

DOI: 10.11830/ISSN.1000-5013.202009029



# 特征房价空间分析及连续型 深度置信网络预测

吴莞姝<sup>1</sup>, 胡龙超<sup>2</sup>, 赵凯<sup>2</sup>

(1. 华侨大学 建筑学院, 福建 厦门 361021;  
2. 华侨大学 数量经济研究院, 福建 厦门 361021)

**摘要:** 以上海为研究区域,利用数据爬虫手段搜集、整理上海市二手房交易数据,通过空间自相关分析二手房交易价格的空间效应,并使用连续型深度置信网络对二手房交易价格进行分析预测.研究表明:上海市二手房交易价格在空间上具有显著的自相关效应,在上海市核心区域存在高-高集聚效应,在周边区域呈现低-低集聚效应,而在核心与周边交界地区存在高-低集聚和低-高集聚的负向空间效应;特征变量对价格偏高区域的二手房交易价格解释力度较小;除中心区域外,基于连续型深度置信网络的特征变量对上海市二手房交易价格预测能力良好.

**关键词:** 连续型深度置信网络; 建筑特征; 区位特征; 邻里特征; 空间自相关; 上海市

**中图分类号:** TP 18; F 293.35 **文献标志码:** A **文章编号:** 1000-5013(2021)04-0537-10

## Spatial Analysis of Characteristics Housing Price and Prediction With Continuous Deep Belief Neural Network

WU Wanshu<sup>1</sup>, HU Longchao<sup>2</sup>, ZHAO Kai<sup>2</sup>

(1. School of Architecture, Huaqiao University, Xiamen 361021, China;  
2. Institute for Quantitative Economics, Huaqiao University, Xiamen 361021, China)

**Abstract:** Taking Shanghai City as the research area, data crawlers are used to collect and organize the second-hand housing transaction data, the spatial effect of second-hand house prices are analyzed through spatial autocorrelation, and the continuous deep belief neural networks are used to analyze and predict the second-hand housing prices. Research results show that the transaction price of second-hand housing in Shanghai City has a significant spatial autocorrelation effect. There is a high-high agglomeration effect in the core area of Shanghai City, and a low-low agglomeration effect in the surrounding areas. There are negative spatial effects of high-low aggregation and low-high aggregation at the junction of the core and surrounding areas. Characteristic variables have less power to explain the transaction prices of second-hand housing in areas with high prices. Except for the core areas, the characteristic variables based on the continuous deep belief neural network have a good ability to predict the transaction price of second-hand housing in Shanghai City.

**Keywords:** continuous deep belief neural network; architectural characteristics; location characteristics; neighborhood characteristics; spatial autocorrelation; Shanghai City

**收稿日期:** 2020-09-14

**通信作者:** 吴莞姝(1988-),女,讲师,博士,主要从事城市规划、大数据与GIS的研究. E-mail: wuwanshu131@163.com.

**基金项目:** 国家自然科学基金资助项目(51908229, 71603087);福建省自然科学基金面上资助项目(2019J01063);华侨大学青年教师科技创新资助计划(ZQN-816)

房地产行业在国民经济中有着举足轻重的地位,作为房地产价值体现的房价不仅关系到国民经济的健康、平稳发展,其涨跌变动还关系到居民的财富及生活水平<sup>[1-2]</sup>. 影响房价的因素众多,不仅受建筑特征的影响,还会因地区之间公共服务设施的资源配置不均衡而产生差异. 人们根据对公共服务和设施的偏好选择居住区域,特别是在物质生活水平日渐丰富的经济社会环境下,居民往往愿意支付较高的价格以获得优质的公共服务,而这部分优质公共资源的价值就资本化于房价之中<sup>[3]</sup>.

针对房价的研究多基于特征价格模型. Ridker 等<sup>[4]</sup>采用特征价格法分析空气污染对房价的影响. Xiao 等<sup>[5]</sup>利用特征向量空间滤波方法消除空间自相关性后,发现北京周围的设施对北京房价的影响参差不齐. 张骥<sup>[6]</sup>以北京市二手房市场上的商品住宅和非商品住宅为研究对象,利用基于特征价格的配对回归模型,研究北京学区房交易价格,发现北京市的学区房溢价已高出 24.3%. 文献[7-8]以北京为例,通过特征价格模型探寻房价影响因素,研究证实地铁、公交等交通基础设施及优质教育资源、高水平的医疗机构、公园等公共服务设施对房价上涨皆具有明显的正效应. Li 等<sup>[9]</sup>则通过整合链家网站、Mobike 网站及百度地图兴趣点的公开数据,分析上海公寓价格的空间模式及其与当地配套属性的关联,研究发现公园、学校、医院和银行等公共服务设施及娱乐、购物等私人服务设施推高了市中心地区的房价.

利用特征价格模型构建房价分析模型需要“先验”地设定函数形式,这往往容易损失房价与其特征变量之间的深层次关系. 近年来,国内外学者尝试应用多种机器学习模型探讨房价的变化趋势和影响因素等问题. 申瑞娜等<sup>[10]</sup>结合主成分分析和支持向量机,综合考察影响上海住房价格的 8 种因素,并对上海房价进行预测. 文献[11-12]利用灰色 GM(1,1)预测模型分别对福州市和周口市的房价走势进行预测,并得到精度度较高的预测结果. 张智鹏等<sup>[13]</sup>利用梯度提升树(GBDT)算法对房价进行预测,实验结果表明,公共设施、生活服务、学校、购物服务等是对房价产生明显影响的因素. 这些文献均采用结构较为简单的机器学习模型,并且分析数据的特征维度偏低. 传统机器学习方法难以全面且精确地挖掘特征因素和房价之间的联系.

