doi:10. 16055/j. issn. 1672-058X. 2021. 0006. 015

基于半参数空间模型的房地产估值数据研究

张琳琳, 黄振生
(南京理工大学 理学院,南京 210094)

摘　要:针对现有的房地产估值模型中不包含空间自相关性以及非线性影响因素的问题,提出了可以灵活解释变量意义的部分线性空间自回归模型来拟合房地产估值数据;对于部分线性空间自回归模型的估计问题,利用局部多项式方法与拟极大似然估计法相结合的两步估计过程得到参数部分的估计;房地产估值数据的拟合结果表明:房地产估值数据确实存在空间相关性,房屋到最近的捷运站的距离与房价呈负相关关系,而步行生活圈中便利店的数量与房价呈正相关关系,这与现实意义上的解释是相通的,另外房屋年龄与房价之间的非线性关系也被体现出来;部分线性空间自回归模型能更加客观和灵活地解释房地产估值数据的现实意义。

关键词:部分线性空间自回归模型;剖面拟极大似然估计;局部多项式估计;空间相关性
中图分类号:O212.7　　**文献标志码**:A　　**文章编号**:1672-058X(2021)06-0114-04

0　引　言

近几年随着市场经济的不断发展壮大,房地产估值数据越来越受到重视,探寻房价的影响因素及房价与这些影响因素间的相关关系已经成为统计学者的研究热点。很多研究人员对房地产估值数据进行了建模分析,Yeh[1]提出了一种创新的房地产估价方法——定量比较法,并通过房地产估值数据证明了定量比较方法比两种经典的享乐价格方法(多元回归分析和神经网络)更准确。同时,利用排序分位数方法,将房价的影响因素按照重要性从高到低的顺序排序为房屋到最近的捷运站的距离、步行圈内便利店的数量、房屋年龄和交易日期;Huang 和 Lou[2]考虑所观测到的房屋位置变量(纬度和经度)带有测量误差,从而利用带测量误差的单指标模型研究了房屋位置和房屋价格之间的关系;Wang[3]对影响房屋价格的因素进行了系统的分析,并针对近些年来的大众评价模型及方法进行了系统的文献综述;Wilhelmsson[4]在传统的房地产模型中考虑房地产数据的空间维度,采用传统的线性空间自回归模型解决问题,结果表明房价数据间存在空间自相关,而空间维度的加入解释了更多的价格变化。

在空间自回归模型中,部分线性空间自回归模型不仅保持了非参数空间自回归模型的灵活性,而且保持了参数空间自回归模型的解释力,从而在近些年引起了人们的广泛关注。Su 和 Jin[5]首次提出了部分线性空间自回归模型,并利用一种剖面拟极大似然方法来估计未知函数 $m(\cdot)$ 和参数向量 $(\boldsymbol{\beta}^{\mathrm{T}},\rho,\sigma^2)^{\mathrm{T}}$,然后系统地研究了所得估计量的渐近性质。为了考虑误差项的异方差性,Zhang[6],Zhang 和 Yang[7]分别针对部分线性空间自回归模型提出了二元差分估计方法和工具变量估计方法;Li 和 Mei[8-9]提出了广义似然比检验法来检验非参数部分 $m(\cdot)$ 是否呈现一些有趣的参数形式,以及参数向量 $(\boldsymbol{\beta}^{\mathrm{T}},\rho)^{\mathrm{T}}$ 是否满足某些线性约束条件;Wei 和 Guo[10]提出了一种半参数部分线性变系数空间自回归模型,它是标准空间自回归模型和部分线性空间自回归模型的推广。另外,为了估计未知的空间滞后参数、常数系数和系数函数,提出了一种基于局部线性方法的轮廓拟极大似然方法,为了检验空间效应的存在性,提出了一种广义似然比检验统计量,并利用基于残差的自举过程导出了检验的 p 值。

目前大多数对房地产数据研究的文献仍停留在

收稿日期:2020-10-26;**修回日期**:2020-12-10.

作者简介:张琳琳(1997-),女,河北衡水人,硕士研究生,从事非参数统计研究.

