

基于异窗宽 GTWR 模型的商品住宅价格影响因素研究

侯健^{1,2,3} 王芝皓^{1,2,3} 田茂再^{1,2,3} 窦燕³

(1. 中国人民大学应用统计中心, 北京 100872; 2. 中国人民大学统计学院, 北京 100872; 3. 新疆财经大学统计与数据科学学院, 新疆 乌鲁木齐 830012)

摘要: 随着数据可获得性的增强, 时空数据被广泛地应用在各个领域。以中国西部干旱区代表城市乌鲁木齐为研究对象, 本文根据标准差椭圆加权算法与时空地理加权回归模型 (GTWR) 提出异窗宽 GTWR 模型对房屋价格变动影响因素的时空变化进行研究。结果表明乌鲁木齐市住房价格的变动存在着明显的空间异质性, 异窗宽 GTWR 模型对其的解释能力较好, 标准差椭圆加权算法有效的减少了计算量, 房价变动在空间上受交通便利性和绿化率等因素的影响较大, 时间上受建筑年龄因素影响较大。

关键词: 时空数据; 干旱区商品住宅; 空间异质性; 标准差椭圆算法; 异窗宽时空地理加权回归模型

中图分类号: F293.3, F224, O212

文献标识码: A

Research on Influencing Factors of House Prices Based on Multi-Bandwidth GWTR Model

HOU Jian^{1,2,3} WANG Zhi-hao^{1,2,3} TIAN Mao-zai^{1,2,3} DOU Yan³

(1. Center for Applied Statistics, Renmin University of China, Beijing 100872, China; 2. School of Statistics, Renmin University of China, Beijing 100872 , China; 3. School of Statistics and Data Science, Xinjiang University of Finance, Urumqi 830012, China)

Abstract: With the increasing availability of data, spatiotemporal data is widely used in various fields. Taking Urumqi, a representative city in the arid region of western China as the research object, in this paper, the Multi-Bandwidth GWTR model is proposed to study the space and time changes of factors affecting house price changes based on the standard deviation elliptic weighting algorithm and the geographical and temporal weighted regression (GTWR). The results are showed that there is obvious spatial heterogeneity in the changes of urban house prices in the western arid area. A better interpretation performance has been represented by the Multi-Bandwidth GWTR mode. The standard deviation ellipse weighting algorithm effectively reduces the calculation amount, and the housing price changes in space It

收稿日期: 2020 年 8 月 13 日 **收修改稿日期:** 2021 年 2 月 11 日 **通讯作者:** 田茂再, mztian@ruc.edu.cn
基金项目: 中国人民大学科学研究基金 (中央高校基本科研业务费专项资金资助) (22XNL016)。

is greatly influenced by factors such as transportation convenience and greening rate, and time is greatly influenced by building age.

Key words: spatiotemporal data; house price in arid areas; spatial heterogeneity; standard deviation ellipse algorithm; multi-bandwidth GWTR model

0 引言

近年来,随着我国城市化进程不断加快,带动了房地产事业的飞速发展,并逐渐成为我国经济发展的支柱产业之一。众多学者对不同地域的房价特征进行过研究,如陈蜒和张志斌(2015)^[1]利用空间反距离权重法、核密度估计法对干旱区城市兰州市地区的房价空间分布格局进行了计量分析;王新刚和孔云峰(2015)^[2]为了补充 GWR 方法的不足,通过构造局部时空窗口统计量的方法建立了基于时空窗口改进的时空地理加权回归模型,对湖北黄石的住房价格空间异质性进行了探究;孙倩和汤放华(2015)^[3]基于空间扩展模型和地理加权回归模型对长沙市住房价格空间分异性进行了对比分析。最近,越来越多的学者开始关注商品住宅价格的空间异质性,石振武和王喆(2018)^[4]考虑商品住宅的时间属性,在传统影响因素的基础上加入近户年份,建立基于分位数的混合时空地理加权回归模型探究了哈尔滨市商品住宅价格的空间异质性。尹上岗等(2018)^[5]以南京市商品房社区为基本研究单元,运用克里金(Kriging)插值法对住宅价格空间分布进行模拟和估计,并利用地理加权回归(GWR)模型探究社区属性、商业区位、交通区位、服务区位和景观区位等类型变量对住宅价格的影响规律。王少剑等(2018)^[6]利用地理统计方法对土地价格对住房价格空间分异的影响进行了实证分析。苏方林(2010)^[7]则运用地理加权回归方法,对1997—2002年期间中国地级市R&D知识溢出的空间非稳定性进行了实证分析,得出R&D知识生产的不同要素存在空间变异的结论。李斌等(2019)^[8]则利用空间计量经济学中的莫兰指数检验和空间自回归模型等方法探究了房地产业对中国城市金融稳定的影响。

然而国内少有文献研究异窗宽 GTWR 模型,即在模型的时空核函数中可根据空间位置的变化而选取不同的窗宽,国外研究中, Fotheringham 等(2017)^[9]首次提出了可加模型形式的多尺度 GWR 模型,使得样本可以不再使用相同窗宽并将其应用于爱尔兰大饥荒(Irish Famine)的致因研究。传统模型通过交叉验证(Cross Validation, CV)或赤池信息准则(Akaike information criterion, AIC)选取全局最优核函数窗宽,且所有样本在研究区域内拥有相同的核函数窗宽,这可能会导致在探究某些问题时样本间潜在的差异性被忽略。而在实际研究中,任意空间位置点处的窗宽选择应是变化的,因此本文基于标准差椭圆构建异窗宽时空地理加权回归模型(Multi-Bandwidth GWTR),利用标准差椭圆算法给出异窗宽时空权重矩阵,推导了参数估计的结果,通过对我国西部干旱区代表性城市乌鲁木齐市的房价影响因素进行时空分析并将研究区域划分为不同的子区域,从而得到异窗宽 GTWR 模型,相比较传统模型一定程度上减少了计算量并验证了该方法的优越性和弥补对干旱区城市商品住宅价格影响因素研究的不足。

