基于变量选择的混合时空地理加权回归参数估计

侯 健[1,2], 田茂再[1,2,3,4,5,*]

(1. 新疆财经大学 新疆社会经济统计研究中心, 新疆 乌鲁木齐 830012)
(2. 新疆财经大学 统计与数据科学学院, 新疆 乌鲁木齐 830012)
(3. 中国人民大学 应用统计科学研究中心, 北京 100872)
(4. 中国人民大学 统计学院, 北京 100872)
(5. 兰州财经大学 统计学院, 甘肃 兰州 730101)

摘 要: 混合时空地理加权回归模型作为一种有效处理空间数据全局平稳和局部非平稳的分析方法得到了广泛的应用. 但其参数估计方法中假定固定系数变量已知且不存在时空效应, 这一较强的前提使回归系数的估计值变得极不稳定. 为探究当固定系数变量存在时空效应时的参数估计方法, 本文提出一种变量选择 (Variable Selection) 方法来剔除指标间的交互效应, 并给出相应的算法过程. 通过乌鲁木齐市商品住宅真实价格数据对不同估计方法进行对比验证, 结果表明, 利用变量选择方法后得到的 MGTWR 模型性能和拟合效果得到提升, 固定回归系数的估计更加稳定, 原有参数估计方法得到改善.

关键词: 混合时空地理加权回归模型; 变量选择; 两步估计

1 引言

一般线性回归模型 (Ordinary Linear Regression, OLR) 在处理变量间关系上发挥着重要的作用, 但随着空间分析技术的不断发展, 数据的空间属性越来越受到关注, OLR 模型已经无法适应数据在空间上的变化情况, 从而很难挖掘出变量之间的真实变化规律. 为了解决带有空间位置属性数据的拟合问题, Fotheringham 等 (1996) 提出了一种基于局部最优思想的空间变系数回归模型: 地理加权回归模型 (Geographically weighted regression, GWR), 通过距离衰减函数来衡量数据间的位置变化情况, 从而对每一样本赋予不同的权重, 使得在每一样本点上都存在着回归方程 [1]. Brunsdon C 等 (1999) 为同时考虑数据的局部和全局变化情况, 在 GWR 模型的基础上提出了混合地理加权回归模型 (Mixed Geographically weighted regression, MGWR) 将 OLR 模型与 GWR 模型组合起来, 使其可以同时反映数据的全局平稳性和局部非平稳性 [2]. Huang B(2010) 则在 GWR 模型的基础上增加了时间维度, 提出了时空地理加权回归模型 (Geographically and Temporally Weighted Regression, GTWR) 并首次将其应用于房地产方面的研究, 相较于 OLR 模型进一步为时空数据的分析提供了基础 [3,4]. 赵阳阳等 (2017) 针对全局时空平稳特征和局部时空非平稳特征同时存在的现象, 提出了混合时

收稿日期: 2020-03-28
资助项目: 中国人民大学科研基金 (中央高校基本科研业务专项资金资助) 项目成果 (18XNL012)
* 通信作者

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

空地理加权回归模型 (Mixed Geographically and Temporally Weighted Regression, MGTWR) 并给出 MGTWR 模型的数学定义和回归参数的两步估计法 [5].

尽管 MGTWR 模型的参数估计与统计诊断体系已经较为成熟, 但其两步估计法对变量的前提假设仍有较强的要求, 即认为固定系数指标不含有任何时空因素且观测值已知. 这就使得两步估计法的迭代过程中不得不利用变系数指标观测值的信息, 从而导致固定回归系数的估计值中包含了时空因素, 减弱了对数据全局时空平稳性的解释能力. 为此, 本文提出一种基于变量选择 (Variable Selection) 的回归系数估计方法, 通过检验指标的时空平稳性, 剔除固定系数指标存在的时空效应, 避免了在迭代过程中时空效应对固定系数估计的干扰.

