldsymbol{l}_{4j,4p}^{\mathrm{T}}\hat{\boldsymbol{B}}(u_0,v_0,t_0)\end{aligned}\right\}\tag{13}$$

因此,回归系数估计值表达式可表示为

$$\hat{\boldsymbol{\beta}}(u_i,v_i,t_i)=\begin{bmatrix}\hat{\beta}_1(u_i,v_i,t_i)\\ \hat{\beta}_2(u_i,v_i,t_i)\\ \vdots\\ \hat{\beta}_p(u_i,v_i,t_i)\end{bmatrix}=\begin{bmatrix}\boldsymbol{l}_{1,4p}^{\mathrm{T}}\hat{\boldsymbol{B}}(u_i,v_i,t_i)\\ \boldsymbol{l}_{5,4p}^{\mathrm{T}}\hat{\boldsymbol{B}}(u_i,v_i,t_i)\\ \vdots\\ \boldsymbol{l}_{4p-3,4p}^{\mathrm{T}}\hat{\boldsymbol{B}}(u_i,v_i,t_i)\end{bmatrix}=\boldsymbol{Q}\hat{\boldsymbol{B}}(u_i,v_i,t_i)\tag{14}$$

局部多项式时空地理加权回归方法因变量拟合值可以表达为

$$\hat{\boldsymbol{Y}}=\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \vdots\\ \hat{y}_n\end{bmatrix}=\begin{bmatrix}\boldsymbol{x}_1\boldsymbol{Q}\hat{\boldsymbol{B}}(u_1,v_1,t_1)\\ \boldsymbol{x}_2\boldsymbol{Q}\hat{\boldsymbol{B}}(u_2,v_2,t_2)\\ \vdots\\ \boldsymbol{x}_n\boldsymbol{Q}\hat{\boldsymbol{B}}(u_n,v_n,t_n)\end{bmatrix}=\boldsymbol{L}\boldsymbol{Y}\tag{15}$$

其中,$\boldsymbol{L}$ 为帽子矩阵,矩阵形式如下

$$\boldsymbol{L}=\begin{bmatrix}\boldsymbol{x}_1\boldsymbol{Q}\left[\boldsymbol{X}^{\mathrm{T}}(u_1,v_1,t_1)\boldsymbol{W}_1\boldsymbol{X}(u_1,v_1,t_1)\right]^{-1}\boldsymbol{X}^{\mathrm{T}}(u_1,v_1,t_1)\boldsymbol{W}_1\\ \boldsymbol{x}_2\boldsymbol{Q}\left[\boldsymbol{X}^{\mathrm{T}}(u_2,v_2,t_2)\boldsymbol{W}_2\boldsymbol{X}(u_2,v_2,t_2)\right]^{-1}\boldsymbol{X}^{\mathrm{T}}(u_2,v_2,t_2)\boldsymbol{W}_2\\ \vdots\\ \boldsymbol{x}_n\boldsymbol{Q}\left[\boldsymbol{X}^{\mathrm{T}}(u_n,v_n,t_n)\boldsymbol{W}_n\boldsymbol{X}(u_n,v_n,t_n)\right]^{-1}\boldsymbol{X}^{\mathrm{T}}(u_n,v_n,t_n)\boldsymbol{W}_n\end{bmatrix}\tag{16}$$

2.2 算法流程

局部多项式时空地理加权回归方法的核心是将回归系数表示为其时空邻域范围内的三元一阶泰勒级数展开,通过剥离随机项方差影响,重新构建满足高斯—马尔可夫假定要求的新时空地理加权回归模型。根据泰勒级数展示式,能推导出原模型回归系数、拟合值与新模型回归系数、拟合值之间的关系,通过加权最小二乘解算出新模型回归系数和拟合值,即可估计原模型的回归系数和拟合值。为了更加清晰地阐述局部多项式时空地理加权回归方法的估计步骤,给出了算法流程如图 1 所示。