传统的参数模型上,但众所周知,空间因素是影响房价的不可或缺的因素。虽然有一些文献[4]在模型中考虑空间自相关性,并基于此对相关变量给出更合理的参数解释,但却没有将自变量的非线性影响考虑在内,而往往解释变量与因变量的关系不是线性的,这就产生了一些局限性。基于上述考虑,针对现有的房地产估值模型中不包含空间自相关性以及非线性影响因素的问题,提出了可以灵活解释变量意义的部分线性空间自回归模型来拟合房地产估值数据,并对拟合结果给出合理解释。

1　数据介绍

房地产估值的市场历史数据集来自台湾新北市新店区。数据集包含 2012-06—2013-05 的 414 个样本,共 6 个变量,变量意义见表 1,相关数据可见文献[1]。首先通过绘制变量间的散点图来判断房价(Y)与各影响因素(X_1,X_2,Z)间的相关关系,其中经度和纬度作为影响房价的位置信息,以空间权重矩阵的形式包含在模型中。从图 1,图 2,图 3 及现实意义中可以看出房价(Y)与房屋到最近的捷运站的距离(X_1)和步行生活圈中便利店的数量(X_2)存在线性关系,而与房屋年龄(Z)存在非线性关系。为了更进一步挖掘出变量数据中的隐含信息,进一步利用部分线性空间自回归模型对数据集进行拟合。

表 1　房地产估值数据变量

Table 1　Real estate valuation data variables

变量名称	变量意义
Y	单位面积的房价
X_1	房屋到最近的捷运站的距离
X_2	步行生活圈中便利店的数量
Z	房屋年龄
D_{LON}	经度
D_{LAT}	纬度

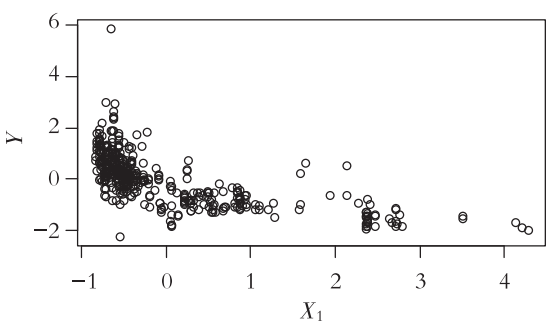

图 1　Y 与 X_1 的散点图

Fig. 1　The scatter plot of Y and X_1

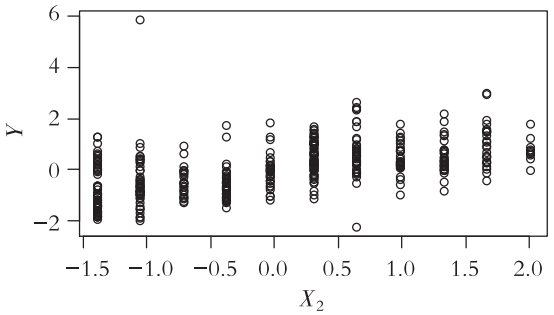

图 2　Y 与 X_2 的散点图

Fig. 2　The scatter plot of Y and X_2

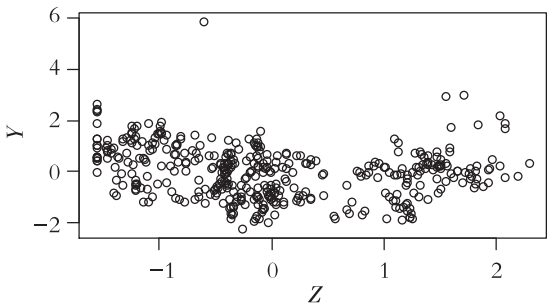

图 3　Y 与 Z 的散点图

Fig. 3　The scatter plot of Y and Z

表 2　变量间的相关系数矩阵

Table 2　Correlation coefficient matrix between variables

	Y	X_1	X_2	Z
Y	1.000 000 0	-0.673 612 9	0.571 004 9	-0.210 567 0
X_1	-0.673 612 9	1.000 000 0	-0.602 519 1	0.025 622 1
X_2	0.571 004 9	-0.602 519 1	1.000 000 0	0.049 592 5
Z	-0.210 567 0	0.025 622 1	0.049 592 5	1.000 000 0

2　模型分析

部分线性空间自回归模型具有形式:

$$\boldsymbol{Y}_n=\boldsymbol{X}_n\boldsymbol{\beta}_0+\boldsymbol{m}_0(\boldsymbol{Z}_n)+\rho_0\boldsymbol{W}_n\boldsymbol{Y}_n+\boldsymbol{U}_n \tag{1}$$