1 研究方法

1.1 模型定义

首先给出同窗宽 GTWR 模型的定义。Wu 等(2010)^[10]在地理加权回归(GWR)模型的基

础上引入样本的时间属性, 提出了时空地理加权回归 (GTWR) 模型, 有如下形式:

$$y_i = \beta_0(u_0, v_0, t_0) + \sum_{j=1}^p \beta_{ij}(u_i, v_i, t_i)x_{ij} + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

其中 (u_i, v_i, t_i) 为第 i 个样本的空间和时间坐标, 通常 u_i 和 v_i 分别表示经度和纬度而 t_i 表示样本当前的时间戳, $\beta_{ij}(u_i, v_i, t_i)$ 则是第 i 个样本对应第 j 个指标的回归系数, p 是指标数量。时空地理加权回归模型采用局部最优思想下的加权最小二乘估计法, 根据其原理, 对于任意样本点处的回归系数 β 需要满足如下目标函数:

$$\hat{\beta} = \operatorname{argmin} \frac{1}{2} \|W_U^{\frac{1}{2}}(Y - X\beta)\|^2. \quad (1)$$

上式中 W_U 是同窗宽时空权重矩阵, 对于第 j 位置样本点 (u_j, v_j, t_j) 的时空权重矩阵为:

$$W_U = \operatorname{diag}(w_1(u_j, v_j, t_j, h), \dots, w_n(u_j, v_j, t_j, h)),$$

其中 $w_i(u_j, v_j, t_j, h)$, $i = 1, 2, \dots, n$ 表示给定 h 窗宽下第 i 个样本点与第 j 个样本点的时空权重:

$$w_i(u_j, v_j, t_j, h) = K_h(d_{ij}^{ST}),$$

d_{ij}^{ST} 表示第 i 位置样本点与第 j 位置样本点之间的时空距离:

$$d_{ij}^{ST} = \lambda d_{ij}^S + \mu d_{ij}^T = \lambda \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} + \mu(t_i - t_j)^2.$$

上式通过调节参数 λ 和 μ 调整空间距离与时间距离的相对重要性, $K_h(d_{ij}^{ST})$ 为时空核函数, 通常可取 Gauss 核函数:

$$K_h(d_{ij}^{ST}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{d_{ij}^{ST}}{h}\right)^2\right),$$

或 Bi-square 核函数:

$$K_h(d_{ij}^{ST}) = \left(1 - \left(\frac{d_{ij}^{ST}}{h}\right)^2\right)^2 I(d_{ij}^{ST} < h).$$

窗宽对样本权重起到放缩的作用, 不难看出, 在该定义下时空权重矩阵令研究区内的所有样本使用相同的窗宽 h , 即在相同水平下衡量样本间的权重。虽然这种设定能够有效简化最优窗宽的计算, 但是忽略了在实际研究中, 样本受地理位置的影响在某些因素上可能会产生较大的差异, 这意味着如果在相同的窗宽下计算样本权重值将会忽略这种差异性。因此, 本文考虑样本在不同地理区域内的差异性, 将样本划分为不同的子区域并给出异窗宽时空权重矩阵的定义:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix}, \quad X = \begin{pmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_m \end{pmatrix}, \quad W_D = \begin{pmatrix} W_1 & 0 & \cdots & 0 \\ 0 & W_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_m \end{pmatrix}. \quad (2)$$

上式表明我们将总研究区域 S 内的样本划分为了 m 个不同的子样本且满足 $m \leq n$, 其中 X_k 和 Y_k ($k = 1, 2, \dots, m$) 分别表示第 k 个子区域内的自变量和结果变量矩阵且 n_k 是该子区域

对应的样本量, $W_k = \text{diag}(w_1(u_i, v_i, t_i, h_k), \dots, w_{n_k}(u_i, v_i, t_i, h_k))$, $k = 1, \dots, m$ 是第 k 个子区域的异窗宽时空权重矩阵。由此能够给出异窗宽 GTWR 模型的定义:

$$y_i = \beta_{D_0}(u_0, v_0, t_0) + \sum_{j=1}^p \beta_{D_{ij}}(u_i, v_i, t_i) x_{ij} + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

其中 $\beta_{D_0}(u_0, v_0, t_0)$ 是异窗宽截距项, $\beta_{D_{ij}}(u_i, v_i, t_i)$ 是第 i 个样本对应的第 j 个异窗宽回归系数, ε_i 是第 i 个随机扰动项。此时对应的 Gauss 核函数和 bi-square 核函数分别改写为:

$$w_i(u_j, v_j, t_j, h_k) = K_h(d_{ij}^{ST}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\left(\frac{d_{ij}^{ST}}{h_k}\right)^2\right), \quad k = 1, 2, \dots, m,$$

和

$$w_i(u_j, v_j, t_j, h_k) = K_h(d_{ij}^{ST}) = \left(1 - \left(\frac{d_{ij}^{ST}}{h_k}\right)^2\right)^2 I(d_{ij}^{ST} < h_k), \quad k = 1, 2, \dots, m.$$

1.2 参数估计及窗宽选择

将式(2)中样本在 m 个区域上的划分改写为: $Y = (Y_1^T, Y_2^T, \dots, Y_m^T)^T$ 和 $X = \text{diag}(X_1, X_2, \dots, X_m)$ 。在同窗宽 GTWR 模型的基础上考虑异窗宽时空权重矩阵 $W_D = \text{diag}(W_1, W_2, \dots, W_m)$ 并将其代入目标函数式(1)后有第 k 个区域内回归系数的估计:

$$\hat{\beta}_{Dk} = \arg\min \frac{1}{2} \|W_{Dk}^{\frac{1}{2}}(Y_k - X_k \beta)\|^2. \quad k = 1, 2, \dots, m, \quad (3)$$