浅层 BP 神经网络模型(BPNN)在预测上优于传统机器学习模型,但仍存在学习速度慢、易陷入局部收敛等问题. 而深度置信网络采用无监督训练方式,具有较好的降维性能. 一方面,深度置信网络能有效克服传统人工神经网络需要大量有监督信号和易陷入局部极小等缺点;另一方面,深度置信网络可高效处理高维的数据并挖掘变量之间的深层关系. 此外,深度置信网络解决了大规模数据计算耗时问题且精度较高,并成功应用于多种人工智能问题的研究,尤其在图像处理、声音辨识和智能网络分析等方面的应用中成效显著<sup>[14-15]</sup>. 由于深度置信网络在运算时使用数据离散化方法进行特征提取,隐藏层和可见层节点均为伯努利值(0 或 1),这使深度置信网络不适用于对连续型变量的高精度预测<sup>[16-17]</sup>. 为提高模型的预测精度,学者们将深度置信网络进行改进,使其能够有效处理连续型的输入变量<sup>[18-19]</sup>.

尽管深度置信网络在人工智能领域特别是模式识别任务中取得了较好的成果,但将其应用于现实经济问题的研究仍较少见. 基于此,本文尝试将连续型深度置信网络扩展至房价问题的研究中,依据特征房价理论并考虑到上海市二手房交易价格可能存在的空间相关性,构建空间计量模型以分析各特征变量对二手房交易价格的影响. 在此基础上,利用连续型深度置信网络建立房价与多维影响因素之间的深度学习预测模型,深层挖掘其潜在规律.

# 1 研究介绍及数据探索

## 1.1 研究区域界定

由于上海是我国“超一线”城市,房地产市场发展较为成熟和完善,具有一定的代表性;同时,上海浦东和浦西在城市化建设和房屋价格上具有明显的差异,这为探讨建筑特征、区位特征和邻里特征对二手房交易价格的影响效果提供了较好的素材,因此,选择上海市作为研究区域.

## 1.2 数据来源

房屋交易数据源于“链家二手房交易平台”(https://m.lianjia.com),链家的楼盘数据库管理着 160 多个城市 1.1 亿套真实的房产数据,依托互联网对数据进行标准化管理,实现信息的无差别共享.

基于 Python 语言的爬虫技术,按照不同行政区对链家上海市二手房交易平台上的数据进行收集. 利用 Beautiful Soup 对网页返回结果进行重构,得到超文本标记语言(HTML)的树状结构,再使用正则

表达式对所需信息进行提取,进而获取变量数据.最终,所爬取的数据涉及房屋交易额、交易单价、百度地理坐标(BD08)、房屋户型、所在楼层、建筑面积、户型结构、建筑类型、建成年代、装修情况、梯户情况等变量信息,共计 45 131 条原始信息.同时,搜集整理上海市全部 40 家三甲医院(及其分院)、上海市 34 所重点中小学及上海市所有地铁站出口的地理坐标数据.上海市的主要三甲医院及重点中小学的地理位置信息,如表 1 所示.

表 1 上海市的主要三甲医院及重点中小学的地理位置信息  
Tab.1 Geographical location information of 3A grade hospitals  
and key primary and secondary schools in Shanghai City

三甲医院			重点中小学		
名称	经度	纬度	名称	经度	纬度
上海市第六人民医院	121.422 469 7	31.179 077 22	上海世界外国语小学	121.418 047 9	31.150 941 73
上海市同济医院	121.431 119 6	31.266 642 94	上海实验学校	88.889 645 4	29.234 825 55
上海交通大学医学院 附属仁济医院北院	121.551 264 8	31.232 570 46	上海市第一师范 附属小学	121.438 955 9	31.225 738 96
上海市东方医院南院	121.512 785 8	31.148 284 10	静安区第一中心小学	121.449 001 2	31.230 231 46
复旦大学附属华山医院	121.443 540 2	31.216 425 81	明珠小学 A 区	121.518 584 2	31.226 320 58
上海市第一人民医院	121.489 342 4	31.253 388 28	明珠小学 B 区	121.524 537 1	31.223 813 82
上海市东方医院	121.512 268 3	31.237 665 20	明珠小学 C 区	110.981 519 9	36.169 960 69

1.3 数据清洗

上海市各行政区的原始数据交易时间跨度不一,为剔除时间跨度的影响,对各行政区数据进行切割,保留共同交易时间跨度的数据为研究样本.另外,由于在房屋的交易权属中个税的收取不同,因而,删除交易权属中的售后公房.为避免极端价格对数据分析的影响,房屋用途中删除别墅、车库和商业办公类房屋,只保留普通住宅.剔除含有缺失值的数据条目,最终获得 9 058 个样本,涉及房屋户型、建筑面积、建成年代、所在楼层、装修情况、配备电梯等变量.

房屋户型变量的形式以字符数字组合为主,采用正则表达式将卧室、大厅、厨房和卫生间这几个数值提取出来,并分别作为建筑特征的变量.建筑面积的原始数据中带有面积单位 m<sup>2</sup>,利用正则表达式剔除该单位,并把面积值变为浮点型数据.原始数据中建成年代为房屋建成年代,采用爬取数据的年份(2019 年)减去建成年代的方式,计算房屋建成年数.所在楼层的原始数据为高楼层、中楼层和低楼层.对所在楼层进行数值化处理,把高楼层、中楼层和低楼层分别赋值为 2,1 和 0.装修情况和配备电梯为二值数据,装修情况为已装修或未装修,配备电梯为有或无,利用 1 和 0 进行数值化处理.

链家官网的坐标数据来源于百度地图,利用 ArcGIS 软件将清洗后样本数据的地理坐标转化为 WGS84 坐标,其分布情况如图 1 所示.