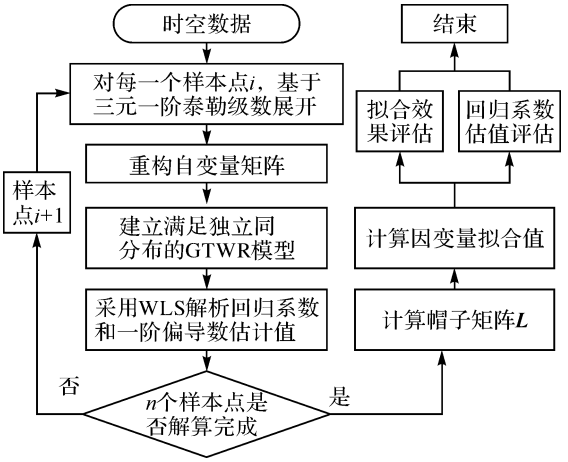


图 1 LPGTWR 流程图

Fig.1 Flow chart of LPGTWR

步骤 1:对每个样本点(u_i,v_i,t_i)第 j 个回归系数$\beta_j(u_i,v_i,t_i)$,$j=1,2,\cdots,p$ 进行泰勒级数展开,表示为该点较小邻域范围内任意点(u_0,v_0,t_0)的空间横坐标 u_0、纵坐标 v_0、时间 t_0 一阶偏导 $\beta_j^{(u)}(u_0,v_0,t_0)$、$\beta_j^{(v)}(u_0,v_0,t_0)$及 $\beta_j^{(t)}(u_0,v_0,t_0)$的线性组合。

步骤 2:基于泰勒级数展开结果,重新设计自变量矩阵 $\boldsymbol{X}$。

步骤 3:基于加权最小二乘估计方法,计算回归系数估计值,包括回归系数估计值 $\hat{\beta}_j(u_0,v_0,t_0)$和空间横坐标一阶偏导 $\hat{\beta}_j^{(u)}(u_0,v_0,t_0)$、纵坐标一阶偏导 $\hat{\beta}_j^{(v)}(u_0,v_0,t_0)$、时间方向一阶偏导 $\hat{\beta}_j^{(t)}(u_0,v_0,t_0)$。

步骤 4:循环完所有样本点,计算帽子矩阵 $\boldsymbol{L}$。

步骤 5:求解因变量的拟合值 $\hat{\boldsymbol{y}}$。

步骤 6:计算拟合值评价指标和回归系数估计值评价指标,分析方法拟合精度。

算法结束。

3 试验及结果分析

为了测试 LPGTWR 方法的性能,本文采用模拟数据和真实数据,以 LGWR、GTWR 作为对比方法,进行试验分析。模拟数据的回归系数和因变量的真值是已知,可以分析估计值与真实值的偏差,真实数据回归系数的真值是未知的,可以从模型估计的角度,评价 LPGTWR 模型的整体性能,从而多角度分析 LPGTWR 的性能,为方法应用提供参考和依据。

3.1 模拟数据试验

3.1.1 试验设置

本文以 u、v 为平面坐标轴、以 t 为时间轴建

立一个三维立体空间。设空间左下角为原点，立体空间每个坐标轴长度均为 12 单位长度，令 u、v、t 的取值分别为 $0,1,2,\cdots,m-1$，观测点均匀地分布在 $m\times m\times m$ 的格点上，则空间内共有 $n=m^3$ 个观测点，观测点的坐标取值可以按照以式(17)计算

$$\begin{aligned}(u_i,v_i,t_i)=&(\mathrm{mod}(i-1,m),\mathrm{mod}(\mathrm{int}(i-1)/\\&m,m),\mathrm{int}((i-1)/m^2))\\&i=1,2,\cdots,m^3\end{aligned}\tag{17}$$

