

# 基于机器学习方法的 一线城市房价影响因素研究

陈小亮 程 硕 陈 衍 肖争艳

**摘要:**一线城市的房地产市场在全国房地产市场中具有举足轻重的地位,因此,维持一线城市房价平稳从而对保障全国房价平稳健康发展具有重要意义。运用XGBoost等机器学习方法和SHAP值可解释性方法,对四大一线城市房价的主要影响因素及其在2012年前后的动态变化进行测算并分析,研究发现:第一,预期因素是一线城市房价上涨的主要影响因素,并且其影响在不断增强。第二,供给因素和需求因素对一线城市房价上涨也起到了较为重要的作用,不过其作用呈现出减弱态势。第三,货币政策等因素对一线城市房价上涨的影响相对偏弱,并且近年来其影响进一步下降。考虑到预期因素是一线城市房价上涨的最主要因素,因此对一线城市而言,稳房价的关键在于稳预期。进一步地,结合实证结果可知,需要让房价更多地由基本面因素来决定,并且通过稳定房地产调控政策来稳预期。一是从供给端发力,构建一线城市土地供给与房价以及土地供给与常住人口之间的联动机制。二是从需求端发力,缩小一线城市与其周边城市以及其他三四线城市之间的公共服务差距,从而减轻一线城市的外来人口压力以及由此引发的住房需求增长。三是保持房地产调控政策的连续性、一致性、稳定性,通过稳定政策来稳定房地产市场的预期。

**关键词:**一线城市;房价;稳预期;机器学习方法

## 一、引 言

房改以来的二三十年间,一线城市房价已经出现过多次轮上涨态势<sup>①</sup>,并且对全国房地产市场运行和货币政策操作产生了重要影响。在2012年之前,一线城市房价上涨通常会带动二三线等其他城市的房价上涨,进而引发全国范围内的房价上涨。而在2012年之后,一线城市与二三线城市房价走势呈现明显分化,而且城市间房价分化在多数情况下是由一线城市房价上涨所引发的,集中表现为一线城市房价涨幅明显超过二三线城市。需要强调的是,2012年之后中国经济持续面临下行压力,货币政策本应适时发力“稳增长”,但是当货币政策发力时又极易引发一线城市房价上涨风险。在此情形下,货币政策被动收紧以“控房价”,但是也导致其“稳增长”的力度有所不足,这在

收稿日期:2023-06-28

基金项目:国家自然科学基金面上项目(72073141);国家自然科学基金专项项目(72141306)

作者简介:陈小亮,中国社会科学院经济研究所(北京 100836)副编审,博士,主要从事经济增长与宏观政策研究;程硕,中国人民大学统计学院(北京 100086)博士生,主要从事大数据与经济统计研究;陈衍,中国人民大学统计学院博士生,主要从事大数据与经济统计研究;肖争艳,中国人民大学统计学院教授,博士生导师,主要从事大数据与经济统计、保险精算研究。

① 仅2006年以来,一线城市就已经经历了五轮房价上涨,分别发生在2006年7月至2008年5月、2009年10月至2010年6月、2013年3月至2014年6月、2015年6月至2017年8月、2020年6月至2021年12月期间。

2020—2021年应对疫情期间表现得尤为明显。当前一线城市房价增速相对平缓,面对经济下行压力,货币政策等房地产调控政策呈现放松迹象<sup>①</sup>,由此也引发了一线城市房价再次上涨的担忧。因此,深入探寻一线城市房价上涨的影响因素,不仅能够“有的放矢”保障一线城市房价平稳,而且能够带动全国房地产市场平稳健康发展,同时帮助货币政策摆脱“控房价”与“稳增长”之间的两难选择,更好地发力“稳增长”。