其中 Y_k, X_k 分别是取自第 k 个子区域的结果变量和自变量矩阵且 $Y_k = (y_{k1}, y_{k2}, \dots, y_{kn_k})$, 定义第 k 个子区域矩阵 $A_k = (0 : I_{n_k} : 0)_{n_k \times p}$, $k = 1, 2, \dots, m$, 当 $k = 1$ 时, $A_1 = (I_{n_1} : 0)$; 当 $k = m$ 时, $A_m = (0 : I_{n_m})$ 。记式(3)为 $J(\beta_{Dk})$, 此时目标函数可写为:

$$\begin{aligned} J(\beta_{Dk}) &= \frac{1}{2} \|W_{Dk}^{\frac{1}{2}}(Y_k - X_k \beta)\|^2 \\ &= \frac{1}{2} (W_{Dk}^{\frac{1}{2}}(Y_k - X_k \beta))^T (W_{Dk}^{\frac{1}{2}}(Y_k - X_k \beta)) \\ &= \frac{1}{2} (Y_k - X_k \beta)^T W_{Dk}^{\frac{1}{2}} W_{Dk}^{\frac{1}{2}} (Y_k - X_k \beta) \\ &= \frac{1}{2} (Y_k^T - \beta^T X_k^T)^T W_{Dk} (Y_k - X_k \beta) \\ &= \frac{1}{2} (Y_k^T W_{Dk} Y_k - \beta^T X_k^T W_{Dk} Y_k - Y_k^T W_{Dk} X_k \beta + \beta^T X_k^T W_{Dk} X_k \beta). \end{aligned}$$

为最小化目标函数, 上式对 β 求偏导数:

$$\begin{aligned} \frac{\partial J(\beta_{Dk})}{\partial \beta} &= \frac{1}{2} \frac{\partial (Y_k^T W_{Dk} Y_k - \beta^T X_k^T W_{Dk} Y_k - Y_k^T W_{Dk} X_k \beta + \beta^T X_k^T W_{Dk} X_k \beta)}{\partial \beta} \\ &= \frac{1}{2} (0 - X_k^T W_{Dk} Y_k - X_k^T W_{Dk} Y_k + 2X_k^T W_{Dk} X_k \beta) \\ &= \frac{1}{2} (-2X_k^T W_{Dk} Y_k + 2X_k^T W_{Dk} X_k \beta). \end{aligned}$$

令上式等于零整理后得:

$$X_k^T W_{Dk} X_k \beta = X_k^T W_{Dk} Y_k,$$

则对于第 k 个区域内回归系数的估计 $\hat{\beta}_{Dk}$ 为:

$$\begin{aligned}\hat{\beta}_{Dk} &= (X_k^T W_{Dk} X_k)^{-1} X_k^T W_{Dk} Y_k \\ &= ((A_k X A_k^T)^T A_k W A_k^T (A_k X A_k^T))^{-1} (A_k X A_k^T)^T A_k W A_k^T (A_k Y) \\ &= ((A_k X^T A_k^T) A_k W A_k^T (A_k X A_k^T))^{-1} (A_k X^T A_k^T) A_k W A_k^T (A_k Y).\end{aligned}$$

上述参数估计过程中时空核函数的最优窗宽和调节参数 μ, λ 可在无先验知识的情况下通过交叉验证 (Cross-Validation) 方法或广义赤池信息准则 (AICc) 确定^[11-13], 即给定 (h_k, μ_k, λ_k) , $k = 1, 2, \dots, m$ 使:

$$CV(h_k, \mu_k, \lambda_k) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_{\neq i}(h_k, \mu_k, \lambda_k))^2, \quad k = 1, 2, \dots, m, \quad (4)$$

值达到最小, 其中 $\hat{Y}_{\neq i}(h_k)$ 是在给定窗宽和调节参数 (h_k, μ_k, λ_k) 的情况下, 去掉第 i 组观测值后, 拟合 GTWR 模型后得到的 (u_i, v_i, t_i) 位置处因变量的估计值。对于广义赤池信息准则^[14]有:

$$AICc_{s_k} = 2n_k \ln \hat{\sigma}_{s_k} + n_k \ln(2\pi) + \frac{n_k + \text{tr}(H_{s_k})}{n_k - 2 - \text{tr}(H_{s_k})}, \quad (5)$$

其中 $H_{s_k} = X_k (X_k^T W_{Dk} X_k)^{-1} X_k^T W_{Dk}$ 被称为第 k 个子区域 S_k 中结果变量的帽子矩阵, $\hat{\sigma}_{s_k}$ 是子区域 S_k 中回归模型随机误差项标准差的估计, $\text{tr}(H_{s_k})$ 是帽子矩阵的迹, 为获得最佳窗宽则需要最小化 $AICc$ 值。

2 标准差椭圆分类算法

异窗宽 GTWR 模型的参数估计中, S_k 是需要被确定的参数, 但对于如何划分 S 目前无文献可供参考, 因此本文提出利用标准差椭圆算法来实现对 S 的一种划分。标准差椭圆是用来度量数据分布和方向的工具, 椭圆心为样本的算数中心, 第一标准差椭圆包含 66.7% 的样本, 二标准差椭圆包含 95% 的样本, 三标准差椭圆包含 99% 的样本。对于标准差椭圆的建立有如下算法, 首先给出无方向标准差椭圆方程:

$$\left(\frac{x}{SDE_x}\right)^2 + \left(\frac{y}{SDE_y}\right)^2 = s,$$

且:

$$SDE_x = \left(\frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2\right)^{\frac{1}{2}}, \quad SDE_y = \left(\frac{1}{n} \sum_{i=1}^n (v_i - \bar{v})^2\right)^{\frac{1}{2}}. \quad (6)$$