2 理论方法

2.1 模型定义

为了解决全局平稳特征和局部时空非平稳特征同时存在的问题, MGTWR 模型将指标变量分成固定系数指标和变系数指标两个部分, 前者用于解释数据的全局稳定性, 后者用于解释局部时空非平稳性. 其基本形式可以表示为 OLR 模型与 GTWR 模型的线性组合:

$$\begin{aligned} y_j &= \sum_{k_0=1}^{p^{(f)}} \theta_{k_0}^{(f)} x_{jk_0}^{f} + \sum_{k_1=1}^{p^{(r)}} \theta_{jk_1}^{(r)}(u_j, v_j, t_j) x_{jk_1}^{(r)} + \varepsilon_j, j = 1, 2, \cdots, n; \\ & p^{(f)} + p^{(r)} = p; \varepsilon_j \sim N(0, \sigma^2) \end{aligned} \tag{1}$$

其中 y_j 是第 j 个响应变量. 我们称 $\theta_{k_0}^{(f)}$ 是第 k_0 个固定系数 (fixed coefficient), 将 $\theta_{jk_1}^{(r)}(u_j, v_j, t_j)$ 称为第 k_1 个变系数 (variable coefficient) 且是关于时空坐标 (u_j, v_j, t_j) 的连续函数. p 表示由 $p^{(f)}$ 个固定指标和 $p^{(r)}$ 个变系数指标组成的指标总数. $x_{jk_0}^{(f)}$、$x_{jk_1}^{(r)}$ 分别是第 j 个样本的第 k_0、k_1 个固定指标和变系数样本观测值. n 是样本总量.ε_j 是第 j 个服从 $N(0, \sigma^2)$ 的随机误差项, 且 σ^2 未知可用样本方差 $\hat{\sigma}^2$ 代替. 若将式 (1) 以矩阵形式表达, 则有:

$$Y = A^{(f)}\boldsymbol{\theta}^{(f)} + A^{(r)}\boldsymbol{\theta}^{(r)} + \boldsymbol{\varepsilon} \tag{2}$$

其中:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{A}^{(f)} = \begin{bmatrix} 1 & x_{11}^{(f)} & \cdots & x_{1p^{(f)}}^{(f)} \\ 1 & x_{21}^{(f)} & \cdots & x_{2p^{(f)}}^{(f)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1}^{(f)} & \cdots & x_{np^{(f)}}^{(f)} \end{bmatrix}, \boldsymbol{\theta}^{(f)} = \begin{bmatrix} \theta_0^{(f)} \\ \theta_1^{(f)} \\ \vdots \\ \theta_{p^{(f)}}^{(f)} \end{bmatrix},$$

$$\boldsymbol{A}^{(r)} = \begin{bmatrix} x_{11}^{(r)} & x_{12}^{(r)} & \cdots & x_{1p^{(r)}}^{(r)} \\ x_{21}^{(r)} & x_{21}^{(r)} & \cdots & x_{2p^{(r)}}^{(r)} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1}^{(r)} & x_{n2}^{(r)} & \cdots & x_{np^{(r)}}^{(r)} \end{bmatrix}, \boldsymbol{\theta}^{(r)} = \begin{bmatrix} \theta_0^{(r)} \\ \theta_1^{(r)} \\ \vdots \\ \theta_{p^{(r)}}^{(r)} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

显然, 当式 (1) 中 $t_j = 0, j = 1, 2, \cdots, n$ 时, MGTWR 模型就不含有时间维度并退化为 MGWR 模型; 当式 (1) 中 $t_j = 0, j = 1, 2, \cdots, n$ 且 $\theta_{k_0}^{(f)} = 0, k_0 = 1, 2, \cdots, p^{(f)}$ 时, MGTWR 模型就退化为不含全局因素的 GWR 模型; 进而当 GWR 模型不存在样本的空间位置信息时就退化为 OLR 模型. 这一关系表明 OLR、GWR、MGWR、GTWR 模型都可以看做是混合时空地理加权回归模型的特殊形式.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.2 参数估计

在给出参数估计方法之前, 我们首先定义一些概念: 固定系数指标所带来的影响称为固定效应, 若固定系数指标含有部分时空因素的干扰则称其带来的影响为交互效应, 变系数指标所带来的影响称为时空效应. 现在 $\boldsymbol{\theta}^{(f)}$、$\boldsymbol{\theta}^{(r)}$ 是需要被估计的参数, 根据两步估计法我们假设 $\boldsymbol{\theta}^{(f)}$ 已知且 $A^{(f)}\boldsymbol{\theta}^{(f)}$ 能够解释 Y 的所有固定效应, 那么由指标 $A^{(r)}$ 单独得到的空间变系数回归模型就不存在任何固定效应. 现将式 (2) 改写为如下形式:

$$Y - A^{(f)}\boldsymbol{\theta}^{(f)} = A^{(r)}\boldsymbol{\theta}^{(r)} + \boldsymbol{\varepsilon} \tag{3}$$