已经有不少文献深入研究了一线城市房价的影响因素。就研究对象而言,一部分文献直接研究了所有一线城市房价的影响因素,如刘仁和等,张浩等,余华义和黄燕芬。<sup>②</sup>还有一部分文献重点研究了北京或上海等某个一线城市房价的影响因素,如张勋等,冯皓和陆铭<sup>③</sup>;或者是将一线城市与二三线城市房价的影响因素进行对比分析,如李迎星等,王松涛<sup>④</sup>。就影响因素而言,已有研究主要关注了四大类因素,分别是教育等公共资源和人口等需求侧因素,土地供给和住宅建造成本等供给侧因素,如屠佳华和张洁,张立建<sup>⑤</sup>;货币政策和限购限贷等政策因素,如张清源等,张牧扬等<sup>⑥</sup>,以及房价预期因素,如孙伟增和林嘉瑜<sup>⑦</sup>,陈建等<sup>⑧</sup>。就研究方法而言,已有文献主要使用的是VAR模型及其拓展模型,如杨思群和董美<sup>⑨</sup>、截面回归、面板回归、GMM等传统实证方法,如孙伟增和林嘉瑜<sup>⑩</sup>。

已有研究对于一线城市房价影响因素进行了较好的探索,不过仍然存在一些不足之处。其一,已有研究主要使用的是VAR模型和面板回归等传统实证方法,这些方法所能考察的影响因素个数较为有限,而且通常只考察了相关因素对房价的线性影响。但是,现实中影响房价的因素众多,很多因素对房价的影响可能呈现复杂的非线性关系。其二,2012年前后中国宏观经济走势发生了明显变化,不同地区的房地产市场的运行也随之出现了一些趋势性的变化,因此2012年前后一线城市房价的主要影响因素很可能是不同的,但是已有研究很少从动态视角探析一线城市房价影响因素的变化。

与VAR模型和面板回归等传统实证方法相比,机器学习方法具有一些明显优势,更适合用来研究一线城市房价的影响因素。机器学习方法具有更强大的计算和求解能力,不仅可以考察多个因素的影响,而且可以将各个因素对房价的非线性影响考虑在内。更重要的是,将机器学习方法与SHAP(Shapley Additive Explanations, SHAP)值可解释性方法相结合,可以针对不同时期一线城

① 其中,广州和深圳于2023年8月30日、上海和北京于2023年9月1日分别宣布执行“认房不认贷”政策。

② 刘仁和、陈奕、陈英楠:《租金房价比失效了吗——基于四大城市的经验研究》,《财贸经济》2011年第11期,第105—111页;张浩、李仲飞、邓柏峻:《教育资源配置机制与房价——我国教育资本化现象的实证分析》,《金融研究》2014年第5期,第193—206页;余华义、黄燕芬:《利率效果区域异质性、收入跨区影响与房价溢出效应》,《经济理论与经济管理》2015年第8期,第65—80页。

③ 张勋等:《学区房溢价的影响因素:教育质量的视角》,《金融研究》2021年第11期,第97—116页;冯皓、陆铭:《通过买房而择校:教育影响房价的经验证据与政策含义》,《世界经济》2010年第12期,第89—104页。

④ 李迎星、田露、杨梦:《限购政策是否降低房地产价格增速》,《系统工程理论与实践》2019年第4期,第906—921页;王松涛:《中国住房市场政府干预的原理与效果评价》,《统计研究》2011年第1期,第27—35页。

⑤ 屠佳华、张洁:《什么推动了房价的上涨:来自上海房地产市场的证据》,《世界经济》2005年第5期,第28—37页;张立建:《利润最大化区位理论与广州高房价的根源》,《统计研究》2008年第9期,第16—23页。

⑥ 张清源、梁若冰、朱峰:《货币政策加剧城市房价的冷热不均吗》,《统计研究》2018年第5期,第75—87页;张牧扬、陈杰、石薇:《交易环节税率差别对房价的影响——来自上海二手房数据的实证研究证据》,《财政研究》2016年第2期,第79—93页。

⑦ 孙伟增、林嘉瑜:《教育资源供给能够降低学区房溢价吗?——来自北京市新建小学的证据》,《经济学(季刊)》2020年第2期,第499—520页。

⑧ 陈建、陈英楠、刘仁和:《所有权成本、投资者预期与住宅价格波动:关于国内四大城市住宅市场的经验研究》,《世界经济》2009年第10期,第14—24页。

⑨ 杨思群、董美:《信贷政策对房价波动的影响分析》,《投资研究》2017年第4期,第90—102页。

⑩ 孙伟增、林嘉瑜:《教育资源供给能够降低学区房溢价吗?——来自北京市新建小学的证据》,《经济学(季刊)》2020年第2期,第499—520页。

市房价的主要影响因素进行动态分析。有鉴于此,本文将综合使用机器学习方法和SHAP值可解释性方法,对一线城市房价的影响因素进行重新识别。具体而言,本文将使用2006—2021年的月度数据开展实证研究,进而详细回答一线城市房价的主要影响因素分别是哪些,在2012年前后房价的影响因素是否存在显著差异等重要问题,在此基础上给出能够促进一线城市房价平稳发展的应对策略。

