基于大数据技术的住宅价格空间分异研究

赵 彬，陈 超

四川轻化工大学计算机科学与工程学院 四川自贡

【摘要】本文选取宜宾市 5 县区做重点研究样本，在 ArcGIS、Python、Echarts、Pyecharts 等软件技术支持下对链家平台 2021 年宜宾市在售二手房房产数据获取、清洗及可视化研究，科学、客观的认识宜宾市住宅价格分异的规律,探寻各影响因素对住宅价格的影响情况,促进城市规划建设,促进房地产市场稳健发展。

【关键词】数据分析；数据可视化；住宅价格；多视图协同

A Study on Spatial Visualization of Residential Prices in Yibin City with Big Data Technology

Bin Zhao, Chao Chen

School of Computer Science and Engineering, Sichuan University of Science & Engineering

【**Abstract**】In this paper, five counties and districts of Yibin City are selected as key research samples. With the support of ArcGIS, python, echorts, pyechards and other software technologies, this paper studies the data acquisition, cleaning and visualization of second-hand houses sold in Yibin City in 2021 on the chain home platform, scientifically and objectively understands the law of housing price differentiation in Yibin City, explores the impact of various influencing factors on housing prices, promotes urban planning and construction, and promotes the steady development of the real estate market.

【**Keywords**】Arcgis; Data visualization; Residential prices; Real Estate

1 引言

宜宾市辖 10 个县级行政区划单位，选取叙州区（宜宾县）、翠屏区、南溪区、长宁县、江安县共五个行政区作为重点研究范围。数据采集与可视化研究具有广泛应用范围，涉及交通可视化、生物医学可视化、股市技术分析等多个领域[1-2]。将数据转换为图形表达，辅以交互手段，以增强人对数据的认知能力、决策辅助等方面展示了强大的赋能作用[3]。故将大数据技术引入至房价分析予以可视化方式呈现，探究城市房价的分异规律。在可视化领域多视图协同作业的方式能够更好地表达数据信息的潜在价值，宜宾市发展迅速，“宜宾大学城”的大力引进政策使得宜宾房产呈现温热态势。为此做可视化研究意义颇深。

2 数据获取

2.1 宜宾市二手房价数据采集

房产信息数据：通过网络爬虫模拟浏览器客户端批量的请求服务器数据。具体是将链家房产网宜宾市二手房信息抓取包括地址、每平米单价、名称等信息。该网络爬虫共得到房产类别标题、经纪人、户型、面积、单价、地址、建筑时间等共 18 个维度信息。

地理信息数据：致力于完善的可视化效果，基于爬虫得到的地址字段信息对每条房产信息进行经纬度添加。使用 Map Location 工具批量转化地址信息至经纬度坐标，使用中国坐标偏移标准 BD-09 坐标系以提高精度的可靠性。从而生成兼具价格与地理双重属性的样本数据。

2.2 数据清洗

对房价数据的清洗主要有：对列表元素进行合

作者简介：赵彬（1997-）男，硕士，主要研究方向为可视化与可视分析、数据分析等；
陈超（1980-）男，博士，教授，主要研究方向为人工智能、大数据分析等。

并，对数据缺失值、异常值、重复值等的处理。对于异常值和缺失值的处理需有先后顺序，不能将异常值作为缺失值填补的依据，故先进行异常值的处理。格拉布斯(Grubbs)检验法：一组测量数据中，如果个别数据偏离平均值很远，那么这个)数据称作“可疑值”。如果用统计方法一例如格拉布斯(Grubbs)法判断，能将“可疑值”从此组测量数据中剔除而不参与平均值的计算，那么该“可疑值”就称作“异常值(粗大误差)”。计算步骤如下：

（1）计算样本单价的平均值:

$$\mu=\frac{(X_1+X_2+\cdots+X_n)}{n} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1)$$

（2）计算样本标准差:

$$s=\sqrt{\frac{\sum(X_i-\mu)^2}{n-1}}(i=1,2\cdots n) \cdots\cdots\cdots\cdots\cdots (2)$$

（3）计算格拉布斯检验统计量:

$$G_i=\frac{(X_i-\mu)}{s}(i=1,2\cdots n)\#(3)$$

（4）将计算值 G_i 与格拉布斯表 $G_p(n)$给出的临界值比较，如果计算的值大于临界值则视为异常值。

（5）根据本房价数据设置临界值 $G_p(n)=3$,故当 $G_i>G_p(n)$时，判别 X_i为异常值，反之正常。

表 2 部分处理后数据

No	Address	Floor	Type	Area	Price	Face	Time
1	翠屏西郊滨江国际	中层(共 29 层)	3 室 2 卫	155 ㎡	8259 元/㎡	南	2009
2	翠屏上江北红丰东路 8-1 号院	低层(共 9 层)	3 室 1 卫	90.6 ㎡	5611 元/㎡	东南	2003
3	叙州柏溪镇毕加索庄园	高层(共 33 层)	3 室 2 卫	84.5 ㎡	8876 元/㎡	南	2021
4	翠屏上江北金科集美江山	中层(共 29 层)	3 室 2 卫	82 ㎡	7781 元/㎡	东南	2021
5	翠屏下江北大学府文澜院	中层(共 25 层)	2 室 1 卫	80 ㎡	8525 元/㎡	西北	2013