式中，$\mathrm{mod}(a,b)$ 表示 a 除以 b 后的余数；$\mathrm{int}(a/b)$ 表示 a 除以 b 后取整。

本文设计 3 组模型数据，其中自变量 x_{1i}、x_{2i} 是分布在(−4,4)之间的随机数；残差 ε_i 服从标准正态分布；回归系数 β_0、β_1、β_2 与 u、v、t 相关，3 组公式如下所示。

一组：

$$\left.\begin{aligned}&\beta_0=(u+v+t)/6\\&\beta_1=2t\\&\beta_2=1/324[36-(6-u)^2][36-(6-v)^2]\end{aligned}\right\}$$

二组：

$$\left.\begin{aligned}&\beta_0=(u+v)/6\\&\beta_1=u/6\\&\beta_2=(u+v+t)/12\end{aligned}\right\}$$

三组：

$$\left.\begin{aligned}&\beta_0=(u+v)/6\\&\beta_1=2t\\&\beta_2=(u+v+t)/12\end{aligned}\right\}$$

3.1.2　结果分析

为了消除数据生成时产生的随机误差，每组数据生成 10 套，每个方法重复 10 次。本文采用 Akaike 信息法则(Akaike information criterion，AIC)[16]、平均均方误差(mean square error，MSE)和回归系数估计偏差作为评价指标，分析方法的适用性、整体估计效果和回归系数估计偏差。其中，回归系数估计偏差是指是回归系数的真实值和估计值之间的偏差统计量[10]。试验采用交叉验证法(cross validation，CV)确定最优带宽和时空参数[17]，表 1 记录了 10 组模拟数据在 LGWR、GTWR 和 LPGTWR 方法估计下的 AIC 平均值。一般地，当两个模型的 AIC 值相差3 时，说明模型之间存在明显差异，且 AIC 最小值对应的模型是最优模型[18]。

由表 1 可知，3 组模拟数据下，LPGTWR 方法都取得了最小的 AIC 值，说明 LPGTWR 比 LGWR 和 GTWR 的适用性更好。对于同一个方法，二组模拟数据结果较好，一组数据结果最差，说明数据复杂程度对方法有影响，数据越简单，模拟效果越好，数据越复杂，模拟效越果差。从 LPGTWR 性能提升情况看，三组数据 AIC 值相差均大于 3，且 LPGTWR/LGWR 性能提升幅度比 LPGTWR/GTWR 幅度大，说明除了异方差外，时间也是影响估计精度的重要因素。

表 1　LGWR、GTWR 和 LPGTWR 方法平均 AIC 统计值

Tab.1　The mean AIC value of the LGWR, GTWR and LPGTWR models

模拟数据	LGWR	GTWR	LPGTWR	LPGTWR 与 LGWR 对比	LPGTWR 与 GTWR 对比
一组	5 708.245	2 981.257	2 582.006	3 126.239	399.251
二组	2 264.925	2 239.549	2 231.584	33.341	7.965
三组	5 701.105	2 478.47	2 218.583	3 482.522	259.887

表 2 给出了 10 套模拟数据在 LGWR、GTWR 和 LPGTWR 下的平均 MSE 值。平均 MSE 值越小，说明估计值越接近真实值，方法的整体估计效果越好。由表 2 可知：首先，LPGTWR 方法在三组数据中取得了最小平均 MSE，说明 LPGTWR 方法整体估计效果优于 GTWR、GTWR。其次，二组模拟数据在 3 种方法下估计效果比较稳定，而一组和三组模拟数据，LGWR 方法误差很大，说明数据复杂性和时间因素，给 LGWR 方法带来了干扰，而相对于 LGWR，GTWR 和 LPGTWR 方法考虑了时间因素，能给出较稳定的估计结果。最后，LPGTWR 比 LGWR 和 GTWR 平均 MSE 性能提升比率大于 14%，说明 LPGTWR 方法在处理异方差后，明显提升了整体估计精度。