本文的边际贡献主要有三点:第一,本文较早地将机器学习方法应用到一线城市房价影响因素的研究领域。与传统方法相比,机器学习方法能够考察的影响因素更多而且还能够挖掘变量背后的非线性关系,从而更准确地识别一线城市房价的影响因素。第二,传统方法很难对不同影响因素的相对重要性进行排序,也就很难识别一线城市房价上涨的主要影响因素是哪些,而机器学习方法可以对影响因素的相对重要性进行排序,能够更准确地识别一线城市房价的主要影响因素究竟是哪些。在此基础上,决策部门可以给出更有针对性、更高效的应对策略。第三,本文不仅将所有一线城市作为整体加以分析,而且还分析了每个一线城市房价的主要影响因素及其在2012年前后的动态变化,能够更好地为房地产“因城施策”调控理念的实施提供决策参考。

## 二、相关文献评述

自房改以来一线城市房价已经出现过多轮上涨,而且一线城市的房价上涨往往会带动二三线等城市的房价上涨,因此已经有大量文献围绕一线城市房价上涨的影响因素进行了深入研究。总体而言,已有文献按照研究对象的不同,可以归纳为三类。

第一类文献将一线城市与二三线城市房价上涨的影响因素进行了对比分析,考察了货币政策和限购限贷等调控政策、预期、人口密度等因素对一线城市与二三线城市房价影响的差异。就货币政策而言,张清源等研究发现,货币供给量的增长对一线城市房价的上涨作用显著高于二三线城市;类似地,杨思群和董美指出,相对于二线和三线城市,一线城市的房价对信贷政策的敏感性更高。从限购等调控政策的角度,李迎星等研究发现,2014年底限购政策的解除导致一线和二线城市房价增速大幅反弹,三线城市反而因为虹吸效应并没有出现房价反弹;这种城市间的差异性在郭金金和夏同水的研究中得到了进一步印证,他们揭示了一线、二线、三线城市在面对租购并举政策时的敏感程度分别呈现出逐级降低的趋势<sup>①</sup>;对于这种房价差异性,王松涛指出,全国性政策工具对不同城市施加的影响力度存在显著差异,导致二线城市房价受到的抑制效果明显超过一线城市。在研究房价预期的影响上,况伟大发现,房价预期对一线城市房价上涨的推动作用要比二三线城市更显著。就人口因素而言,王洁和张继良证实,与二三线城市相比,一线城市人口过密对房价产生了显著推动作用。<sup>②</sup>此外,赵伟等还从媒体视角研究了媒体关注度对房价的影响,发现当期媒体关注度对一线城市的影响要小于二三线城市。<sup>③</sup>

第二类文献聚焦研究了某个一线城市房价上涨的影响因素。针对北京,张勋等使用带有边界固定效应的Hedonic特征价格模型研究发现,以优质教育资源带来的学区房溢价是北京房价上涨的重要因素,孙伟增和林嘉瑜开展了类似研究而且得到了相似的结论;梁军辉等则全面分析了地

① 郭金金、夏同水:《租购并举政策下住宅适应性预期价格演化及仿真——兼论项目建设开发速度与土地供给速度对住宅价格的影响》,《中国软科学》2019年第9期,第51—60页。