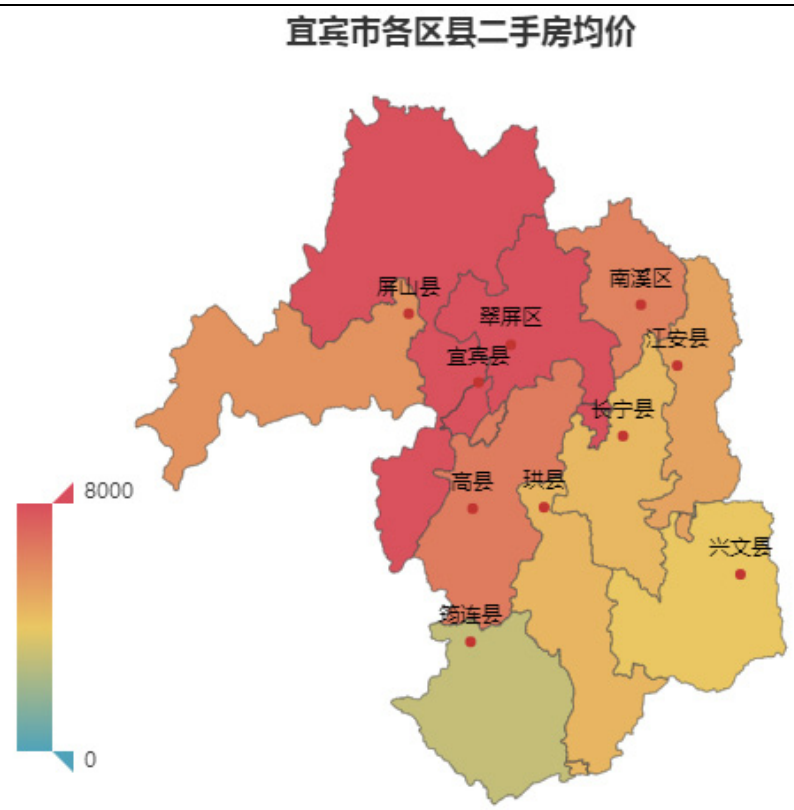

图 1 各区县房均价热力图

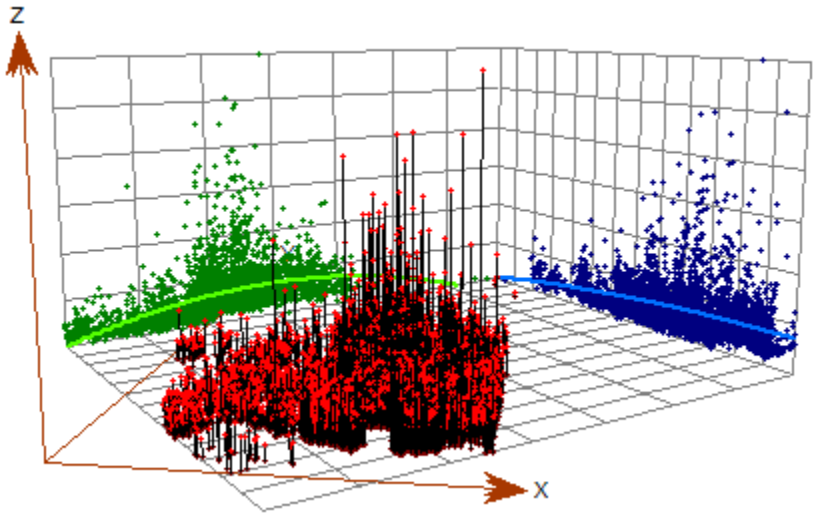

图 2 空间趋势图

3 房产可视化分析

3.1 房价数据分组

将房产数据分类的主要依据是按照影响房价因素的重要程度进行聚类和分组，对于样本数据中宜宾整体二手房价平均为 7652 元/㎡。但整体均价对异常值并不敏感，并不能全然展现宜宾市房价。所以无法以整体均价做为在宜宾购房的参考价格。故将样本数据聚类以区县为单位分为叙州区、翠屏区、南溪区、长宁县、江安县五个行政区着重研究，对其他区县概括研究。箱形图最大的优点就是不受异常值的影响，能够准确稳定地描绘出数据的离散分布情况，同时也利于数据的清洗。城市房价热力图以地图形式分类展示房价信息的高低程度，帮助购房者根据需求在现波动范围内合理决策。如图 1 所示。