1.4 特征变量的选取

借鉴以往研究,将二手房的特征梳理为建筑特征、区位特征和邻里特征 3 类.二手房建筑特征为房屋本身的属性,涉及的变量包括房屋户型、所在楼层、建筑面积、建成年代、装修情况及配备电梯.区位特征量化了二手房区位对整个城市的可达性,如出行成本等.

将二手房到城市中心的距离作为二手房的区位特征变量;以陆家嘴金融贸易中心区域的质心为城市中心点.邻里特征通常指房屋周围的环境及配套,如交通站点、学校、医院等.选取到最近三甲医院的距离、到最近重点中小学的距离及到最近地铁站的距离体现邻里特征.距离计算方式取大圆距离,大圆距离是将地球看作一个球形,计算球面上两点的最短路径.特征变量的相关说明,如表 2 所示.

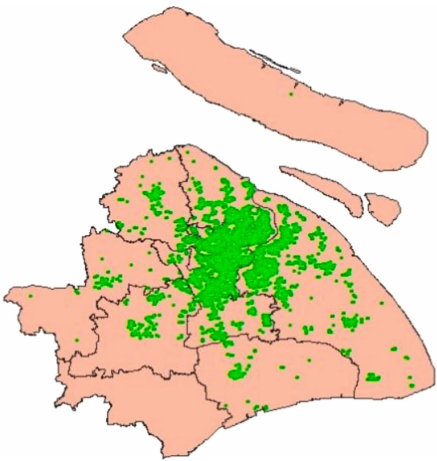


图 1 样本数据的空间分布  
Fig.1 Spatial distribution of sample data

表 2 特征变量说明  
Tab.2 Description of characteristic variables

类别	变量符号	变量名称	变量说明
建筑特征	area	面积	房屋的建筑面积
	room	卧室	房屋所拥有的卧室数量
	hall	客厅	房屋所拥有的客厅数量
	kitchen	厨房	房屋所拥有的厨房数量
	bath	洗手间	房屋所拥有的洗手间数量
	age	楼龄	房屋建筑年限(利用建筑的建成时代计算)
	decoration	装修	房屋是否已装修(否=0,是=1)
	elevator	电梯	房屋是否配备电梯(否=0,是=1)
	floor	楼层	房屋的相对楼层(低楼层=0,中楼层=1,高楼层=2)
	totalfloor	总楼层	建筑的总楼层
区位特征	D_center	距市中心距离	距陆家嘴金融贸易中心区域质心的距离
	D_subway	地铁站	住宅与最近地铁站的大圆距离
邻里特征	D_hospital	三甲医院	住宅与最近三甲医院的大圆距离
	D_school	重点中小学	住宅与最近重点中小学的大圆距离

1.5 描述性统计

清洗后数据的描述性统计,如表 3 所示.由表 3 可知:二手房成交单价最大值为 147 668 元·m<sup>-2</sup>,最小值为 7 059 元·m<sup>-2</sup>,均值为 49 577 元·m<sup>-2</sup>,偏度为 1.011 452,属于右偏数据,大多数交易价格集中在 50 000 元·m<sup>-2</sup>左右.

表 3 数据的描述性统计  
Tab.3 Descriptive statistics of datas

变量符号	变量名称	最小值	最大值	均值	单位
unitprice	交易单价	7 059	147 668	49 577	元·m <sup>-2</sup>
area	面积	17	343	73.67	m <sup>2</sup>
room	卧室	1	6	1.92	个
hall	客厅	0	5	1.32	个
kitchen	厨房	0	2	1	个
bath	洗手间	0	5	1.12	个
age	楼龄	1	107	20.2	年
decoration	装修	0	1	0.90	—
elevator	电梯	0	1	0.38	—
floor	楼层	0	2	1.11	—
totalfloor	总楼层	1	61	10	层
D_center	距市中心距离	0.75	56.81	14.64	km
D_subway	地铁站	0.01	32.86	3.66	km
D_hospital	三甲医院	0.03	48.66	5.56	km
D_school	重点中小学	0.02	52.85	6.75	km

绘制 QQPlot 分布图,将数据分布与正态分布进行对比,结果如图 2 所示.图 2 中:S 为标准正态值;Q 代表数据的分位数.由图 2 可知:对数成交单价数据近似服从于正态分布.

化后的利用空间趋势分析将二手房成交单价投影到 XZ 和 YZ 平面上,绘制上海市二手房交易价格分布的空间趋势,如图 3 所示.图 3 中:X 为经度;Y 为纬度;Z 为单价.由图 3 可知:不论在东西方向还是南北方向,上海市二手房交易单价都呈现由中心向两头递减的趋势.综上可知,上海市二手房市场具有中心高价的特点.

2 特征房价空间分析

2.1 空间自相关

由于上海市二手房交易价格在空间分布上呈现“中心高价、边缘低价”的特点(图 3),故二手房交易

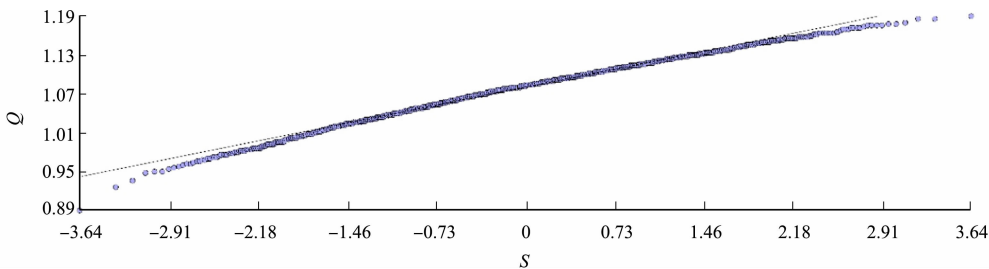


图 2 对数化后的二手房成交单价分布

Fig. 2 Transaction unit price distribution of second-hand housing after logarithm