基于已经给定的假设, 等式 (3) 的左侧应当只存在时空效应, 那么式 (3) 整体上就可以看做是 GTWR 模型. 接下来记 $Y - A^{(f)}\boldsymbol{\theta}^{(f)} = Y^{(r)}$. 对于 $\boldsymbol{\theta}^{(r)}$ 的估计可以采用加权最小二乘估计得到, 假定 ${A^{(r)}}^T A^{(r)}$ 的逆矩阵存在, 此时有如下目标函数:

$$J(\hat{\boldsymbol{\theta}}^{(r)}) = \mathbf{minimize}\frac{1}{2}(A^{(r)}\boldsymbol{\theta}^{(r)} - Y^{(r)})^T W(A^{(r)}\boldsymbol{\theta}^{(r)} - Y^{(r)}) \tag{4}$$

其中 W 是时空核函数在 h 窗宽下得到的时空权重矩阵:

$$W = \begin{bmatrix} w_1(h) & 0 & 0 & 0 \\ 0 & w_2(h) & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & w_n(h) \end{bmatrix},$$

式 (4) 对 $\boldsymbol{\theta}^{(r)}$ 求偏导数有:

$$\begin{aligned} \frac{\partial J(\hat{\boldsymbol{\theta}}^{(r)})}{\partial \boldsymbol{\theta}^{(r)}} &= \frac{1}{2}\frac{\partial((A^{(r)}\boldsymbol{\theta}^{(r)} - Y^{(r)})^T W(A^{(r)}\boldsymbol{\theta}^{(r)} - Y^{(r)}))}{\partial \boldsymbol{\theta}^{(r)}} \\ &= \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\theta}^{(r)}}({\boldsymbol{\theta}^{(r)}}^T {A^{(r)}}^T W A^{(r)}\boldsymbol{\theta}^{(r)} - {\boldsymbol{\theta}^{(r)}}^T {A^{(r)}}^T W Y^{(r)} - {Y^{(r)}}^T W A^{(r)}\boldsymbol{\theta}^{(r)} + {Y^{(r)}}^T W Y^{(r)}) \\ &= \frac{1}{2}(2{A^{(r)}}^T W A^{(r)}\boldsymbol{\theta}^{(r)} - 2{A^{(r)}}^T W Y^{(r)}) \end{aligned}$$

令上式整体等于 0, 整理后得:

$${A^{(r)}}^T W A^{(r)}\boldsymbol{\theta}^{(r)} = {A^{(r)}}^T W Y^{(r)} \tag{5}$$

进而有 $\boldsymbol{\theta}^{(r)}$ 的估计:

$$\hat{\boldsymbol{\theta}}^{(r)} = ({A^{(r)}}^T W A^{(r)})^{-1}{A^{(r)}}^T W Y^{(r)} \tag{6}$$

同时记 $({A^{(r)}}^T W A^{(r)})^{-1}{A^{(r)}}^T W = \boldsymbol{S}$, 则有 $\hat{\boldsymbol{\theta}}^{(r)} = \boldsymbol{S}Y^{(r)}$, 通常 $\boldsymbol{S}$ 被称为回归系数的帽子矩阵. 实际上, 式 (6) 就定义了一个基于一种时空核函数的 GTWR 算法过程, 再将其代入到式 (3) 整理后得:

$$(\boldsymbol{I} - \boldsymbol{S})Y = (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{\theta}^{(f)}A^{(f)} \tag{7}$$

$\boldsymbol{I}$ 是与 $\boldsymbol{S}$ 同阶的单位矩阵, 式 (7) 可以看做是普通线性回归模型, 则 $\boldsymbol{\theta}^{(f)}$ 的估计可通过最小二乘估计法 (OLS) 给出:

$$\hat{\boldsymbol{\theta}}^{(f)} = [{A^{(f)}}^T(\boldsymbol{I} - \boldsymbol{S})^T(\boldsymbol{I} - \boldsymbol{S})A^{(f)}]^{-1}{A^{(f)}}^T(\boldsymbol{I} - \boldsymbol{S})^T(\boldsymbol{I} - \boldsymbol{S})Y \tag{8}$$

进而可以给出 $\boldsymbol{\theta}^{(r)}$ 在利用固定系数变量观测值信息时估计的完整形式:

$$\hat{\boldsymbol{\theta}}^{(r)} = \boldsymbol{S}W(Y - \hat{\boldsymbol{\theta}}^{(f)}A^{(f)}). \tag{9}$$