上式中 SDE_x 和 SDE_y 分别为无方向标准差椭圆的长半轴与短半轴, u_i 和 v_i 是第 i 个样本的坐标, (\bar{u}, \bar{v}) 为样本的平均算数中心。长短半轴的长度单位与计算数据的单位相同, 接下来需要确定椭圆的方向以符合数据的空间分布。以正北方向为零度, 按顺时针旋转, 旋转角度 θ 为:

$$\theta = \arctan \frac{(\sum_{i=1}^n \tilde{u}_i^2 - \sum_{i=1}^n \tilde{v}_i^2) + ((\sum_{i=1}^n \tilde{u}_i^2 - \sum_{i=1}^n \tilde{v}_i^2)^2 + 4(\sum_{i=1}^n \tilde{u}_i^2 \tilde{v}_i^2)^2)^{\frac{1}{2}}}{2 \sum_{i=1}^n \tilde{u}_i^2 \tilde{v}_i^2}.$$

\tilde{u}_i 和 \tilde{v}_i 表示第 i 样本空间位置坐标与平均算数中心 (\bar{u}, \bar{v}) 之差。则 x 轴与 y 轴的标准差为:

$$\sigma_x = \sqrt{2} \left(\frac{1}{n} \sum_{i=1}^n (\tilde{u}_i \cos \theta - \tilde{v}_i \sin \theta)^2\right)^{\frac{1}{2}}, \quad \sigma_y = \sqrt{2} \left(\frac{1}{n} \sum_{i=1}^n (\tilde{u}_i \sin \theta - \tilde{v}_i \cos \theta)^2\right)^{\frac{1}{2}}.$$

上式代入式(6)后可以得到不同标准差下的长短半轴长度。可以看出,一般的椭圆方程长短半轴与标准差椭圆长短半轴的区别在于后者需要以样本的算术中心为圆心进行方向旋转,从而反映数据的空间分布方向。获得标准差椭圆后就能对样本进行分类,步骤如下:

(1) 给定总样本区域 S ,选择用于计算标准差椭圆长短半轴长度的指标值,计算样本平均数中心后分别建立三个标准差的无方向椭圆。

(2) 确定正北方向,按照样本的空间分布趋势对标准差椭圆进行方向旋转,分别记此时被标准差椭圆覆盖的样本子区域为 S_1, S_2, S_3, S_4 ,且满足 $\cup_{k=1}^4 S_k = S$, $\cap_{k=1}^4 S_k = \emptyset$ 。绘出样本空间分布的散点图来验证旋转的方向是否合理。

(3) 得到符合样本空间分布趋势的标准差椭圆后对不同区域的样本进行识别,然后在 S_1, S_2, S_3, S_4 区域内分别拟合 GTWR 模型,则各区域的总模型形式可写为:

$$\hat{Y}_k = X_k (X_k^T W_{Dk} X_k^T)^{-1} X_k^T W_{Dk} Y_k, \quad k = 1, 2, 3, 4.$$

由于此时样本已被标准差椭圆划分,再次考虑 W_D 的定义,则此时的异窗宽时空权重矩阵为: $W_D = \text{diag}(W_1, W_2, W_3, W_4)$ 其中 W_1 可以表示为:

$$W_1 = \text{diag}(w_1(u_1, v_1, t_1, h_1), \dots, w_{n_1}(u_{n_1}, v_{n_1}, t_{n_1}, h_1)),$$

W_2, W_3, W_4 以此类推,此时样本所属的区域已经确定,当选定某一时空核函数并按照式(4)、式(5)中的准则给出局部最优窗宽 h_1, h_2, h_3, h_4 后就能进一步确定异窗宽时空权重矩阵 W_D 。

2 模拟实验

我们在 Python 环境下实现了异窗宽 GTWR、GTWR、GWR 和 OLS 模型的拟合,采用模拟数据来分别评估和比较模型之间的性能。本文模拟数据参考赵阳阳等(2017)^[14]所使用的数据集,并在该数据集的基础上修改了样本的时空属性,增强时间变量的强度以更好的区分不同模型的性能差距。模拟过程为:设定研究区域是边长为 $m+1$ 个单位的时空正方体,其中分别用 u, v 来表示生成样本点的空间坐标,用垂直于水平面方向的 t 轴表示时间变化,令随机误差项 $\varepsilon \sim N(0, 1)$ 。模拟数据与协变量取值规则见表 1。

表 1 模拟数据生成规则

| 实验组号 | 响应变量取值 | 协变量与坐标取值 |
|------|--|--|
| Sim1 | $y = 5 + uvt x_1 + x_2 + x_3 + x_4 + \varepsilon$ | $x_1, x_2 \sim U(0, 1); u, v, t \sim U(0, 20)$ |
| Sim2 | $y = \sqrt{uv} + tx_1 + tx_2 + x_3 + \varepsilon$ | $x_1, x_2, x_3 \sim U(0, 1); u, v, t \sim U(0, 20)$ |
| Sim3 | $y = 0.76 + 2.47x_1 + utx_2 + vtx_3 + uvt x_4 + \varepsilon$ | $x_1, x_2, x_3, x_4 \sim U(0, 1); u, v, t \sim U(0, 20)$ |
| Sim4 | $y = 1 + x_1 + x_2 + x_3 + x_4 + \varepsilon$ | $x_1, x_2, x_3, x_4 \sim U(0, 1)$ |

在实验组 Sim1 中,设定较弱的时空变化属性和较强的全局线性属性,用于反映各模型是否能够捕捉到数据的时空变化;在 Sim2 中设定较强的时间变化数据和较弱的空间变化数据,用于反映模型对时间属性较强的数据的探查能力;在 Sim3 中同时设定较强的时空变化,是现实应用中经常出现的数据类型,用于反映模型的真实性能。利用 AIC_c 作为计算上述中带有时空属性模型窗宽、时空因子的准则,在拟合结果中给出各模型的 MSE 、 R^2 、 AIC_c 平均指标。为了避免实验的偶然性,将上述实验过程重复 100 次并取 m 为 20, 样本量为 500, 取各指标的平均值作为最终结果。我们另外加入了对照组 Sim4,其原因是该组数据完全不含有时空变量,异窗宽 GTWR、GTWR、GWR 模型都会退化为普通线性模型,如果他们在 Sim4 中的表现情况与 OLS 模型一致,那么就说明它们正确的进行了退化,即保证模拟过程没有设计错误。