价格之间很可能具有空间关联性. 利用莫兰指数进行空间自相关检验, 包括建立空间权重矩阵和确定判断指标. 关于建立空间权重矩阵, 如果二手房交易数据为  $\{x_i\}_{i=1}^n$ , 则可定义基于反向距离的空间权重矩阵为

$$W = \begin{pmatrix} 0 & 1/d_{1,2} & \cdots & 1/d_{1,n} \\ 1/d_{2,1} & 0 & \cdots & 1/d_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 1/d_{n,1} & 1/d_{n,2} & \cdots & 0 \end{pmatrix}. \quad (1)$$

式(1)中:  $d_{i,j}$  为样本  $i$  与样本  $j$  之间的距离, 若样本  $i$  与样本  $j$  的平面坐标分别为  $P_i(x_i, y_i)$  和  $P_j(x_j, y_j)$ , 则有  $d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ ;  $n$  为总样本数.

由于二手房交易价格数据在空间上是点要素的形式, 没有多边形的拓扑关系, 在空间上的分布也较不均衡, 故整体的空间关联程度可以利用全局莫兰指数判断, 莫兰指数( $I$ )的表达式为

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n W_{i,j} (P_i - \bar{P})(P_j - \bar{P})}{S^2 \sum_{i=1}^n \sum_{j=1}^n W_{i,j}}. \quad (2)$$

式(2)中:  $P_i$  和  $P_j$  分别为样本  $i, j$  的交易价格;  $\bar{P}$  为交易单价  $P$  的平均值;  $S^2$  为样本交易价格的方差.

计算得到全局空间自相关的检验结果如下: 莫兰指数  $I$  为 0.236 862 3; 统计量为  $-0.000 271$ ;  $p$  近似为 0. 由检验结果可知: 二手房交易价格的莫兰指数显著为正, 说明上海市二手房交易价格具有空间集聚效应.

上海市二手房交易价格全局空间自相关散点图, 如图 4 所示. 通过莫兰散点图将空间自相关分为高-高集聚、高-低集聚、低-高集聚、低-低集聚这 4 种类型. 图 4 中:  $L$  为  $\ln P$  的空间一阶滞后; 第 1, 3 象限是高-高集聚、低-低集聚区域, 即同质化明显的区域; 而第 2, 4 象限是高-低集聚、低-高集聚区域, 即异质性较强的区域.

由图 4 可知: 绝大多数样本落入第 1, 3 象限, 少部分样本落入第 2 象限, 空间集聚特点较为明显.

## 2.2 空间异质性

借助 ArcGIS 软件绘制上海市二手房交易价格的 LISA 集聚状况, 如图 5 所示. 由图 5 可知: 中心城区的房价呈现高-高集聚的空间效应, 且越靠近城市中心点, 高-高集聚的特征越显著; 高-低集聚区域沿着高-高集聚区域的边缘分布; 而低-低集聚效应区域大多分布在上海周边地区.

进一步, 通过局部空间自相关检验探讨分析上海市二手房交易价格的空间异质性. 局部空间自相关

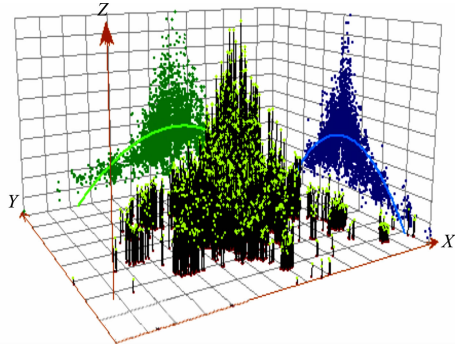


图 3 二手房成交单价的空间趋势

Fig. 3 Spatial trend of transaction unit price of second-hand housing

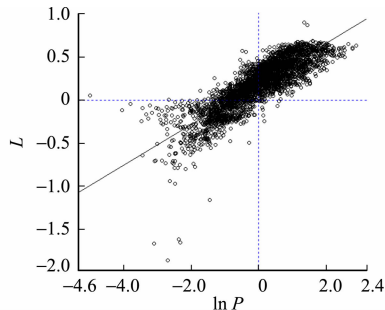


图 4 上海市二手房交易价格全局空间自相关散点图

Fig. 4 Global spatial autocorrelation scatter diagram of second-hand housing transaction price in Shanghai City



水平的冷热点分布,如图 6 所示.由图 6 可知:热点区域和冷点区域均在 99%的置信水平上显著.上海市二手房交易价格呈现“中间高、四周低”的空间格局,相较于 LISA 集聚,冷热点分布更宽,涉及更多边缘样本.城市中部的浦西七区、宝山区及闵行区的二手房交易价格为高-高集聚,环绕四周的嘉定区、青浦区、松江区、奉贤区和浦东新区外环城区的二手房交易价格为低-低集聚.

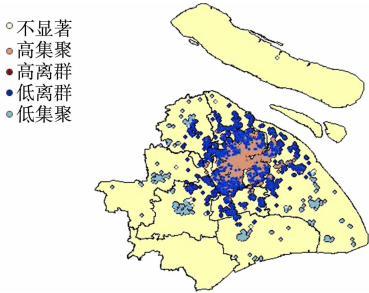


图 5 LISA 集聚状况  
Fig. 5 LISA agglomeration

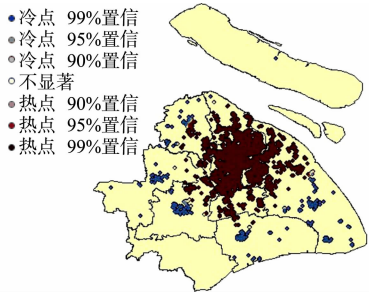


图 6 冷热点分布  
Fig. 6 Cold-hot spot distribution

2.3 空间计量模型估计

上海市二手房交易价格具有空间关联性,应选择空间计量模型进行分析.拉格朗日乘数检验项  $L\text{-}Mlag$ ,  $L\text{-}Merr$  及其稳健值  $R\text{-}L\text{-}Mlag$ ,  $R\text{-}L\text{-}Merr$  的检验结果,如表 4 所示.由于  $L\text{-}Mlag$ ,  $L\text{-}Merr$  均显著,需进一步比较  $R\text{-}L\text{-}Mlag$  和  $R\text{-}L\text{-}Merr$  的显著性,又因为  $R\text{-}L\text{-}Merr$  显著而  $R\text{-}L\text{-}Mlag$  不显著,故选择空间误差模型(SEM)进行分析.