2.3 时空核函数

由式 (6) 不难发现时空权重矩阵的设定很大程度上决定了 $\boldsymbol{\theta}^{(r)}$ 估计的好坏. 为计算 W 我们需要通过时空核函数来反映观测值间的距离关系. 定义样本 i 与 j 的空间距离为闵可

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

夫斯基距离: $d_{ij}^S = \|\sum_{k=1}^m (x_i - x_j)\|^P$, 当 $P=1$ 时 d_{ij}^S 为曼哈顿距离, 当 $P=2$ 时 d_{ij}^S 时为欧式距离, $P \to \infty$ 时 d_{ij}^S 为切比雪夫距离. 为保证核函数的有效性, 取时间距离的平方:$d_{ij}^T = (t_i - t_j)^2$, 那么样本 j 与 i 间的时空距离可以表达为:

$$d_{ij}^{ST} = \sqrt{\lambda \sum_{k=1}^{m} (x_i - x_j)^2 + \mu (t_i - t_j)^2} \tag{10}$$

其中 λ 和 μ 分别是空间和时间因子, 用来衡量两样本点间空间和时间距离的相对重要性且 $\lambda + \mu = 1$. 此时需要通过时空核函数来为样本分配权重, 常用的核权函数有:

1) 高斯核函数:

$$\mathbf{K}(h) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{d_{ij}^{ST}}{h} \right)^2 \right\} \tag{11}$$

其中 h 是窗宽. 这种核函数在不同样本上的变化较为稳定的同时解决了空间邻接距离所存在间断的问题, 但在样本量较大的情况下, 使用高斯核函数的 GTWR 算法寻找最优窗宽的收敛速度较慢.

2) 双平方核函数:

$$\mathbf{K}(h) = \begin{cases} \left[1 - \left(\dfrac{d_{ij}}{h} \right)^2 \right]^2, & \text{当}j\text{属于}i\text{最近的}h\text{个样本时} \\ 0, & \text{否则} \end{cases} \tag{12}$$

双平方核函数也称截尾型高斯核函数, 因其仅考虑窗宽范围内的所有样本点, 因此收敛速度较快. 通过合理的设定核函数窗宽和时空因子, 就能得到时空权重矩阵. 核函数的最优窗宽可以由留一交叉验证 (Leave-One-Out Cross Validation) 方法确定 [6,7], 即选取 h, 使:

$$\mathbf{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_{\neq 1}(h))^2 \tag{13}$$

值达到最小. 其中 $\hat{Y}_{\neq 1}(h)$ 是在窗宽 h 下, 去掉第 i 组观测值后, 利用 GTWR 算法所得到的 (u_i, v_i, t_i) 处响应变量的估计值. 留一法交叉验证使用的训练集与原始数据集相比仅相差一个样本, 这就使得在绝大多数情况下留一法所训练的模型往往被认为更加准确. 同样的, 留一法在数据集较大时的计算成本是难以接受的. 用于衡量样本间距离关系的核函数还有很多种, 本文不再赘述.

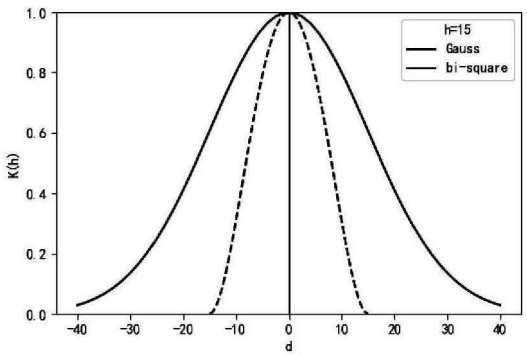

图 1　$h = 15$ 时两种核函数随距离 d 的变化情况

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

图 1 反映了两种核函数在相同窗宽下 ($h = 15$) 对距离的敏感程度, 双平方核函数在距离到达 20 个单位后权重就已经衰减为 0, 而高斯核函数则在距离到达 40 个单位后逐渐衰减至 0. 因此对于样本间空间距离较远的问题, 通常会考虑使用双平方核函数来加快交叉验证的收敛速度从而减少计算量.