② 王洁、张继良:《住房空置率对房价的影响——基于35个重点城市的面板数据》,《资源科学》2020年第6期,第1135—1147页。

③ 赵伟、耿勇、何雅静:《媒体报道、预期与房价波动》,《经济评论》2018年第2期,第133—146页。

铁站点、商场、休闲娱乐设施、小区绿化率、医疗资源等多方面公共服务因素对北京房价的影响。<sup>①</sup>针对上海,冯皓和陆铭以及张牧扬等都分析了教育资源对上海学区房溢价的影响;屠佳华和张洁则采用VAR模型,从经济角度出发,关注了房地产投资占比、人均可支配收入、空置面积、滞后一期房价等因素对上海房价的影响;在政策调整的影响上,张牧扬等基于双重差分思想,深入研究了2014年11月上海市普通住房认定标准的调整对二手房价格所带来的变化。在广州和深圳的房价影响因素研究中,学者们的探索路径略有不同,比如张立建关注了住房供给与收入差距两大因素,探讨它们是如何共同影响广州房价的,蒲火元等则从人口规模这一角度分析人口流入对广州房价的作用。<sup>②</sup>而林江等则使用协整和VAR模型检验了信贷政策和房价预期对深圳房价的影响。<sup>③</sup>

第三类文献重点关注了某个因素或少数几个因素对四大一线城市房价上涨的影响及其差异。刘仁和等基于住房使用成本模型,构建了租金和房价变动的预测模型,分析了租金房价比对一线城市房价的影响,发现北京、上海的租金房价比与未来房价负相关,深圳的租金房价比与未来房价正相关,而广州的租金房价比与未来房价不相关。张浩等分析了教育资源对一线城市房价的影响,发现基础教育资源通过“资本化”不断蕴含在房价之中,进而对房价产生持续影响。余华义和黄燕芬基于GVAR模型检验了居民收入对房价的影响,结果表明一线城市的房价波动不仅受本城市人均收入变动的影 响,还在很大程度上受其他城市收入变动的影 响。此外,多数学者对于投资者预期对四大一线城市的影响做了重点讨论:如陈建等检验了住宅所有权成本与投资者预期对一线城市房价的影响,发现广州与深圳的住宅所有权成本可能不是房价上涨的主要因素,而北京与上海的住宅所有权成本则推升了房价,投资者预期对四个城市的房价上涨都有较为显著的解释能力。张浩和李仲飞同样分析了投资者预期对一线城市房价的影响,发现投资者预期对上海房价的推动力超过广州。<sup>④</sup>

这三类文献在研究过程中大都进行了实证检验,而且使用的方法主要是VAR模型及其拓展模型、截面回归、面板回归、GMM等传统实证方法。就VAR模型及其拓展模型,屠佳华和张洁等分别研究了上海和深圳房价的影响因素<sup>⑤</sup>,余华义和黄燕芬使用GVAR模型研究了居民收入对一线城市房价的影响<sup>⑥</sup>,杨思群和董美使用FAVAR模型研究了信贷政策对一线城市房价的影响<sup>⑦</sup>。就截面回归、面板回归、GMM等回归方法而言,张牧扬等和蒲火元等使用截面回归方法分析了上海和广州房价的影响因素<sup>⑧</sup>,冯皓和陆铭等人的文献使用面板回归方法分析了一线城市房价的影响因素<sup>⑨</sup>。

已有文献对本文探寻一线城市房价上涨的影响因素大有裨益,不过仍然存在几方面不足和有待改进之处。第一,影响一线城市房价的因素众多,已有文献主要分析了某一类或少数几类因素

① 梁军辉、林坚、吴佳雨:《北京市公共服务设施配置对住房价格的影响》,《城市发展研究》2016年第9期,第82—87页。

② 蒲火元、曹宗平、李超:《人口流动对中心城市房价的影响:以广州为例》,《南方人口》2018年第5期,第27—40页。

③ 林江、徐世长、黄建新:《适应性预期、金融加速器效应与房价波动——以深圳为例》,《现代财经(天津财经大学学报)》2015年第1期,第22—31页。

④ 张浩、李仲飞:《房价预期、土地价格与房地产商行为》,《管理评论》2016年第4期,第52—61页。

⑤ 屠佳华、张洁:《什么推动了房价的上涨:来自上海房地产市场的证据》,《世界经济》2005年第5期,第28—37页。

⑥ 余华义、黄燕芬:《利率效果区域异质性、收入跨区影响与房价溢出效应》,《经济理论与经济管理》2015年第8期,第65—80页。

⑦ 杨思群、董美:《信贷政策对房价波动的影响分析》,《投资研究》2017年第4期,第90—102页。

⑧ 张牧扬、陈杰、石薇:《交易环节税率差别对房价的影响——来自上海二手房数据的实证研究证据》,《财政研究》2016年第2期,第79—93页;蒲火元、曹宗平、李超:《人口流动对中心城市房价的影响:以广州为例》,《南方人口》2018年第5期,第27—40页。