图 2 绘制了 3 种方法的回归系数估计偏差分布图。其中，图 2(a)、图 2(b)表示局部时空变回归系数估计偏差；图 2(c)、图 2(d)表示局部时间变回归系数估计偏差；图 2(e)、图 2(f)表示局部空间变回归系数估计偏差。图中纵轴表示 LPGTWR 的回归系数估计偏差，横轴表示 LGWR 或 GTWR 的回归系数估计偏差，散点表示 10 次试验结果。当散点靠近横标时，说明横轴的方法估计偏差大于纵轴方法的估计偏差，即纵轴方法优于横轴方法。反之，说明横轴方法优于纵轴方法。分析图 2 可知，对于局部时空变回归

系数和局部时间变回归系数，LPGTWR 优于 GTWR 和 LGWR 方法，GTWR 方法优于 LGWR。对于局部空间变回归系数，LGWR 方法优于 LPGTWR 和 GTWR 方法。这说明在时空回归分析中，时间对模型精度的影响比重大于异方差，在考虑时间因素的前提下，局部多项式时空地理加权回归方法能提升估计精度。

表 2 LGWR、GTWR 和 LPGTWR 方法的 MSE 统计值

Tab. 2 The MSE value of the LGWR, GTWR and LPGTWR models

模拟数据	LGWR	GTWR	LPGTWR	LPGTWR 与 LGWR 对比/(%)	LPGTWR 与 GTWR 对比/(%)
一组	140.498 6	1.558 552	1.224 298	99.13	21.45
二组	1.2565 19	1.0677 17	0.898 352	28.50	15.86
三组	139.88 23	1.027 06	0.882 689	99.39	14.06