2.4 变量选择算法

回顾 2.2 节中的参数估计方法, 虽然我们通过两步估计法给出了 $\boldsymbol{\theta}^{(f)}$ 和 $\boldsymbol{\theta}^{(r)}$ 的估计, 但观察式 (8) 和式 (9) 可以发现 $\hat{\boldsymbol{\theta}}^{(f)}$ 的计算结果依赖于 $\boldsymbol{S}$ 的估计结果, 且 $\boldsymbol{S}$ 的计算会涉及时空权重矩阵的确定, 其中时空核函数在不同的窗宽下会导致 $\boldsymbol{S}$ 的估计值发生显著变动, 这表明 $\boldsymbol{\theta}^{(f)}$ 的两步估计实际上仍然含有时空效应且将不再是固定值, 这显然与我们的前提假设不符. 为此, 本文提出基于变量选择的方法来剔除 $A^{(f)}$ 中的时空效应, 使其估计值不利用变系数指标的观测值信息. 其算法过程如下:

1) 计算固定效应的全残差. 考虑变量 $A^{(f)}$ 存在着时空效应, 且 $\boldsymbol{\theta}^{(f)}$ 的估计需要利用变系数指标的样本观测值 $A^{(r)}$ 的信息, 因此 $\boldsymbol{\theta}^{(f)}$, $\boldsymbol{\theta}^{(r)}$ 之间存在着交互效应, 在计算 $\hat{\boldsymbol{\theta}}^{(f)}$ 之前, 我们首先将固定系数指标中的每一个指标都作为 $\boldsymbol{\theta}^{(r)}A^{(r)}$ 的结果变量并进行 GTWR 算法, 此时有:

$$A_p^{(f)} = \boldsymbol{\theta}^{(r)}A^{(r)} + \varepsilon_p, \quad p = 1, 2, \ldots, p^{(f)} \tag{14}$$

其中 $A_p^{(f)}$ 表示第 p 个固定系数指标, 由此能够从上式获得与第 p 个指标相对应的残差:

$$e_{(p)} = A_p^{(f)} - \hat{A}_p^{(f)} \tag{15}$$

2) 检验变量的时空效应. 将 (1) 步中的每一残差进行时空莫兰指数 (Temporal-Moran' s I) 检验, 检验统计量如下 [8−9]:

$$I = \frac{n\sum_{i=1}^{n}\sum_{i=1}^{n} we_{p,i}e_{p,\neq i}}{\sum_{i=1}^{n}\sum_{i=1}^{n} we_{p,i}^2} \tag{16}$$

其中 w 是时空权重矩阵. $\mathbf{Moran's\ \ I} > 0$ 表示时空正相关性, 其值越强, 相关性越明显. $\mathbf{Moran's\ \ I} < 0$ 表示时空负相关性, 其值越小, 负相关性越强. 否则, $\mathbf{Moran's\ \ I}$=0, 时空分布呈随机性. 若第 p 个指标的残差 e_p 存在着时空效应, 则将该变量剔除, 重复第二步以遍历所有指标直至对任意的 e_p 都不存在时空效应. 当固定系数指标较少且全都含有时空效应时, 可将被剔除的变量重新加入变系数指标部分, 则 MGTWR 模型退化为 GTWR 模型. 重复上述步骤后, 此时我们获得无时空效应的 $A^{(f)}$, 同时记所有不存在时空效应的残差矩阵为 $e(A^{(f)})$.

3) 计算时空效应残差. 将 Y 作为结果变量与 $\boldsymbol{\theta}^{(r)}A^{(r)}$ 进行 GTWR 算法, 通过调整时空核函数的窗宽和时空因子最大程度解释 Y 中的时空效应, 因此这一过程的所得到的残差时空效应已经被完全解释从而 $e(A^{(r)}) = \hat{Y} - Y$ 只包含固定效应. 即:

$$\boldsymbol{\theta}^{(r)}A^{(r)} - Y = e(A^{(r)}) \tag{17}$$

4) 计算固定系数估计值. 此时的 $e(A^{(f)})$, $e(A^{(r)})$ 都不存在时空效应, 再令 $e(A^{(r)})$ 为结果变量, $e(A^{(f)})$ 为自变量利用最小二乘法 (OLS) 方法即可解释 Y 中的固定效应从而得到无时空效应的估计值. 即:

$$e(A^{(r)}) = \boldsymbol{\theta}^{(f)}e(A^{(f)}) + \varepsilon^{(f)} \tag{18}$$