⑨ 冯皓、陆铭:《通过买房而择校:教育影响房价的经验证据与政策含义》,《世界经济》2010年第12期,第89—104页。

的影响是否显著,但是并没有进一步分析究竟哪些因素是最主要的影响因素,也就无法给出更加有效、更加明确的决策参考。这主要是因为,已有文献使用的是VAR模型和面板回归等传统实证方法,所能考察的变量个数有限,而且无法对不同因素的相对重要性进行对比分析。第二,房价预期等因素对房价上涨的影响很可能是复杂的非线性形式,但是已有文献通常只考虑了这些因素对房价的线性影响。即便是少数文献考虑了某些因素对房价的非线性影响,也要事先假定方程的非线性形式,在此过程中很可能产生偏误。第三,伴随着时间的推移,经济增长态势和房地产市场的发展趋势存在显著差异,这意味着不同时期一线城市房价影响因素可能是不同的,但是已有研究大都没有从动态视角分析一线城市房价影响因素在不同时期的变化。

机器学习方法相比传统实证方法具有多方面优势,为本文更准确地识别一线城市房价影响因素提供了方法支撑。首先,机器学习方法的测算框架更加灵活,求解和计算能力更加突出,因此所能涵盖的变量个数显著增加,从而能够更全面地考虑一线城市房价的影响因素。其次,机器学习方法具有数据驱动的特征,不需要事先设定函数形式,而是利用数据找到最优的函数形式,能够更好地解决非线性问题,有效地挖掘各个因素对房价的非线性影响。第三,SHAP值可解释性方法等可解释性机器学习方法在动态分析方面具有显著优势,如Lundberg<sup>①</sup>和Lee,Aas等<sup>②</sup>,能够细致分析一线城市房价影响因素在不同时期的差异。得益于上述优势,机器学习方法相比传统方法更适合用来研究一线城市房价的影响因素。虽然已经有少数文献尝试使用机器学习方法对房价进行预测研究,例如,唐晓彬等<sup>③</sup>基于蝙蝠算法SVR方法,对北京市二手房价进行了预测研究。李仲达等<sup>④</sup>利用高维稀疏VAR模型,对35个大中城市的住宅销售价格进行了预测研究。崔明明等<sup>⑤</sup>基于TEI@I思想,采用综合集成预测方法,对70个大中城市的住房价格进行了预测研究。董倩等<sup>⑥</sup>使用Bagging、Boosting等多种方法,对北京、上海等16个大中城市的房价进行了预测研究。但是尚无文献使用机器学习方法深入研究一线城市房价的影响因素。有鉴于此,本文将使用机器学习方法重新识别一线城市房价上涨的影响因素及其差异,并对2012年前后房价影响因素的动态变化进行对比分析,从而有针对性地给出促进一线城市房价平稳运行的政策建议。

### 三、方法选取与模型构建

本文的主要目的是,综合使用XGBoost方法和SHAP值可解释性方法,分析一线城市房价的影响因素。为了实现这一目标,本文在模型构建方面主要包括两部分工作。一是基于一线城市房价及其影响因素数据训练并运用交叉验证方法挑选最优的XGBoost方法并进行建模。二是运用Aas等提出的SHAP值可解释性方法揭示XGBoost方法所识别的一线城市房价主要影响因素。

#### 1.XGBoost方法介绍与模型训练

首先需要阐明本文选择使用XGBoost方法的理由。本文将XGBoost方法在数据集上的预测性能与Random Forest(随机森林)、SVR(支持向量回归)、Lasso回归等常用机器学习方法以及普通

① Scott M. Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, December, 2017, pp.4765-4774.

② Aas Kjersti, et al., "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values," *Artificial Intelligence*, Vol. 9, 2021, pp.1-24.