表 4 拉格朗日乘数的检验结果

Tab. 4 Lagrange multiplier test results

诊断依据	统计值	$p$
$L\text{-}Mlag$	1 089.4	0
$L\text{-}Merr$	6 530	0
$R\text{-}L\text{-}Mlag$	0.212 26	0.645
$R\text{-}L\text{-}Merr$	5 440.7	0

基于特征价格法建立 SEM,探讨影响上海市二手房交易价格的可能因素. SEM 回归结果,如表 5 所示.表 5 中:  $E_s$  为标准误差;  $\lambda$  为空间自相关系数; \*, \*\*, \*\*\* 分别表示在 10%, 5%, 1% 水平上影响有统计学意义.对数似然值为 1 004.224;赤池信息准则 AIC 为 -1 974.4.

由表 5 可知:除厨房数量外,其他特征变量对二手房交易价格的影响皆有统计学意义;已装修、带电梯、有客厅且洗手间数量较多的二手房交易价更高;临近重点中小学、医院和市中心的二手房交易价格较高;然而,卧室数量及建筑面积在一定程度上会对二手房交易价格产生一定的抑制作用,原因可能是上海市过高的单价抑制了人们对大面积住宅的需求;楼龄与楼层均在 1% 的显著性水平下对房价有反向影响,但系数较小.

表 5 SEM 回归结果

Tab. 5 SEM regression results

特征变量	系数估计值	$E_s$	$Z$	特征变量	系数估计值	$E_s$	$Z$
常数项	7.785 190***	1.080 750	7.203 5	area	-0.000 580**	0.000 246	-2.374 6
room	-0.016 990**	0.007 863	-2.160 4	hall	0.036 266***	0.008 882	4.083 1
kitchen	-0.007 580	0.044 481	-0.170 3	bath	0.061 285***	0.012 953	4.731 2
age	-0.001 420***	0.000 485	-2.931 8	decoration	0.118 799***	0.014 439	8.227 8
elevator	0.084 272***	0.011 930	7.063 6	floor	-0.025 040***	0.004 709	-5.317 4
totalfloor	0.003 172***	0.000 731	4.339 3	D_center	-0.025 760***	0.001 026	-25.116 6
D_subway	0.003 701**	0.001 622	2.281 9	D_hospital	-0.004 090***	0.001 373	-2.982 6
D_school	-0.010 300***	0.001 372	-7.509 3	$\lambda$	0.997 080***	0.001 969	506.43

3 连续型深度置信网络房价预测

3.1 连续型深度置信网络简介

连续型深度置信网络(CDBNN)改造于深度置信网络(DBN). DBN 是由多个受限玻尔兹曼机(RBM)逐层堆叠而成,其核心思想是自底向上每一层 RBM 对输入数据进行提取、抽象,尽可能保留重要信息,训练过程一般采用贪婪无监督方式,即逐层对 DBN 中的每一个 RBM 进行训练.

RBM 是一种基于能量的概率生成模型,生成模型是对特征和标签之间的联合分布进行建模. 当可见层的状态  $v$  和隐藏层的状态  $h$  确定后,RBM 模型中的能量可以表示为

$$E(v, h | \theta) = - \sum_{i=1}^n a_i v_i - \sum_{j=1}^m b_j h_j - \sum_{i=1}^n \sum_{j=1}^m v_i \omega_{i,j} h_j. \tag{3}$$

式(3)中: $\theta$  为参数向量; $a_i$  和  $b_j$  分别为可见层第  $i$  个神经元上的偏置和隐藏层第  $j$  个神经元上的偏置; $\omega_{i,j}$  为可见层神经元和隐藏层神经元之间的连接权重值.

基于能量函数,可得  $v$  和  $h$  的联合概率分布为

$$P_{\theta}(v, h) = \frac{1}{z(\theta)} e^{-E(v, h | \theta)}, \quad z(\theta) = \sum_{v, h} e^{-E(v, h | \theta)}.$$

上式中: $z(\theta)$  为归一化函数,使得概率之和为 1.

依据联合概率分布,可以得到在可见层状态  $v$  确定时,隐藏层每个神经元被激活的概率,以及在隐藏层状态  $h$  确定时,可见层每个神经元被激活的概率分别为

$$P(v_i = 1 | h, \theta) = \sigma(a_i + \sum_j \omega_{i,j} h_j), \tag{4}$$

$$P(h_j = 1 | v, \theta) = \sigma(b_j + \sum_i v_i \omega_{i,j}). \tag{5}$$

式(4),(5)中: $\sigma(x)$  为 sigmoid 激活函数, $\sigma(x) = \frac{1}{1 + e^{-x}}$ .

对 RBM 的训练就是最大化对数似然函数,得到最优的参数值  $L(\theta) = \sum_{t=1}^T \lg P(v^{(t)} | \theta)$ , 其中,  $\{v^{(1)}, v^{(2)} \dots, v^{(T)}\}$  为训练样本集合; $T$  为样本容量.