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

5) 计算变系数估计值. 再次考虑 $Y-\boldsymbol{\theta}^{(f)}A^{(f)}=\boldsymbol{\theta}^{(r)}A^{(r)}+\boldsymbol{\varepsilon}$, 由于等式左侧已经剔除了时空效应, 此时利用式 (8) 就可以得到 $\boldsymbol{\theta}^{(r)}$ 的估计值. 最后模型的总残差为:

$$\boldsymbol{\varepsilon}=\boldsymbol{\varepsilon}^{(f)}+\boldsymbol{\varepsilon}^{(r)}. \tag{19}$$

3 实证分析

为验证本文提出的变量选择方法的有效性, 利用网络爬虫获取乌鲁木齐市商品住宅的价格数据, 选取影响房价变化的指标进行验证分析. 通过查阅文献共选取包括房屋基本属性、交通便利性、医疗便利性、生活便利性在内的 12 个指标 [10−15], 总计爬取同一研究区域内的 275 个住宅点, 并将指标分为固定系数指标和变系数指标, 见表 1. 显然房价在空间上的分布会受到其附近各类指标的影响而呈现出空间差异性, 通常称为区位效应. 但房价的影响因素中同样也存在着全局稳定的指标, 例如本文所选取的人口密度、就业率、建筑面积, 在同一研究范围内它们并不会受到时空效应的影响. 首选将数据集随机分为总量 80% 的训练集用于建立模型, 20% 的交叉验证集用于对模型性能进行验证. 借助 Python 编程分别实现 GTWR 算法、两步估计法 (Two-Step)、变量选择法 (Variable Selection) 对数据进行拟合. 在执行算法时通过对比迭代 CV 值的收敛情况选择最优模型, 并将三种算法所得到的的最优模型用于验证集进行性能对比. 三种算法的性能如表 2 所示, 其交叉验证的收敛情况如图 2 所示.

表 1　指标选取与估计结果

指标	指标类型/系数估计值 *	指标	指标类型/系数估计值 *
单价	变系数/0.013	最近公园距离	变系数/–1.286
建筑面积	固定系数/0.371	容积率	变系数/0.187
楼龄	变系数/1.845	绿化率	变系数/0.376
最近公交站距离	变系数/–1.025	人口密度	固定系数/0.335
最近三甲医院距离	变系数/–0.078	就业率	固定系数/1.338
最近药房距离	变系数/–0.136	最近购物点距离	变系数/–0.076

表 2　算法性能对比

算法	CV	R^2	RSS	h
GTWR	514.51	0.78	141489.48	0.315
TS-GTWR	505.62	0.81	109842.85	0.258
VS-GTWR	489.21	0.86	75335.11	0.173

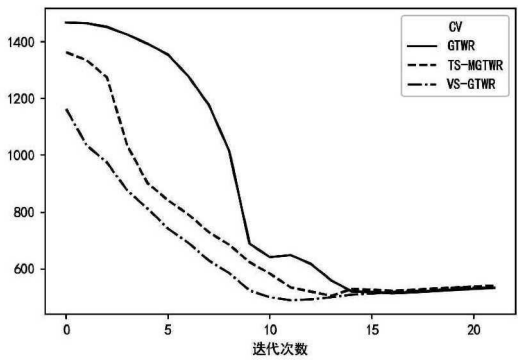

图 2　三种算法下选取最优模型时的 CV 值

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

变系数指标在每一样本点上都存在着回归系数, 限于篇幅本文仅给出回归系数估计值的均值. 从验证集得到的结果来看, VS-MGTWR 模型显著优于 TS-MGTWR 模型. 基于变量选择方法的 MGTWR 模型的核函数最优窗宽为 0.173, 由于在拟合数据时已经充分去除了固定指标的时空效应, 其 CV 值的收敛过程较为平稳, 表明数据的全局属性和局部属性都得到了较好的解释, 因此相比较两步估计法和 GTWR 法拥有较小的窗宽. 从交叉验证的收敛情况来看, 变量选择法的 CV 值显著优于另两种算法且下降收敛的速度更快. 变量选择法的 R^2 值相较于两步估计法有了提升, 由于样本量与指标数量的限制, 性能提升并不明显. 实验结果表明, 变量选择的方法使得 MGTWR 模型的参数估计变得更加合理, 同时也可以在高维数据情况下将不符合参数估计假设的指标予以剔除从而达到对模型进行简化的目的. 由于所选指标过多和固定系数指标在全局上并不发生变化, 本文将部分指标的回归系数进行可视化, 结果如图 3 所示.