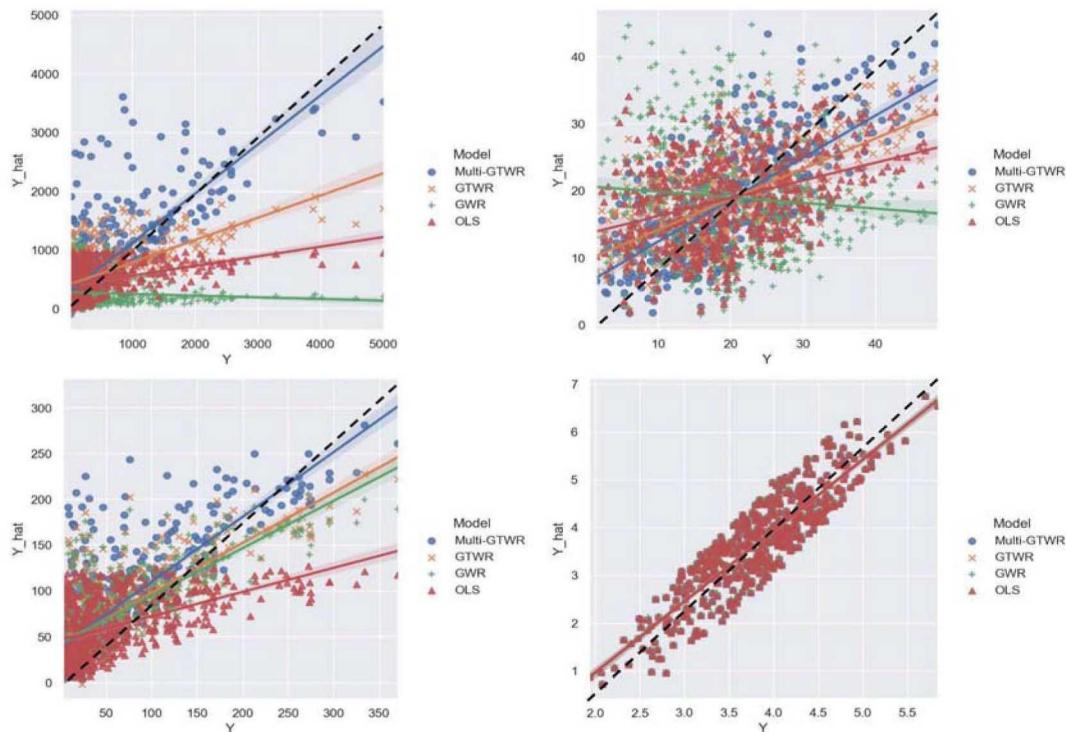


图 1 各模型随机模拟实验拟合结果

图 1 给出了各模型在不同实验组中拟合情况的可视化结果, 将横坐标设定为模拟数据的真实值, 纵坐标设定为模型基于模拟数据给出的预测值, 并加入一条 45° 趋势线(黑色虚线)来反映模型的拟合程度。可以直观的看出本文提出的异窗宽 GTWR 模型在各实验组上的拟合趋势线相比都更加接近 45° 线, 性能表现优于以往的地理加权回归模型。Sim1 (左上) 生成的数据时空变化较弱, 结果显示各类模型的拟合效果并不完美, 各模型之间的差异并不明显; Sim2 (右上) 生成中的数据更加强调时间属性的变化, 因此 Multi-GTWR、GTWR 模型的拟合效果较好, 而 GWR 模型中并不含有时间项, 因此表现较差; Sim3 (左下) 是生成规则最复杂的数据, OLS 模型在此数据上的表现最差, Multi-GTWR 的表现较好, GTWR 与 GWR 的差异不大。Sim4 (右下) 中各模型拟合结果的趋势线正确重合, 表明对无时变属性的数据拟合情况一致, 模型没有设计错误。上述模拟过程的详细结果见表 2。

表 2 各模型性能对比

| | Multi-GTWR | | | GTWR | | | GWR | | | OLS | | |
|------|------------|---------|-------|-------|---------|-------|-------|---------|-------|--------|---------|-------|
| | MSE | AICc | R^2 | MSE | AICc | R^2 | MSE | AICc | R^2 | MSE | AICc | R^2 |
| Sim1 | 20.39 | 4136.39 | 0.81 | 32.70 | 4429.22 | 0.65 | 33.54 | 4635.13 | 0.54 | 384.76 | 4587.28 | 0.59 |
| Sim2 | 68.73 | 2660.54 | 0.94 | 78.26 | 2767.79 | 0.88 | 81.90 | 2868.63 | 0.74 | 107.26 | 2908.01 | 0.66 |
| Sim3 | 17.62 | 2602.14 | 0.95 | 28.88 | 2729.14 | 0.91 | 33.73 | 2826.78 | 0.82 | 62.25 | 3021.08 | 0.45 |
| Sim4 | 0.341 | 969.260 | 0.98 | 0.341 | 969.260 | 0.98 | 0.341 | 969.260 | 0.98 | 0.341 | 969.260 | 0.98 |

表 2 给出了各模型的具体指标数值, Multi-GTWR 在各组数据中的 R^2 值都大于 0.80, 拟合程度较高, MSE 和 AICc 方面也在不同的实验组中优于其他模型, 另外也说明 Multi-GTWR 能够根据数据不同的分散情况来适应最优窗宽, 从而拟合效果更优且更加稳定。

3 实证分析

在商品住宅价格影响因素的相关文献中, 其中主要考虑了建筑所在小区的基本属性、周边地区的交通便利性、医疗便利性、生活便利性等指标^[15-19]。综合以上文献选取以下 10 个指标: 单价、面积、楼龄、距最近公交站距离、800 米内公交站数量、距最近三甲医院距离、距最近药店距离、距最近综合商场距离、容积率、绿化率作为影响房价时空变动的主要因素, 见表 3。其中指标数据来源于权威互联网房屋交易平台“房天下”, 共获取同一研究区域内的 388 个住宅样本点, 但少量住宅由于建筑年代久远无绿化率的相关数据, 本文通过谷歌地球影像和图像色彩识别技术对缺失绿化率进行了近似识别, 填补了缺失数据。