一般采用梯度下降方法求取最优参数值,过程中涉及难以求解的归一化函数  $z(\theta)$ ,常用吉布斯(Gibbs)采样方法近似计算<sup>[20]</sup>. CDBNN 是在 DBN 的基础上改进,对式(4),(5)和激活函数  $\sigma(x)$  进行改进,使其适用于连续型数据,即

$$P(v_i = 1 | h, \theta) = \sigma(a_i + \sum_j \omega_{i,j} h_j + \varphi N_i(0, 1)), \tag{6}$$

$$P(h_j = 1 | v, \theta) = \sigma(b_j + \sum_i v_i \omega_{i,j} + \varphi N_j(0, 1)), \tag{7}$$

$$\sigma(x) = \theta_L + (\theta_H - \theta_L) \frac{1}{1 + e^{-x}}. \tag{8}$$

式(6)~(8)中: $N_i(0, 1), N_j(0, 1)$  表示均值为 0 且方差为 1 的高斯随机变量; $\varphi$  为常量; $\theta_H$  和  $\theta_L$  为渐近线,一般取样本中的最大值和最小值.

由于连续型深度置信网络是在深度置信网络的基础上衍生而来,因此,该方法同样采用误差反向传播的算法进行网络调优. CDBNN 算法主要有以下 8 个步骤.

**步骤 1** 准备训练数据  $D = (x_1, x_2, \dots, x_n)$ , 共  $n$  个样本,假设所有神经元的状态使用状态集  $\{S_i\}$  表示,随机初始化所有神经元的参数,设训练的最大次数为  $K$  次.

**步骤 2** 随机选择样本  $x_i$ , 输入可见层,由  $P(h_j = 1 | v, \theta) = \sigma(b_j + \sum_i v_i \omega_{i,j})$  计算第一个隐藏层的所有神经元的状态  $S_j$ .

**步骤 3** 根据步骤 2 得出的  $S_j$ , 同步步骤 2, 由  $P(v_i = 1 | v, \theta) = \sigma(a_i + \sum_j \omega_{i,j} h_j)$  计算可见层的重构神经元状态  $S_i$ .

**步骤 4** 根据步骤 3 所得的  $S_i$ , 同步步骤 2, 计算隐藏层的重构神经元状态  $S_j$ .

**步骤 5** 继续随机选择下一个训练样本,返回步骤 2, 如果样本集中的样本都选完毕,则依据式(8)计算参数变化量,更新方式为  $w_{i,j}(k+1) = w_{i,j}(k) + \Delta w, a_i(k+1) = a_i(k) + \Delta a_i$ .

**步骤 6** 进行第  $k+1$  次训练,当权重的变化量落入预定的范围内,即  $|\Delta w_{i,j}| < \epsilon$ , 其中,  $\epsilon$  是预先设定的误差范围,或者训练次数达到  $k$  次,则训练停止.

**步骤 7** 将训练好的 RBM 的输出作为下一层 RBM 的输入层输入数据,按照步骤 1~6 进行训练,直到训练完 DBN 的所有 RBM 层.

**步骤 8 网络调优:**完成 DBN 的训练后,需进一步优化深度神经网络权值.将训练好的 DBN 网络作为网络的初始状态,训练得出的参数作为 DBN 的初始参数;然后,使用反向传播的方法,运用梯度下降法对网络的整体权值进行有监督的学习.

3.2 连续型深度置信网络结构的确定

连续型深度置信网络结构的确定实质上就是选择深度置信网的超参数.待确定的神经网络结构的超参数包括神经网络的层数、神经网络隐藏层的节点数、学习率的确定、高斯随机变量中的方差值和样本迭代次数的选择及其他参数的选择.

超参数调优即选择超参数使网络结构达到最优的效果,是训练神经网络的核心任务.目前,常用的超参数调优方法有网格搜索与随机搜索.前者基于整个超参数空间进行搜索,速度较慢,但可获得最优的超参数组合.后者速度快,但可能会错过搜索空间中最优的超参数值.借鉴 Snoek 等<sup>[21]</sup>的思路,利用贝叶斯思想自动优化超参数,不仅能有效兼顾上述两种方法的优点,还能借助 Python 的 hyperopt 模块轻松实现优化超参数.主要超参数的估计值,如表 6 所示.

3.3 预测模型性能的比较

在建立连续型深度置信网络的过程中,将所有样本按 7 : 3 的比例随机分成训练集和测试集,先通过训练集对模型进行训练,再使用训练后的模型对测试集数据进行预测.将文中的预测结果与现有文献采用的支持向量机(SVM)、集成模型(采用 Adaboost 算法)和 BP 神经网络模型的预测结果进行对比.连续型深度置信网络、SVM、集成模型、BP 神经网络的预测误差分别为 0.006 67,0.007 61,0.008 42,0.029 03.BP 神经网络的预测误差远高于其他 3 个模型,这是由于随机初始化使其难以达到全局最优值.而连续型深度置信网络可预先对 BP 神经网络进行预处理,有效缓解随机初始化对最优预测的阻碍.此外,CDBNN 的预测结果也略优于 SVM 和集成模型,表明 CDBNN 有更高的复杂度,能够更加深入且全面地进行特征分析.

4 种模型测试集样本点的预测残差绝对值,如图 7 所示.图 7 中: $\varepsilon$  为残差绝对值.由图 7 可知:与其他模型相比,CDBNN 的预测残差总体情况更优,CDBNN 能够有效地解决 BP 神经网络在预测模型上存在的不足.