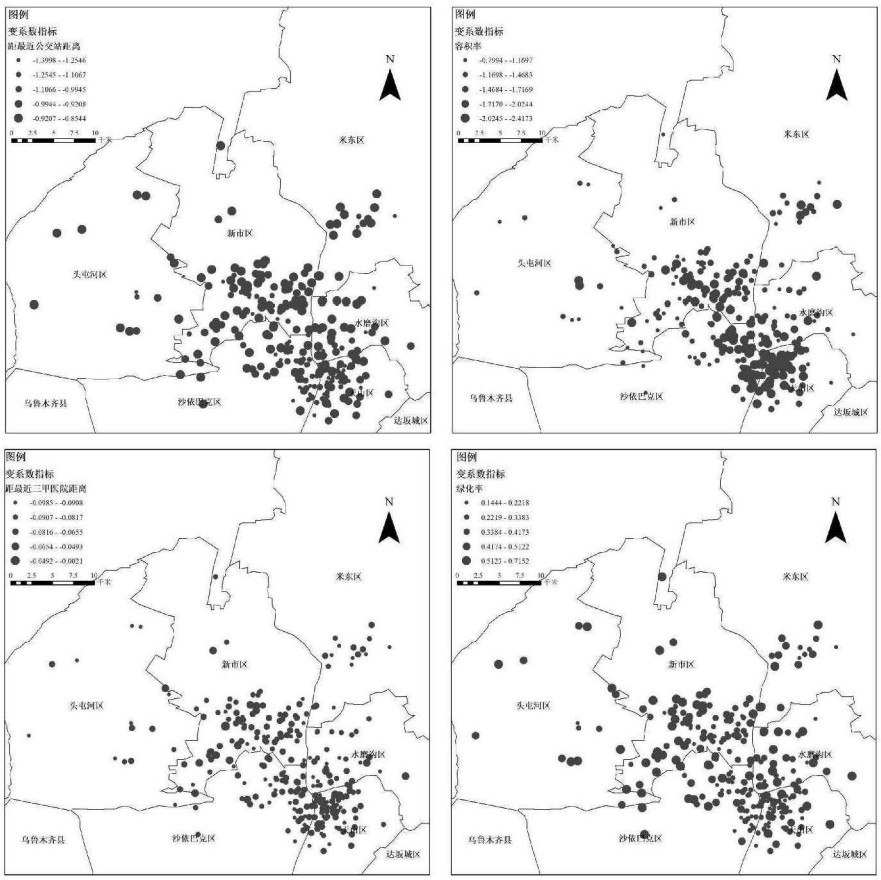

图 3 部分指标回归系数可视化

从可视化的回归结果来看, 回归系数的符号同预期保持一致. 与 OLS 方法不同的是, 混合时空地理加权回归模型在估计过程中考虑了指标在局部和全局上的变化情况, 受空间位置和时间变化情况的影响, 变系数指标回归系数的空间分布具有明显的异质性. 其中受最近三甲级医院距离影响较大的样本点多分布于各市区的周边且均是受到负向影响, 它们通常距离

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

三甲级医院较远, 房价会随这种距离的增加而下降的更快. 容积率的影响情况有正有负, 这同天山区、沙依巴克区人口密集, 新市区头屯河区人口稀疏有关, 人口过于密集的同时建筑密度较大导致天山区、沙依巴克区的住房价格受容积率的影响较为显著, 而城市北部的新市区由于建筑密度低受容积率影响较小. 绿化率总体上均呈现对房价的正向影响且在各区差异不大, 但天山区南部绿化率较高的住宅区受绿化率的影响较小, 这表明绿化率正成为影响房价构成的重要通用指标. MGTWR 模型很好的反映了各指标影响能力在时空上的变化情况, 能够为商品住宅开发的区位选择提供一定的参考.

4 结论

针对 MGTWR 两步估计法中固定系数指标可能存在时空效应, 导致参数估计的假设无法成立的问题, 本文提出变量选择方法来检验和去除时空效应, 并且给出了具体的算法步骤, 一定程度上为解决固定系数存在时空效应时的参数估计给出了办法, 该方法同时可以在高维数据情况下对模型进行简化, 避免直接将所有指标直接带入计算导致过拟合. 以乌鲁木齐市真实房价数据对算法的性能进行验证, 从模型性能的角度来看, 通过变量选择法所获得的最优模型拥有更低的 CV 值和更高的拟合优度, 显著优于两步估计法的性能. 尽管该算法有效去除了固定系数的时空效应, 但在两步估计法的基础上再一次加入了新的迭代过程, 当指标过多时计算量会显著提升, 还需要进一步考虑超高维数据时变量选择算法的优化问题.