其中 $\boldsymbol{X}_n\equiv(x_{n,1},\cdots,x_{n,n})^{\mathrm{T}}$ 为 p 维协变量,$\boldsymbol{Z}_n\equiv(z_{n,1},\cdots,z_{n,n})^{\mathrm{T}}$ 为 q 维协变量,$\boldsymbol{W}_n$ 是一个指定的 $n\times n$ 空间权重矩阵,$\boldsymbol{U}_n\equiv(u_1,\cdots,u_n)^{\mathrm{T}}$ 是具有零均值和有限方差 σ_0^2 的独立同分布扰动项的 n 维向量,$\boldsymbol{m}_0(\boldsymbol{Z}_n)\equiv(m_0(z_{n,1}),..,m_0(z_{n,n}))^{\mathrm{T}}$,并且 $\boldsymbol{m}_0(\cdot)$是一个定义在 R^q 上的未知函数。令 $\boldsymbol{\theta}_0=(\boldsymbol{\beta}_0^{\mathrm{T}},\rho_0,\sigma_0^2)^{\mathrm{T}}$ 是实值参数向量,$\boldsymbol{U}_n(\boldsymbol{\delta})=\boldsymbol{Y}_n-\boldsymbol{X}_n\boldsymbol{\beta}-\boldsymbol{m}_0(\boldsymbol{Z}_n)-\rho\boldsymbol{W}_n\boldsymbol{Y}_n$,$\boldsymbol{T}_n(\rho)=\boldsymbol{I}_n-\rho\boldsymbol{W}_n$,$\boldsymbol{Y}_n=\boldsymbol{T}_n^{-1}(\boldsymbol{X}_n\boldsymbol{\beta}_0+\boldsymbol{m}_0(\boldsymbol{Z}_n)+\boldsymbol{U}_n)$。通过以下两步来对 θ 进行估计:(1) 基于给定的 $\boldsymbol{\theta}$ 对 $\boldsymbol{m}_0(z)$进行估计,记为 $\boldsymbol{m}_{\boldsymbol{\theta}}(z)$;(2) 将所得到的 $\boldsymbol{m}_{\boldsymbol{\theta}}(z)$带入 $\boldsymbol{U}_n(\boldsymbol{\delta})$中,并分别得到 $\boldsymbol{\theta}$ 和 $\boldsymbol{m}_0(z)$的剖面拟极大似然估计(QMLE)$\hat{\theta}$,$\boldsymbol{m}_{\hat{\boldsymbol{\theta}}}(z)$。

首先利用局部线性估计[11]给出非参数函数 $\boldsymbol{m}_0(z)$ 的估计。令 $\boldsymbol{K}(\cdot)$ 表示定义在 R^q 上的核函数，$h=h_n$ 是带宽，定义 $K_h(z)=h^{-q}K(z/h)$，$\boldsymbol{Y}_n^*(\rho)=\boldsymbol{T}_n(\rho)\boldsymbol{Y}_n$，则对于给定的 θ，$\boldsymbol{m}_0(z)$ 的剖面局部线性估计量 $\boldsymbol{m}_\theta(z)$ 为

$$\boldsymbol{m}_\theta(z)=\boldsymbol{s}(z)'(\boldsymbol{Y}_n^*(\rho)-\boldsymbol{X}_n\boldsymbol{\beta}) \tag{2}$$

其中 $\boldsymbol{s}(z)^{\mathrm{T}}=\boldsymbol{e}^{\mathrm{T}}[\vec{\boldsymbol{Z}}(z)^{\mathrm{T}}\boldsymbol{K}_h(z)\vec{\boldsymbol{Z}}(z)]^{-1}\vec{\boldsymbol{Z}}(z)^{\mathrm{T}}\boldsymbol{K}_h(z)$，$\boldsymbol{e}=(1,0,\cdots,0)^{\mathrm{T}}$ 是一个 $N\times1$ 的单位向量，$\boldsymbol{K}_h(z)=\mathrm{diag}(K_h(z_{n,1}-z),\cdots,K_h(z_{n,n}-z))$，$\vec{\boldsymbol{Z}}(z)=(\vec{Z}_1(z),\cdots,\vec{Z}_n(z))^{\mathrm{T}}$，$\vec{\boldsymbol{Z}}_i(z)=(1,(z_{n,i}-z)^{\mathrm{T}}/h)^{\mathrm{T}}$。

在第二步中，考虑最大化下列似然函数

$$\log\boldsymbol{L}_n(\boldsymbol{\theta},\boldsymbol{m}_\theta(\boldsymbol{Z}_n))=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\sigma^2+\log|\boldsymbol{T}_n(\rho)|-\frac{1}{2\sigma^2}(\boldsymbol{T}_n(\rho)\boldsymbol{Y}_n-\boldsymbol{X}_n\boldsymbol{\beta}-\mathbf{m}_\theta(\boldsymbol{Z}_n))^{\mathrm{T}}(\boldsymbol{T}_n(\rho)\boldsymbol{Y}_n-\boldsymbol{X}_n\boldsymbol{\beta}-\boldsymbol{m}_\theta(\boldsymbol{Z}_n))$$