表 3 房价影响因素指标

| 指标类型 | 指标名称 | 标签 | 指标类型 | 指标名称 | 标签 |
|-------|-------------|-------|--------|----------|----------|
| 基本属性 | 单价 | X_2 | 医疗便利性 | 最近三甲医院距离 | X_7 |
| | 面积 | X_3 | | 最近药店距离 | X_8 |
| | 楼龄 | X_4 | 生活便利性 | 最近购物中心距离 | X_9 |
| 交通便利性 | 最近公交站距离 | X_5 | 所属小区属性 | 容积率 | X_{10} |
| | 800 米内公交站个数 | X_6 | | 绿化率 | X_{11} |
| 因变量 | 总价 | Y | | | |

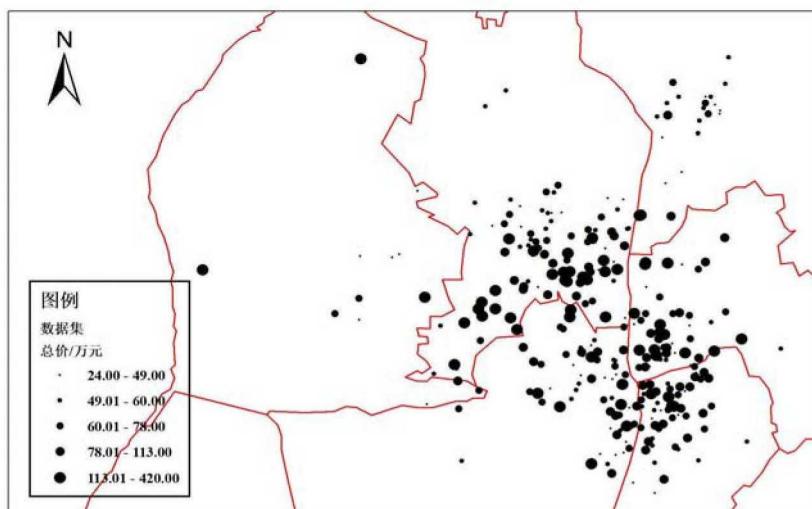


图 2 房价空间分布图

建立异窗宽 GTWR 模型需要确保数据存在着空间自相关性, 首先对乌鲁木齐市商品住宅价格进行空间自相关的检验。按分析角度的不同, 空间自相关检验为全局莫兰指数检验和局部莫兰指数检验 (Local Moran's I), 其原假设认为数据不存在着空间自相关性, 莫兰指数为正表示空间正自相关, 为负时表示空间负自相关。检验结果如表 4 所示。

表 4 空间自相关检验结果

| 全局莫兰指数 | Z 得分 | P 值 | 局部莫兰指数 | Z 得分 | P 值 |
|--------|--------|-------|--------|-------|-------|
| 0.763 | 60.004 | 0.000 | 0.373 | 2.677 | 0.007 |

表 4 的检验结果中全局和局部莫兰指数的数值都大于零, 表明存在着全局的空间正相关与局部的高 - 高或低 - 低聚集。其 P 值均小于显著性水平 0.05, 拒绝了数据不存在空间自相

关性的原假设。因此从全局角度来看, 乌鲁木齐市商品住宅价格在空间上具有正自相关性, 从局部角度来看这种相关性存在着集聚的现象, 符合图 2 所反映的分布情况, 可以进一步建立时空地理加权回归模型。用 Y_i 表示第 i 个商品房的住宅价格, x_{ij} 表示第 i 个商品住宅的第 j 个影响因素指标值, 由此建立如下异窗宽 GTWR 模型:

$$y_i = \beta_{D0}(u_0, v_0, t_0) + \sum_{j=1}^p \beta_{Dij}(u_i, v_i, t_i)x_{ij} + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

其中 $n = 388$, $j = 11$, (u_i, v_i, t_i) 分别为第 i 个样本经度、维度、建筑年龄时间戳。根据经纬度 - 欧式距离计算公式, 我们可以计算出任意两个样本 a, b 之间的时空距离:

$$d_{ab}^{ST} = 2\lambda R \arcsin \left(\left(\sin^2 \left(\frac{u_1 - u_b}{2} \right) + \cos u_a \cos u_b \sin^2 \left(\frac{v_a - v_b}{2} \right) \right)^{\frac{1}{2}} \right) + \mu(t_a - t_b)^2,$$

其中 $R = 6431.76$ (km) 是地球赤道平均半径, 且 R 的单位能够显著影响回归系数与时空核函数窗宽的数量级。选择指标 X_6 作为建立标准差椭圆的权重项, 且满足结果中更靠近 S_1 中心的样本交通便利性更强。对建立的标准差椭圆对不同区域的样本进行识别, 通过本文给出的算法流程建立异窗宽 GTWR 模型, 其中样本的标准差椭圆如图 3 所示。