参考文献

[1] Brunsdon C, Fotheringham A S, Charlton ME. Geographically weighted regression: a method for exploring spatial nonstationarity[J]. Geographical Analysis, 1996, 28(4): 281-298.

[2] Brunsdon C, Fotheringham A S, Charlton ME. Some notes on parametric signficance tests for geographically weighted regression[J]. Journal of Regional Science, 1999, 39(3): 497-524.

[3] 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(3): 383-401.

[4] A Stewart Fotheringham, Ricardo Crespo, Jing Yao. Geographical and temporal weighter regression(GTWR)[J]. Geographical Analysis, 2015, 47(4): 431-452.

[5] 赵阳阳, 刘纪平, 杨毅, 张福浩, 仇阿根. 混合时空地理加权回归及参数的两步估计 [J]. 计算机科学, 2017, 44(3): 274-277+312.

[6] 玄海燕, 罗双华, 王大斌. GWR 模型中权函数的选取与窗宽参数的确定 [J]. 甘肃联合大学学报 (自然科学版), 2008(3): 10-12.

[7] 张琰, 梅长林. 基于地理加权回归的我国中东部城市商品房价格的空间特征分析 [J]. 数理统计与管理, 2012, 31(5): 898-905.

[8] Goodchild M F. Spatial Autocorrelation[M]. Norwich, England: Geo Books, 1986.

[9] Ord J K, Arthur Getis. Local spatial autocorrelation statistics: distributional issues and an application[J]. Geographical Analysis, 1995, 27(4): 286-306.

[10] 尹上岗, 宋伟轩, 马志飞, 李在军, 吴启焰. 南京市住宅价格时空分异格局及其影响因素分析——基于地理加权回归模型的实证研究 [J]. 人文地理, 2018, 33(3): 68-77.

[11] 王少剑, 王婕妤, 王洋. 土地价格对住房价格空间分异的影响——基于中国县域单元的实证分析 [J]. Journal of Geographical Sciences, 2018, 28(6): 725-740.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[12] 汤庆园, 徐伟, 艾福利. 基于地理加权回归的上海市房价空间分异及其影响因子研究 [J]. 经济地理, 2012, 32(2): 52-58.
[13] 周祥, 王丽娅. 城市交通便利度对房价影响研究——基于 14 座新一线城市面板数据分析 [J]. 价格理论与实践, 2019(10): 48-51.
[14] 汪佳莉, 季民河, 邓中伟. 基于地理加权特征价格法的上海外环内住宅租金分布成因分析 [J]. 地域研究与开发, 2016, 35(5): 72-80.
[15] 吕萍, 甄辉. 基于 GWR 模型的北京市住宅用地价格影响因素及其空间规律研究 [J]. 经济地理, 2010, 30(3): 472-478.

Parameter Estimation of Miexd-GTWR Baesd on Variable Selection

HOU Jian[1,2], TIAN Mao-zai[1,2,3,4,5]

(1. Xinjiang Center for Socio-Economic Statistics, Xinjiang University of Finance and Economics, Urumqi 830012, China)
(2. School of Statistics and Data Science, Xinjiang University of Finance and Economics, Urumqi 830012, China)
(3. Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing 100872, China)
(4. School of Statistics, Renmin University of China, Beijing 100872, China)
(5. School of Statistics, Lanzhou University of Finance and Economics, Lanzhou 730101, China)

Abstract: Mixed-Geographical and Temporal weighted regression model as a kind of effective processing spatial data of global smooth and local non-stationary analysis method has been widely used. But its parameter estimation method has a hypothesis that we have already known the fixed variable and there is no geographical and temporal effect, the hypothesis of the strong make estimates of the regression coefficient too unstable. To explore when fixed coefficient of Variable have space-time effects of the parameter estimation method, this paper proposes a variable selection method to eliminate the interaction effect between indicators, and the corresponding algorithm process is given. Through Urumqi commodity residential real price data for comparing different estimation methods validation, the results show that the use of variable Selection method after MGTWR model performance and improved fitting effect, fixed estimate of regression coefficients is more stable, the original parameter estimation methods of improvement.

Keywords: MGTWR; variable selection; two-step estimation

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net