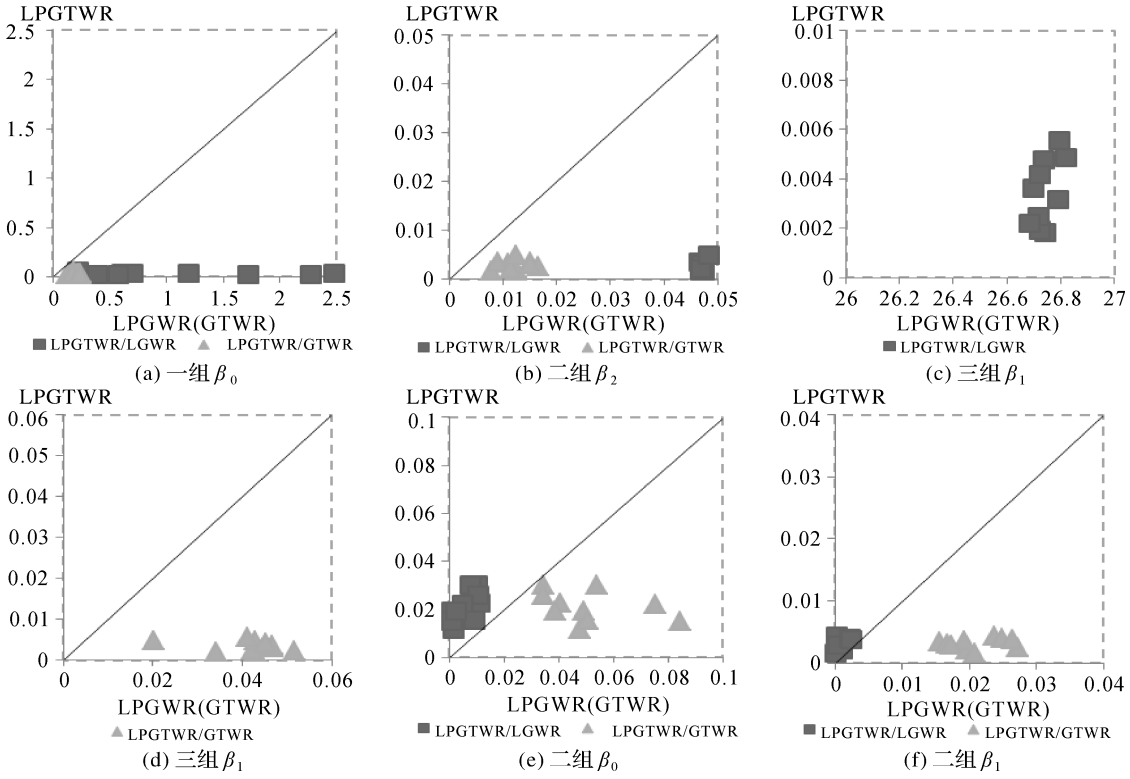


图 2 回归系数估计偏差分布图

Fig.2 The coefficient estimation bias of simulated data

3.2 真实数据试验及结果分析

本文以长江中下游地区人口密度分布与影响因素关系作为真实数据进行试验。研究发现人口密度分布与自然条件、土地利用以及社会经济等因素相关[19-21]。本文收集了 2000 年、2005 年、2010 年、2015 年长江中下游各地市人均 GDP(元/人)、年均降水量(mm)、年均气温(℃)、耕地面积(km^2)、林地面积(km^2)、城乡工矿居民用地面积(km^2)和平原面积(km^2)等 7 个特征变量。其中，人均 GDP 指区域 GDP 总产值除以区域常住人口数。时空因素包括空间位置坐标和时间，空间上采用 WGS84 坐标系，高斯克里格投影，时间设置 2000 年为基准年用 1 表示，每增加 1 年增加一个单位。长江中下游地区行政区划矢量数据来源于地图出版社；各地市的人口数和 GDP 来源于各省统计年鉴；土地利用数据来源于中国科学院资源环境科学数据中心。

为了建立可靠的人口密度分布模型，需要选择相关性强的变量。在普通线性回归分析中，一般采用逐步回归方法建立最优模型[22]。在地理加权回归分析中，文献[18]基于最小 AIC 准则，穷举了所有变量组合进行建模，以最小 AIC 值模型为最优模型。本文采用逐步回归的迭代方式，以 AIC 最小作为评价准则选取变量进行建模。经过多重共线性分析和变量选取，确定人均 GDP、年均气温和城乡工矿居民用地面积为自变量，人口密度为因变量，分别建立 LGWR、GTWR 和 LPGTWR 模型。文本采用拟合优度(R^2)、调