3.4 预测结果与讨论

通过绘制残差绝对值,可对 CDBNN 模型的预测结果进行评价.残差绝对值与二手房交易价格的关系,如图 8 所示.由图 8 可知:对于房价偏低或偏高的区域,影响上海市二手房交易价格的因素较为复杂,不仅局限于房屋建筑特征变量;房价偏低的区域大部分偏离市中心,距上海市重点中小学、三甲医院

表 6 主要超参数的估计值

Tab. 6 Estimated values of main hyper-parameters

超参数	估计值
RBM 层数	1
RBM 中隐含层神经元节点数	5
RBM 训练的学习率	0.402
高斯随机变量中的方差	0.006
RBM 训练中样本迭代次数	2
网络调优中训练的学习率	0.36
网络调优中样本迭代次数	4

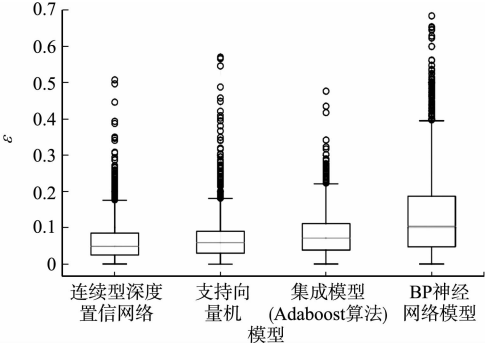


图 7 测试集样本点的预测残差绝对值  
Fig. 7 Absolute value of predicted residuals of test set sample points

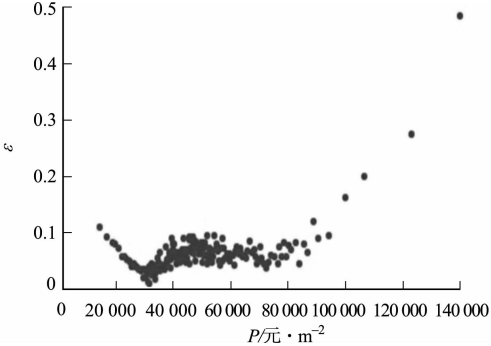


图 8 残差绝对值与二手房交易价格的关系  
Fig. 8 Relationship between absolute residual and transaction price of second-hand housing

以及地铁站距离较远,利用特征价格法选取的变量对房价偏低区域的房价预测能力相对较差.这些区域的二手房交易价格可能会更多地受到其所在区域的亚中心及该区域所配套的基础设施的影响.对于二



手房交易价格偏高的区域,预测的残差绝对值相对较大,可以认为当房价过高( $P \geq 80\,000$  元  $\cdot$   $m^{-2}$ )时,二手房交易价格的影响因素更加复杂.这其中除特征变量之外,还可能与购房者的购房目的等因素有关.对上海市高房价区域的购买者来说,房价弹性相对较低,他们对高房价并不敏感;高房价的购买者对房屋的消费不仅在于其本身的价值,而可能是出于政策便利性和高房价周围的邻里交际环境.

各行政区预测结果的残差绝对值平方,如图 9 所示.由图 9 可知:崇明区、黄浦区和静安区的预测结果的残差绝对值较大;崇明区的二手房交易价格偏低,而交易价格偏高的区域大多集中在黄浦区和静安区这两个市中心区域;浦东新区预测结果的残差绝对值较小,基于特征房价探讨的变量对浦东新区这样的非市中心的二手房交易价格的预测效果较好;杨浦区和浦东新区的预测结果的残差绝对值很相近,而静安区和黄浦区的残差绝对值相对较大.这可能是因为杨浦区和浦东新区隔海相望,杨浦区的经济和浦东新区的经济相互影响较大,黄浦区和静安区作为上海一直以来的市中心,其二手房交易价格的影响因素较为复杂;而浦东新区是改革开放后繁荣的区域,受上个世纪 90 年代开放的房地产市场影响较大,所以,预测效果较好,特征价格法所选的建筑特征、邻里特征和区位特征对新区房价的解释力度更强;对于黄浦区和静安区这样的老中心区域,其房价的解释力度相对较小;黄浦区和静安区的二手房交易市场的影响因素已经超出特征价格变量的解释范围.

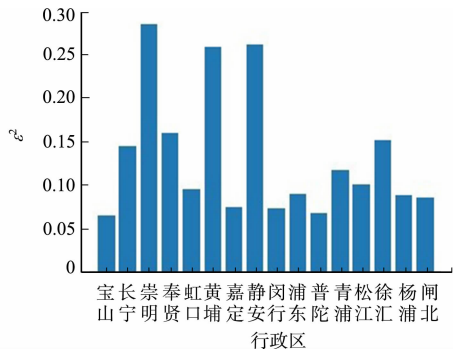


图 9 各行政区预测结果的残差绝对值平方  
Fig. 9 Square of residual absolute value of prediction results of each administrative region

### 4 结论

以上海市二手房交易市场为例,通过空间自相关分析,发现上海市二手房交易单价在空间上具有显著的自相关效应.二手房交易价格在上海市核心区域存在高-高集聚效应,在周边区域呈现低-低集聚效应,而在核心与周边交界地区存在高-低集聚和低-高集聚的负向空间效应.与此同时,基于连续型深度置信网络对特征二手房交易价格进行分析预测,发现特征变量对价格偏高区域的二手房交易价格解释力度较小,价格偏高区域的二手房交易价格影响因素较为复杂.从区域角度分析,除中心区域外,基于深度置信网络的特征变量对上海市二手房交易价格预测能力良好.连续型深度置信网络不仅能有效地解决 BP 神经网络在预测模型上存在的不足,而且与其他机器学习模型相比,连续型深度置信网络能更精准地对房价进行预测,从而为政府部门进行房价预测提供理论支持和政策导引.

文中采用一种能够处理大数据的深度学习模型,但由于获取数据的难度大,仅选取十余个解释变量.现实中,影响二手房交易价格的因素非常多,如加入更多的解释变量,基于连续型深度置信网络对特征二手房交易价格模型的预测将会更加精准.采用空间分析及深度学习技术对二手房交易价格进行研究,在各大城市均具有普遍适用性,可应用于其他城市的房价研究.