其中 $\boldsymbol{m}_\theta(\boldsymbol{Z}_n)=(m_\theta(z_{n,1}),\cdots,m_\theta(z_{n,n}))^{\mathrm{T}}$，令 $\boldsymbol{S}_n=(s(z_{n,1}),\cdots,s(z_{n,n}))$，在拟对数似然函数中，对于给定的 ρ，β 的 QMLE 估计为

$$\hat{\boldsymbol{\beta}}(\rho)=\left[\boldsymbol{X}_n^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{S}_n)^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{S}_n)\boldsymbol{X}_n\right]^{-1}\times\boldsymbol{X}_n^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{S}_n)^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{S}_n)\boldsymbol{T}_n(\rho)\boldsymbol{Y}_n$$

σ^2 的 QMLE 估计为

$$\hat{\sigma}^2(\rho)=\frac{1}{n}Y_n^{\mathrm{T}}T_n(\rho)^{\mathrm{T}}(I_n-S_n)^{\mathrm{T}}M_n(I_n-S_n)T_n(\rho)Y_n$$

令 $P_n=(I_n-S_n)X_n$，$M_n=I_n-P_n[P_n^TP_n]^{-1}P_n^T$，关于 ρ 的集中对数似然函数为

$$\log\boldsymbol{L}_n^c(\rho)=-\frac{n}{2}(\log(2\pi)+1)-\frac{n}{2}\log\hat{\sigma}^2(\rho)+\log|\boldsymbol{T}_n(\rho)|$$

最大化上式可以得到 ρ 的 QMLE 估计 $\hat{\rho}$，以及 $\hat{\boldsymbol{\beta}}(\hat{\rho})$，$\hat{\sigma}^2(\hat{\rho})$，$\boldsymbol{m}_{\hat{\theta}}(z)$。

3 模型拟合

采用台湾新北市新店区 2012-06—2013-05 的房地产估值数据，共有 414 个样本，6 个变量。利用变量中的经度($\boldsymbol{D}_{\mathrm{LON}}$)和纬度($\boldsymbol{D}_{\mathrm{LAT}}$)来建立空间权重矩阵，将数据应用于部分线性空间自回归模型 $\boldsymbol{Y}=\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{X}_2\boldsymbol{\beta}_2+\boldsymbol{m}(\boldsymbol{Z})+\rho\boldsymbol{W}_n\boldsymbol{Y}+\boldsymbol{U}$ 中，并对指标变量进行标准化处理。令 $\boldsymbol{W}_n=(w_{ij})$，$w_{ij}=\max(1-d_{ij}/d_0,0)$，其中 d_{ij} 为欧几里得距离，选择阈值距离 d_0 为 0.05，并对空间权重矩阵 $\boldsymbol{W}_n$ 进行标准化，使其行和为 1。此外利用 EPANECHNIKOV 核函数，通过大拇指法则选取带宽 $h=0.317\ 6$ 进行估计。部分线性空间自回归模型指标参数的估计值见表 3。

表 3 未知参数估计值

Table 3 The estimators of unknown parameters

$\hat{\rho}$	$\hat{\beta}_1$	$\hat{\beta}_2$
−0.149 8	−0.474 3	0.288 9

未知函数的拟合结果如图 4 所示。

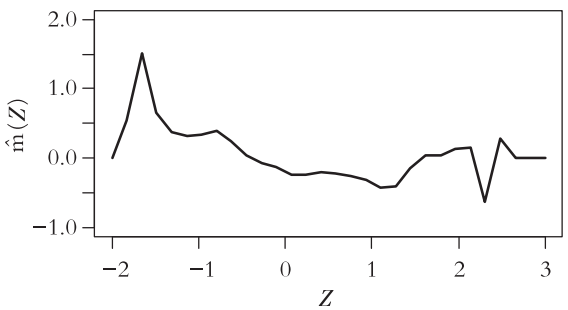

图 4 $m(Z)$ 的估计

Fig. 4 The estimation of $m(Z)$

从表 3 可以看出，单位面积房价(Y)指标之间存在负空间相关性，房屋到最近的捷运站的距离(X_1)与房价(Y)呈负相关关系，即离最近的捷运站距离越近，房价越高；而步行生活圈中便利店的数量(X_2)与房价(Y)呈正相关关系，便利店数量越多，房价越高；这与实际意义相符，同时也证明模型具有现实意义。从图 4 可以看出房价(Y)与房屋年龄(Z)之间存在非线性关系，部分线性空间自回归模型可以较好地拟合房地产估价数据。