整型拟合优度(R^2adj)、MSE、误差项平方和(sum of squares for error，SSE)作为评价指标[23]，结果如表 3 所示。

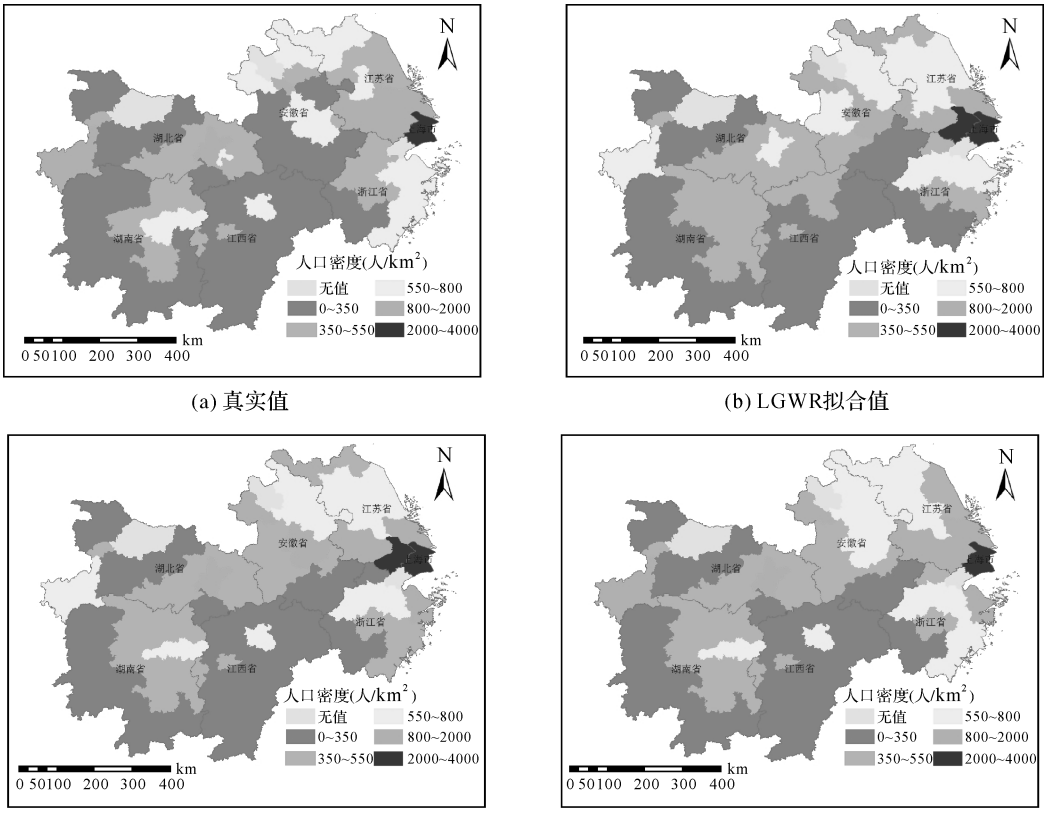
表 3　真实数据试验结果

Tab.3　The test results of real data

方法	R^2	R^2adj	MSE	SSE
LGWR	0.711 4	0.707 2	0.036 3	7.556 8
GTWR	0.815 2	0.811 6	0.023 4	4.839 6
LPGTWR	0.882 8	0.880 5	0.014 8	3.068 6
LPGTWR/LGWR improvement/(%)	24.09	24.51	59.23	59.39
LPGTWR/GTWR improvement/(%)	8.29	8.49	36.75	36.59

结果显示，3 种方法 R^2 均大于 0.71，说明 3 种方法都能建立可靠的模型来估算人口密度。从整体建模情况看，LPGTWR 方法的拟合优度达 0.882 8，比 LGWR 方法提升了 24.09%，比 GTWR 方法提升了 8.29%，调整拟合优度为 0.880 5，比 LGWR 方法提升了 24.51%，比 GTWR 方法提升了 8.49%，说明 LPGTWR 建立的模型最优。从估计偏差情况看，LPGTWR 方法的均方差为 0.014 8，比 LGWR 方法提升了59.23%，比 GTWR 方法提升了 36.75%，LPGTWR 方法的误差项平方和为 3.068 6，比 LGWR 方法提升了 59.39%，比 GTWR 方法提升了 36.59%，上述指标均说明 LPGTWR 方法的整体估计偏差小，结果优于 GTWR、LGWR 方法。综上所述，真实数据试验说明 LPGTWR 方法相比 GTWR 和 LGWR 方法，能建立更优的模型，减小整体估计偏差，得到更可靠的结果。

图 3 绘制了 2015 年长江中下游地区人口密度的真值、LGWR 拟合值、GTWR 拟合值和 LPGTWR 拟合值。观察图可知，在上海、湖北西部地区，LPGTWR 方法的预测结果相比 LGWR 和 GTWR 方法更接近真实情况。但 3 种方法在安徽西部、北部地区预测误差较大。通过进一步研究发现，安徽北部蚌埠市、宿州市和亳州市 2015 年人口相对 2000 年分别增长了 14.45%、17.74%和 25.01%，而 15 年来上海市增长了 50.09%。相对上海市该地区人口的时空变化波动程度并不明显，而且该地区的时空非平稳特征也不显著，即异方差和时空非平稳性对该地区人口密度变化影响程度并不显著，因此预测结果存在偏差。



(a) 真实值　(b) LGWR拟合值

(c) GTWR拟合值　(d) LPGTWR拟合值

图 3　2015 年长江中下游地区人口密度分布图

Fig.3　The population distribution of middle and lower reaches of the Yangtze River