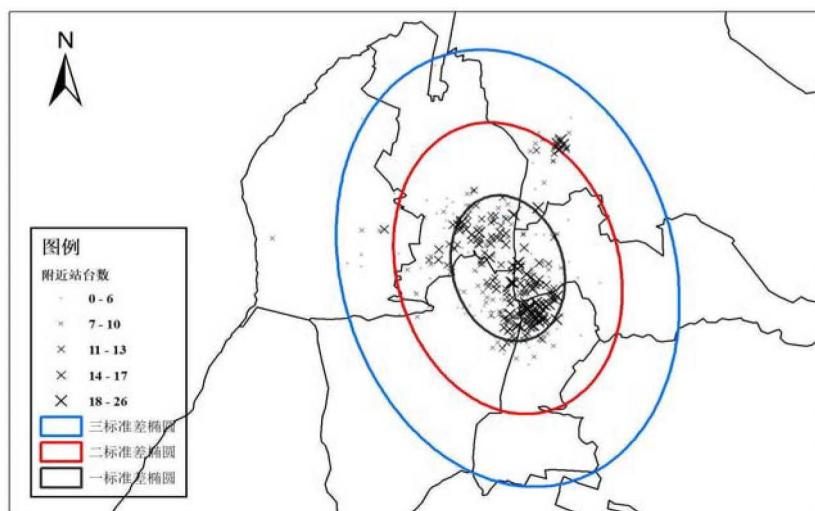


图 3 样本标准差椭圆分布图

图 3 所显示的标准差椭圆是按照数据的空间分布趋势进行旋转后的结果, 椭圆的方向与乌鲁木齐市的城市南北空间布局一致并将数据集划分为四个子区域。处于一标准差椭圆内的样本较为聚集且多处于城市中心, 拥有更高的交通便利性, 而位于三标准差椭圆外的样本会被认为是拥有较低的交通便利性, 因此它们只需要考虑附近极少数样本的影响。以 Bi-square 核函数为例, 本文数据集的样本量为 388 且时空距离需要样本进行两两计算, 一标准差内的样本最大考虑 66.7% 的邻近点, 理论计算量为 258^2 , 二标准差内的样本最大考虑 28.3% 的邻近点, 理论计算量为 110^2 , 三标准差内样本理论计算量为 15^2 , 此时权重的理论总计算量为 78670 相较于同窗宽高斯核函数的 150544 减少了一半。在拟合 Multi-GTWR 模型之前为消除多重共线性对模型的影响, 首先对指标进行主成分降维, 结果见表 5。

表 5 主成分分析结果

| 主成分 | 特征值 | 方差解释比 | 累积方差解释比 |
|-----|--------|--------|---------|
| P1 | 5.3414 | 0.5628 | 56.28% |
| P2 | 2.6849 | 0.2661 | 82.89% |
| P3 | 1.4840 | 0.0727 | 90.16% |
| P4 | 1.0557 | 0.0663 | 96.79% |
| P5 | 0.8057 | 0.0215 | 98.94% |
| P6 | 0.6778 | 0.0106 | 100.00% |
| P7 | 0.6548 | 0.0000 | 100.00% |
| P8 | 0.6021 | 0.0000 | 100.00% |
| P9 | 0.5245 | 0.0000 | 100.00% |

由主成分分析结果可以看出,前三个主成分的累积方差解释比已经达到了 90% 且特征值大于 1,选择前三项主成分作为拟合 Multi-GTWR 模型的响应变量。接下来进一步按照 AIC_c 准则对模型进行迭代后确定最佳的时空因子比为 $\frac{\lambda}{\mu} = 3.128$ 。由于局部最优思想下的加权最小二乘法使得每个样本点都存在着一个回归方程,因此回归系数的总数达到了 4×388 个,本文仅给出回归系数的描述统计,如表 6 所示。

表 6 Multi-GTWR 回归系数描述统计

| 指标 | C | P1 | P2 | P3 |
|---------|----------|----------|----------|-----------|
| 均值 | 16.93682 | -0.00192 | 0.015736 | 5.65E-06 |
| 标准差 | 0.256326 | 3.47E-05 | 2.87E-05 | 1.53E-05 |
| 最小值 | 16.04267 | -0.00216 | 0.015571 | -2.83E-05 |
| 25% 分位数 | 16.75916 | -0.00195 | 0.015717 | -3.46E-06 |
| 50% 分位数 | 16.97205 | -0.00193 | 0.015733 | 4.62E-06 |
| 75% 分位数 | 17.11877 | -0.00190 | 0.015752 | 1.24E-05 |
| 最大值 | 18.44732 | -0.00181 | 0.015829 | 7.68E-05 |

为比较不同模型的性能和验证本文所提出模型的优越性,本文将异窗宽 GTWR 模型与同窗宽 GTWR 模型、GWR 模型、多元线性回归模型的各项指标进行了比较,结果如表 7 所示。

表 7 模型性能对比

| 模型 | MSE | AIC _c | AdjR ² |
|------------|-----------|------------------|-------------------|
| Multi-GTWR | 979.87 | 2732.02 | 0.8988 |
| GTWR | 1557.13 | 2792.83 | 0.8474 |
| GWR | 189305.39 | 2611.84 | 0.71182 |
| OLS | 200955.16 | 2619.06 | 0.68240 |

比较结果后不难得出,Multi-GTWR 模型相较于同窗宽 GTWR、GWR 模型、多元线性回归模型在拟合优度、残差平方和、AIC_c 值上都有较大的提升。表明传统的 GWR 模型只强调了空间距离对权重的影响,采用同窗宽时空权重矩阵的 GTWR 与 GWR 模型考虑了回归点周围所有样本的影响后仍没有达到较好的拟合效果也说明在全局上使用相同的窗宽值存在着计算冗余,而本文中根据研究区域实际情况建立标准差椭圆来重新衡量空间距离概念,使模型的解释能力大大提升。Multi-GTWR 模型解释了房价数据 89% 以上的变动情况,相比同窗宽 GTWR 模型很好的提高了模型性能。