4 结 论

采用部分线性空间自回归模型对房地产估值数据进行统计研究，与之前文献研究所不同的是，加入了空间和非线性因素的影响，对于部分线性空间自回归模型的估计问题，首先利用局部多项式估计法对非参数部分进行估计，从而得到含参数形式的非参函数估计，再将此估计表达式代入对数似然函数中，通过极小化对数似然函数得到参数部分的估计。在房地产估值数据的拟合过程中，不同于之前文献中计算空间相关矩阵的方法，利用经度、纬度计算房屋之间的欧几里得距离，从而将空间位置因素以空间相关矩阵的形式引入模型中。结果表明，房价数据之间存在着空间相关性，房屋到最近的捷运站的距离与房价呈负相关关系，而步行生活圈中便利店的数量与房价呈正相关关系，这与现实意义上的解释是相通的。另外房屋年龄与房价之间的非线性关系也被体现出来。部分线性空间自回归模型能更加客观和灵活地解释房地产估值数据的现实意义。

参考文献(References):

[1] YEH I C, HSU T K. Building Real Estate Valuation Models with Comparative Approach through Case-Based Reasoning[J]. Applied Soft Computing, 2018, 65(10): 260—271

[2] HUANG Z, LOU W. Statistical Inferences for Single-Index Models with Measurement Errors[J]. Journal of Applied Statistics, 2020:1-20

[3] WANG D, LI V J. Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review[J]. Sustainability, 2019, 11(24): 7006

[4] WILHELMSSON M. Spatial Models in Real Estate Economics[J]. Housing, Theory and Society, 2002, 19(2): 92—101

[5] SU L, JIN S. Profile Quasi-Maximum Likelihood Estimation of Partially Linear Spatial Autoregressive Models[J]. Journal of Econometrics, 2010, 157(1): 18—33

[6] ZHANG Z. A Pairwise Difference Estimator for Partially Linear Spatial Autoregressive Models[J]. Spatial Economic Analysis, 2013, 8(2): 176—194

[7] ZHANG Y, YANG G. Statistical Inference of Partially Specified Spatial Autoregressive Model[J]. Acta Mathematicae Applicatae Sinica, English Series, 2015, 31(1): 1—16

[8] LI T, MEI C. Testing a Polynomial Relationship of the Non-Parametric Component in Partially Linear Spatial Autoregressive Models[J]. Papers in Regional Science, 2013, 92(3): 633—649

[9] LI T, MEI C. Statistical Inference on the Parametric Component in Partially Linear Spatial Autoregressive Models[J]. Communications in Statistics-Simulation and Computation, 2016, 45(6): 1991—2006

[10] WEI C, GUO S, ZHAI S. Statistical Inference of Partially Linear Varying Coefficient Spatial Autoregressive Models[J]. Economic Modelling, 2017, 64(8): 553—559

[11] 薛留根. 现代统计模型[M]. 北京: 科学出版社,2012
XUE L G. Modern Statistical Model[M]. Beijing: Science Press,2012 (in Chinese)

Study on Real Estate Valuation Data Based on Semi-parametric Spatial Model

ZHANG Lin-lin, HUANG Zhen-sheng

(School of Science, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: In view of the problem that the existing real estate valuation models do not include spatial autocorrelation and nonlinear influencing factors, a partially linear spatial autoregressive model that can flexibly explain the meaning of variables is proposed to fit real estate valuation data. For the estimation of partial linear spatial autoregressive models, the two-step estimation process combining the local polynomial method and the quasi-maximum likelihood estimation method is used to obtain the estimation of the parameter part. The fitting result of real estate valuation data shows that real estate valuation data does have spatial correlation, and the distance between the house and the nearest MRT station is negatively correlated with the housing price, while there is a positive correlation between the number of convenience stores in the living circle on foot and the housing price, which is similar to the practical explanation. In addition, the nonlinear relationship between the age of house and the housing price is also reflected. Partially linear spatial autoregressive model can explain the practical significance of real estate valuation data more objectively and flexibly.

Key words: partially linear spatial autoregressive model; profile quasi-maximum likelihood estimation; local polynomial estimation; spatial correlation

责任编辑:田　静

引用本文/Cite this paper:

张琳琳,黄振生. 基于半参数空间模型的房地产估值数据研究[J]. 重庆工商大学学报(自然科学版),2021,38(6):114—117

ZHANG L L, HUANG Z S. Study on Real Estate Valuation Data Based on Semi-parametric Spatial Model[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2021, 38(6):114—117