4 结 语

本文提出了一种局部多项式时空地理加权回归方法,它利用三元一阶泰勒级数展开式和加权最小二乘估计的原理,将回归系数定义为真值和在空间坐标轴和时间方向的偏差,通过剥离空间和时间维度的偏差来消除随机项方差,进而解决时空地理加权回归无法处理异方差的问题。本文介绍了局部多项式时空地理加权回归方法的原理,给出了算法流程,并通过模拟数据和真实数据对 LPGTWR 进行了测试。模拟数据试验从适用性、整体估计效果和回归系数估计偏差 3 个角度验证了 LPGTWR 方法优于 GTWR 和 LGWR,特别是在时空非平稳情况下,LPGTWR 能显著提升拟合精度。真实数据试验验证了 LPGTWR 方法的有效性,且其拟合结果在时空维度最接近真实情况。局部多项式时空地理加权回归方法是利用数学方法提升时空回归拟合精度的时空回归方法,本文在人口密度及影响因素分析中进行了验证,未来希望能在更多领域推广应用。

参考文献:

[1] HUANG Bo, WU Bo, BARRY M. Geographically and Temporally Weighted Regression for Modeling Spatio-temporal Variation in House Prices[J]. International Journal of Geographical Information Science, 2010, 24(3): 383-401.

[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.

[3] 覃文忠. 地理加权回归基本理论与应用研究[D]. 上海: 同济大学, 2007.
QIN Wenzhong. The Basic Theoretics and Application Research on Geographically Weighted Regression[D]. Shanghai: Tongji University, 2007.

[4] BRUNSDON C, AITKIN M, FOTHERINGHAM A S, et al. A Comparison of Random Coefficient Modelling and Geographically Weighted Regression for Spatially Non-stationary Regression Problems[J]. Geographical and Environmental Modelling, 1999, 3(1): 47-62.

[5] FOTHERINGHAM A S, BRUNSDON C, CHARLTON M. Geographically Weighted Regression: the Analysis of Spatially Varying Relationships[M]. New York: John Wiley & Sons, 2002: 34-45.

[6] WU Bo, LI Rongrong, HUANG Bo. A Geographically and Temporally Weighted Autoregressive Model with Application to Housing Prices[J]. International Journal of Geographical Information Science, 2014, 28(5): 1186-1204.

[7] WRENN D H, SAM A G. Geographically and Temporally Weighted Likelihood Regression: Exploring the Spatio-temporal Determinants of Land Use Change[J]. Regional Science and Urban Economics, 2014, 44: 60-74.

[8] BAI Yang, WU Lixin, QIN Kai, et al. A Geographically and Temporally Weighted Regression Model for Ground-Level $PM_{2.5}$ Estimation from Satellite-derived 500 m Resolution AOD[J]. Remote Sensing, 2016, 8(3): 262.

[9] CHU H J, HUANG Bo, LIN C Y. Modeling the Spatio-temporal Heterogeneity in the PM_{10}-$PM_{2.5}$ Relationship[J]. Atmospheric Environment, 2015, 102: 176-182.

[10] WANG Ning, MEI Changlin, YAN Xiaodong. Local Linear Estimation of Spatially Varying Coefficient Models: An Improvement on the Geographically Weighted Regression Technique[J]. Environment and Planning A: Economy and Space, 2008, 40(4): 986-1005.

[11] ZHANG Huiguo, MEI Changlin. Local Least Absolute Deviation Estimation of Spatially Varying Coefficient Models: Robust Geographically Weighted Regression Approaches[J]. International Journal of Geographical Information Science, 2011, 25(9): 1467-1489.

[12] BRUNSDON C, FOTHERINGHAM A S, CHARLTON M. Some Notes on Parametric Significance Tests for Geographically Weighted Regression[J]. Journal of Regional Science, 1999, 39(3): 497-524.

[13] 刘美玲. 时空地理加权回归模型的统计诊断[D]. 西安: 西安建筑科技大学, 2013.
LIU Meiling. Statistical Diagnostics in Geographically and Temporally Weighted Regression Models[D]. Xi'an: Xi'an University of Architecture and Technology, 2013.