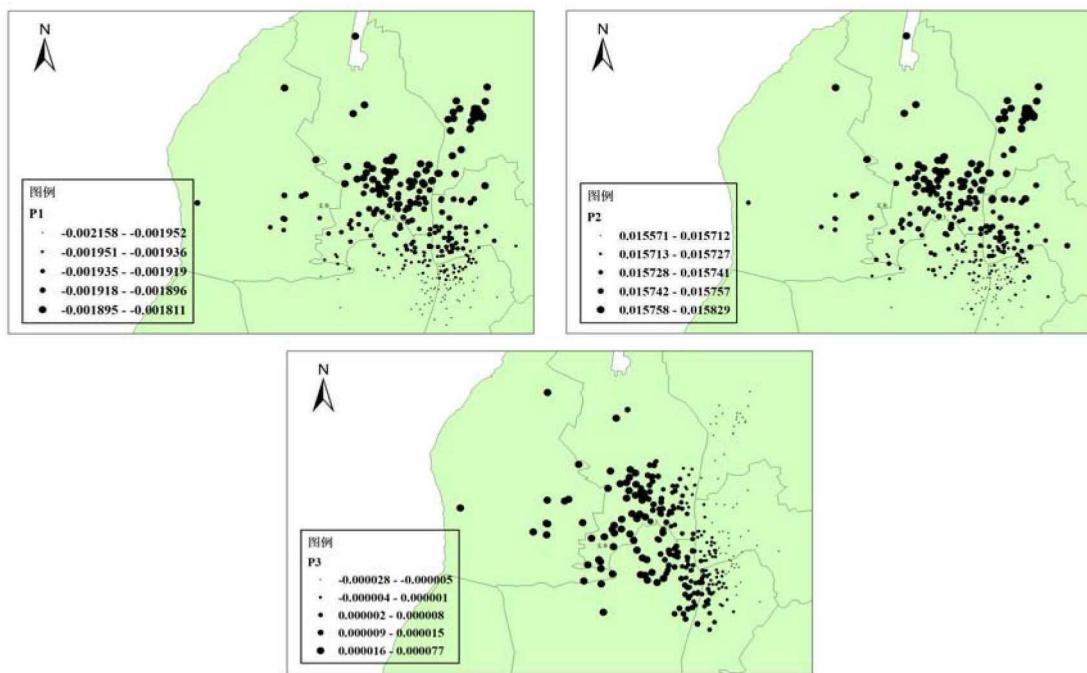


图 4 回归系数分布

从可视化结果中可以得出, Multi-GTWR 模型在考虑了样本的时间因素后, 回归系数的分布变得更符合实际情况, 在建筑年龄较大的区域回归系数会更小, 建筑年龄较小的区域回归系数会更大, 表明在相同的空间属性下房价的变动情况与建筑年龄呈空间正相关性, 房屋的建筑年龄能够显著的影响价格变动。在不同的主成分上, Multi-GTWR 模型回归系数的时空分异性更强, 相比较 GTWR 模型能够更好反映不同区位上房价的变化情况, 商品住宅价格中的空间和时间因素都得到了较好的拟合。

4 结论

本文就干旱区城市乌鲁木齐市的商品住宅价格, 在同窗宽 GTWR 模型的基础上, 考虑干旱区城市的样本潜在差异, 提出异窗宽时空权重矩阵, 通过标准差椭圆算法给出了异窗宽的计算方法, 进而得到了异窗宽时空地理加权回归模型 (Multi-Bandwidth GWTR), 且该模型相比较经典的 GTWR 模型显著减少了计算量。对商品住宅价格影响因素的实证分析结果中, 模型性能较以往的计量模型有了较大提升, 从回归系数空间分布可视化的对比结果中可以得到, 异窗宽时空地理加权回归模型回归系数的符号同实际预期保持一致, GWR 模型对房价的空间分异性的解释能力较为单一, 仅能给出房价的南北或东西空间自相关性, 而异窗宽 GTWR 则模型充分反映了不同区域指标受时间因素影响的分异情况, 建筑年龄对回归系数的变化情况产生了显著的影响。结果表明, 干旱区商品住宅价格的波动在空间上存在显著的南北分异性, 即房价的变动存在着时空正相关性, 但受住宅基本属性影响较小, 而受外在因素影响较大, 交通便利性、医疗保便利性、商品住宅所属小区的属性空间分异性相对较强, 基于地理信息系统的可视化对比分析也显示出异窗宽 GTWR 模型相较于同窗宽 GTWR 模型和 GWR 模型的合理性与优越性。

[参考文献]

- [1] 陈艇, 张志斌. 兰州市商品住房价格空间分布格局及其影响因素 [J]. 干旱区资源与环境, 2015, 29(12): 44–50.
- [2] 王新刚, 孔云峰. 基于时空窗口改进的时空加权回归分析 —— 以湖北省黄石市住房价格为例 [J]. 地理科学, 2015, 35(5): 615–621.
- [3] 孙倩, 汤放华. 基于空间扩展模型和地理加权回归模型的城市住房价格空间分异比较 [J]. 地理研究, 2015, 34(7): 1343–1351.
- [4] 石振武, 王喆. 基于分位数的混合地理加权回归模型的商品住宅价格空间分析 —— 以哈尔滨市为例 [J]. 土木工程与管理学报, 2018, 35(5): 28–33.
- [5] 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.
- [6] 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.
- [7] 苏方林. 地级市 R&D 知识溢出的 GWR 实证分析 [J]. 数理统计与管理, 2010, 29(1): 41–51.
- [8] 李斌, 卢明炜, 张所地, 范新英. 房地产业对中国城市金融稳定的影响研究 —— 基于空间计量模型的比较分析 [J]. 数理统计与管理, 2019, 38(2): 343–356.
- [9] Fotheringham A S, Yang W, Kang W. Multiscale geographically weighted regression (MGWR) [J]. Annals of the American Association of Geographers, 2017, 107(6): 1247–1265.
- [10] Wu B, Li R R, Huang B. A geographically and temporally weighted autoregressive model with application to housing prices [J]. International Journal of Geographical Information Science, 2010, 28(5): 1186–1204.
- [11] Fotheringham A S, Crespo R, Yao J. Geographical and temporal weighted regression (GTWR) [J]. Geographical Analysis, 2015, 47(4): 431–452.
- [12] 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.
- [13] Fotheringham A S, Yang W, Kang W. Multiscale geographically weighted regression (MGWR) [J]. Annals of the American Association of Geographers, 2017, 107(6): 1247–1265.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 吕萍, 鄒辉. 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.
- [19] 龙莹. 空间异质性与区域房地产价格波动的差异 —— 基于地理加权回归的实证研究 [J]. 中央财经大学学报, 2010, (11): 80–85.