文中研究结果可为后续相关研究提供方法参考和模型借鉴.通过对上海市不同行政区及不同价格区间的房价预测模型效果进行差异性分析发现,在价格偏高的区域和上海市中心区域的预测效果较差.这为后续相关研究提供两方面借鉴:一方面,在预测房价时,需要考虑到空间异质性的影响,应针对不同区域构建不同的预测模型;另一方面,为进一步提高房价预测的精度,需要在建筑、区位、邻里等特征变量的基础上纳入更多相关的社会经济要素,从而提升模型的预测能力.对房价走势进行高精度预判,有助于政府制定调控政策.房地产市场调控一直是政府相关部门的工作重点,而稳定房价是调控的主要目标.对房价走势进行高精度预测具有一定的现实意义,可为政府相关部门完善房地产市场、优化城市规划设计提供一定的理论支持.

### 参考文献:

[1] 刘建丰,于雪,彭俞超,等.房产税对宏观经济的影响效应研究[J].金融研究,2020(8):34-53. DOI:1002-7246(2020)08-0034-20.

- [2] 冯苑.城市高房价会抑制居民劳动参与吗? [J]. 财经研究, 2020, 46(10): 154-168. DOI: 10. 16538/j. cnki. jfe. 20200518. 401.
- [3] 丛颖, 杜泓钰, 杨文静. 公共服务资本化对房价影响的空间计量分析: 基于我国 269 个地级市的经验研究[J]. 财经问题研究, 2020(7): 69-77. DOI: 10. 19654/j. cnki. cjwtyj. 2020. 07. 007.
- [4] RIDKER R G, HENNING J A. The determinants of residential property values with special reference to air pollution [J]. The Review of Economics and Statistics, 1967, 49(2): 246-257. DOI: 10. 2307/1928231.
- [5] XIAO Yixiong, CHEN XIANG, LI Qiang, *et al.* Exploring determinants of housing prices in Beijing: An enhanced hedonic regression with open access POI data[J]. ISPRS International Journal of Geo-Information, 2017, 6(11): 358-370. DOI: 10. 3390/ijgi6110358.
- [6] 张骥. 学区房溢价的再估计: 以北京市为例[J]. 经济问题探索, 2017(8): 57-63.
- [7] 王芳, 高晓路, 颜秉秋. 基于住宅价格的北京城市空间结构研究[J]. 地理科学进展, 2014, 33(10): 1322-1331. DOI: 10. 11820/dlkxjz. 2014. 10. 004.
- [8] 沈体雁, 于瀚辰, 周麟, 等. 北京市二手住宅价格影响机制: 基于多尺度地理加权回归模型(MGWR)的研究[J]. 经济地理, 2020, 40(3): 75-83. DOI: 10. 15957/j. cnki. jjdl. 2020. 03. 009.
- [9] LI Han, WEI Y D, WU Yangyi, *et al.* Analyzing housing prices in Shanghai with open data: Amenity, accessibility and urban structure[J]. Cities, 2019, 91: 165-179. DOI: 10. 1016/j. cities. 2018. 11. 016.
- [10] 申瑞娜, 曹昶, 樊重俊. 基于主成分分析的支持向量机模型对上海房价的预测研究[J]. 数学的实践与认识, 2013, 43(23): 11-16. DOI: 10. 3969/j. issn. 1000-0984. 2013. 23. 002.
- [11] 刘琼芳. 基于灰度 GM(1,1)模型的福州市房价预测[J]. 福建金融管理干部学院学报, 2018(1): 44-50. DOI: 1009-4768(2018)01-0044-07.
- [12] 陈娜, 唐晨旭, 刘伟, 等. 周口市住宅商品房价格的分析与预测[J]. 数学的实践与认识, 2019, 49(19): 291-299.
- [13] 张智鹏, 郑大庆. 影响区域房价的客观因素挖掘分析[J]. 计算机应用与软件, 2019, 36(11): 32-38, 85. DOI: 10. 3969/j. issn. 1000-386x. 2019. 11. 006.
- [14] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554. DOI: 10. 1162/neco. 2006. 18. 7. 1527.
- [15] DESELAERS T, HASAN S, BENDER O, *et al.* A deep learning approach to machine transliteration[C]// Proceedings of the 4th EACL Workshop on Statistical Machine Translation. Athens: Association for Computational Linguistics, 2009: 233-241. DOI: 10. 3115/1626431. 1626476.
- [16] SCHMIDHUBER J. Deep learning in neural networks: An overview[J]. Neural Network, 2015, 61: 85-117. DOI: 10. 1016/j. neunet. 2014. 09. 003.
- [17] LANGKVIST M, KARLSSO L, LOUTFI A. A review of unsupervised feature learning and deep learning for time-series modeling[J]. Pattern Recognition Letters, 2014, 42: 11-24. DOI: 10. 1016/j. patrec. 2014. 01. 008.
- [18] ZHANG Ren, SHEN Furao, ZHAO Jinxi. A model with fuzzy granulation and deep belief networks for exchange rate forecasting[C]// Proceedings of the 2014 International Joint Conference on Neural Networks. Beijing: IEEE Press, 2014: 366-373. DOI: 10. 1109/IJCNN. 2014. 6889448.
- [19] CHEN H, MURRAY A F. Continuous restricted boltzmann machine with an implementable training algorithm[J]. IEEE Proceedings-Vision Image Signal Process, 2003, 150(3): 153-158. DOI: 10. 1049/ip-vis: 20030362.
- [20] HINTON G E. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8): 1771-1800. DOI: 10. 1162/089976602760128018.
- [21] SNOEK J, LAROCHELLE H, ADAMS R P. Practical bayesian optimization of machine learning algorithms[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2012: 2960-2968.

(责任编辑: 黄晓楠 英文审校: 吴逢铁)