[14] 黄砚玲. 地理加权空间经济计量模型的 GMM 估计及区域金融发展收敛性实证研究[D]. 广州: 华南理工大学, 2012.
HUANG Yanling. GMM Estimate for SGWR Model and a Spatial Analysis of Regional Financial Convergence in China[D]. Guangzhou: South China University of Technology, 2012.

[15] 杨毅. 顾及时空非平稳性的地理加权回归方法研究[D]. 武汉: 武汉大学, 2016.
YANG Yi. Research on Geographically and Temporally Weighted Regression for Spatial and Temporal Nonstationarity[D]. Wuhan: Wuhan University, 2016.

[16] HURVICH C M, SIMONOFF J S, TSAI C L. Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion[J]. Journal of The Royal Statistical Society: Series B, 1998, 60(2): 271-293.

[17] CLEVELAND W S. Robust Locally Weighted Regression and Smoothing Scatterplots[J]. Journal of the American Statistical Association, 1979, 74(368): 829-836.

[18] LU Binbin, CHARLTON M, HARRIS P, et al. Geographi-

cally Weighted Regression with a Non-euclidean Distance Metric: A Case Study Using Hedonic House Price Data [J]. International Journal of Geographical Information Science, 2014, 28(4): 660-681.

[19] 柏中强，王卷乐，杨雅萍，等. 基于乡镇尺度的中国 25 省区人口分布特征及影响因素[J]. 地理学报，2015，70(8)：1229-1242.
BAI Zhongqiang, WANG Juanle, YANG Yaping, et al. Characterizing Spatial Patterns of Population Distribution at Township Level across the 25 Provinces in China[J]. Acta Geographica Sinica, 2015, 70(8): 1229-1242.

[20] 戚伟，李颖，刘盛和，等. 城市昼夜人口空间分布的估算及其特征——以北京市海淀区为例[J]. 地理学报，2013，68(10)：1344-1356.
QI Wei, LI Ying, LIU Shenghe, et al. Estimation of Urban Population at Daytime and Nighttime and Analyses of Their Spatial Pattern: A Case Study of Haidian District, Beijing[J]. Acta Geographica Sinica, 2013, 68(10): 1344-1356.

[21] 鲁楠，张委伟，陈利军，等. 顾及城乡差异的大区域人口密度估算——以山东省为例[J]. 测绘学报，2015，44(12)：1384-1391. DOI: 10.11947/j.AGCS.2015.20150005.
LU Nan, ZHANG Weiwei, CHEN Lijun, et al. Estimation of Large Regional Urban and Rural Population Density Based on the Differences of Population Distribution between Urban and Rural: Take Shandong Province as Example [J]. Acta Geodaetica et Cartographica Sinica, 2015, 44(12): 1384-1391. DOI: 10.11947/j.AGCS.2015.20150005.

[22] 陈彦光. 基于 Matlab 的地理数据分析[M]. 北京：高等教育出版社，2012：34-45.
CHEN Yanguang. Geographical Data Analysis with Matlab [M]. Beijing: Higher Education Press, 2012: 34-45.

[23] SONG Weize, JIA Haifeng, HUANG Jingfeng, et al. A Satellite-based Geographically Weighted Regression Model for Regional $PM_{2.5}$ Estimation over the Pearl River Delta Region in China[J]. Remote Sensing of Environment, 2014, 154: 1-7.

（责任编辑：陈品馨）

收稿日期：2017-11-28
修回日期：2018-03-12

第一作者简介：赵阳阳(1987—)，女，博士，助理研究员，研究方向为时空数据分析与挖掘。
First author: ZHAO Yangyang(1987—), female, PhD, assistant research fellow, majors in spatio-temporal data analysis and mining.
E-mail: nhyyangyang@126.com
通信作者：张小璐
Corresponding author: ZHANG Xiaolu
E-mail: 397203228@qq.com