根据图 1，将宜宾楼盘按区县分组，可以清晰地看到翠屏和叙州区（宜宾县）房价总体较高，翠屏区位于全市首位。箱型图将房价分成四等份，如翠屏区 25%的楼盘在 3500 元-6400 元相差 2900 元，25%的楼盘在 6400 元-7800 元相差 1400 元，25%的楼盘在 7800 元-9400 元相差 1600 元，剩余 25%在 9400 元-13000 元相差 3600 元。中位数相对较多，总体不同价格分布比较均衡，可以满足不同收入人群购买。虽有疫情影响，但各区县 GDP 总量仍属于正增长状态。其中翠屏区 GDP 较排在第二位的叙州区达两倍之多。宜宾大学城的引进发展也为翠屏区房价提供强劲动力。长宁县和江安县房价相对稳定，价格差距小，若不考虑其他因素而仅是刚需可以优先考虑。

3.2 空间探索性分析

城市地理空间数据的可视化表达能在基础地图上直观反映城市的人口、经济等民生数据。房价作为其标准之一，也是大众最为关心。结合空间趋势分析图清晰地表达宜宾市房价分布。运用 Arcgis 10.5 软件中 Geostatistical Analyst 的数据探索功能，将含有地理坐标信息的房价数据添加至趋势分析模块，因位置信息精确至小区，故出现多个房产信息有相同的坐标数据。为确保数据的完整性对重合样本进行全部包括。如图 2 所示，x 轴 y 轴分别代表地理信息的东西和南北方向，z 轴表示二手房价格。红点表示二手房源的价格点，长短不等的黑色短线至底面表示房源居于的不同地理位置，顶部的红色价格点分别投影至南北和东西的网格上形成绿色和蓝色点，更清楚的看出房源数据的分布状态，且趋势分析根据网格上的投影拟合出一条趋势曲线。由图 2 看出，房源主要集中至市中心位置，四周南北向和东西向房源相对较少，说明宜宾市中心房源密度较大，郊区相对较少。根据价格点投影拟合的趋

势曲线呈现出倒”U”形，随之价格也呈现出”中间高，四周低”的表现趋势。

4 结语

主要介绍了房产专题数据的抓取、清洗、可视化和空间化集成方法。主要针对房产门户网站链家作数据来源，在 Python 网络爬虫技术支持下对宜宾市房产数据进行抓取，得到房产类别户型、面积、单价、等共 18 个维度信息，而由此进一步得到关键的地理信息数据。利用 pandas 等技术对数据二次加工后利用图表以及空间探索方法对数据集成表达，而不仅仅是简单的数据图形化。通过程序和文章的表现情况可以看出，可视化的表达手段能够清晰有效地传递原始数据所隐含的潜在信息。利用 Echarts、Arcgis、Pyecharts 等技术手段对宜宾市房产价格数据深入分析研究并做可视化处理，把房价及其影响要素以一种合理的可视化的形式加以呈现。相比传统分析方法对数据的理解获取有明显优势。目前，改文章仅初步粗浅的实现房产数据的分析与可视化表达，此后将继续完善，集成更有力的分析工具促进数据更加量化和精细化。

参考文献

[1] Tuarob Suppawong et al. DAViS: a unified solution for data collection, analyzation, and visualization in real-time stock market prediction[J]. Financial Innovation, 2021, 7(1)

[2] Qiu Yongjian and Lu Jing. A visualization algorithm for medical big data based on deep learning[J]. Measurement, 2021, 183

[3] 夏佳志,李杰,陈思明,秦红星,刘世霞.可视化与人工智能交叉研究综述[J/OL].中国科学:信息科学:1-25[2021-11-18].http://gffiyc3066c973e9140a1hp90f56p9nupp6fnp.fffb.suse.cwkeji.cn:999/kcms/detail/11.5846.TP.20211115.0847.002.html.

[4] 曾婷婷. 基于机器学习的房价预测模型研究[D].西南科技大学,2020.

[5] 魏迪. 城市二手房交易数据可视化系统研究与实现[D].天津大学,2018.

[6] Dobson Barnaby A. et al. Effects of flood hazard visualization format on house purchasing decisions[J]. Urban Water Journal, 2018, 15(7) : 671-681.

[7] 徐征,洪明月,宋玉. 基于“人才引进”政策的西安楼市现状探究及房价预测[A]. 中国统计教育学会、教育部高等学校统计学类专业教学指导委员会、全国应用统计专业学位研究生教育指导委员会.2019 年（第六届）全国大学生统计建模大赛优秀论文集[C].中国统计教育学会、教育部高等学校统计学类专业教学指导委员会、全国应用统计专业学位研究生教育指导委员会:中国统计教育学会,2019:56.

收稿日期：2022 年 8 月 18 日
出刊日期：2022 年 9 月 6 日
引用本文：赵彬，陈超，基于大数据技术的住宅价格空间分异研究[J]. 国际计算机科学进展, 2022, 2(2) : 17-20.
DOI: 10.12208/j. aics.20220016

检索信息：RCCSE 权威核心学术期刊数据库、中国知网（CNKI Scholar）、万方数据 （WANFANG DATA）、Google Scholar 等数据库收录期刊