③ 唐晓彬、张瑞、刘立新:《基于蝙蝠算法SVR模型的北京市二手房价预测研究》,《统计研究》2018年第11期,第71—81页。

④ 李仲达、林建浩、王美今:《大数据时代的高维统计:稀疏建模的发展及其应用》,《统计研究》2015年第10期,第3—11页。

⑤ 崔明明等:《数据特征驱动的房地产市场集成预测研究》,《管理评论》2020年第7期,第89—101页。

⑥ 董倩、孙娜娜、李伟:《基于网络搜索数据的房地产价格预测》,《统计研究》2014年第10期,第81—88页。

面板混合回归模型进行了比较,发现XGBoost方法在测试集上的预测性能最优(MAE=0.95, R<sup>2</sup>=0.98), Random Forest方法的预测性能次之。<sup>①</sup>有鉴于此,本文在基准分析中使用XGBoost方法,在后文的稳健性检验中还将使用Random Forest方法加以分析。

XGBoost是由Chen和Guestrin<sup>②</sup>在梯度提升树(Gradient Boosting Decision Tree, GBDT)基础上提出的改进方法,相比于Boosting和Bagging等方法具有更优的预测效果。XGBoost方法的改进主要体现在模型训练效率的优化和降低过拟合上。具体训练过程中,假设X是一线城市房价影响因素的相关指标集合,Y为一线城市的房价增速,给定训练数据集为 $D=\{(x_1, y_1), (x_2, y_2), \dots, (x_s, y_s)\}$ ,其中 $x_i=(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)})$ 为输入样本实例,p为影响因素个数(p=18,详见下文),s为样本个数。对于X和Y形成的样本集合,本文进行了以下的预处理:一是对指标集合X进行数据频率转换,使用三次样条插值将其中的季度和年度数据转换为月度数据;二是采用MissForest预测模型方法对一线城市中的少量缺失值进行填补;三是对于部分指标存在的少量异常值,通过Winsorize处理将异常值缩尾至合理区间内。然后,将预处理后的样本输入至不同超参数组合下的XGBoost进行网格搜索(GridSearch),寻找最优的模型超参数,最终本文选取的最优模型训练参数为:学习率等于0.05,子树最大深度等于8,最大迭代次数等于800。

## 2. SHAP值可解释方法介绍

虽然XGBoost方法可以较为准确地预测一线城市房价增速,但是与线性方法相比,XGBoost方法却是“黑箱”模型,难以解释哪些因素对房价具有重要影响。为了进一步分析一线城市房价的影响因素,本文选用SHAP值可解释性方法对一线城市房价的影响因素进行解释和分析。SHAP值可解释性方法不仅克服了变量替换法(Feature Permutation)和LIME等解释性方法在统计性质上的一些不足<sup>③</sup>,还能在每个分样本上对一线城市房价各个影响因素的重要性进行排序,同时计算出各个影响因素的重要性大小及其影响方向,由此可见SHAP值可解释性方法适用于挖掘并解释一线城市房价的影响因素。SHAP值可解释性方法由Lundberg和Lee提出,其核心源于合作博弈论的概念Shapley值,对于每个预测样本,该方法都会产生一个预测值,该样本中每个特征所分配到的数值,即为SHAP值。SHAP值同时反映了各个影响因素对被解释变量的大小和方向,在本文是指各个影响因素对一线城市房价增速的影响大小和方向。

假设第i个样本为 $x_i$ , $x_i$ 的第j个影响因素为 $x_{ij}$ ,XGBoost方法对一线城市房价增速的预测值为 $y_i$ ,模型预测的参考值为 $y_0$ ,根据SHAP值可解释性方法,XGBoost方法在训练样本的预测值 $y_i$ 可以分解为:

$$y_i = y_0 + \sum_{j=1}^p f(x_{ij}) \quad (1)$$

其中,将一线城市房价增速均值(5%)作为预测参考值 $y_0$ , $f(x_{ij})$ 是 $x_{ij}$ 的SHAP值。以 $f(x_{i1})$ 为例, $f(x_{i1})$ 表示第i个样本中第1个特征对最终预测值 $y_i$ 的贡献值。 $\sum_{j=1}^p f(x_{ij}) < 0 (> 0)$ 表明,所有因素共同推动一线城市房价下跌(上涨)。具体关于第i个样本的第j个特征 $x_{ij}$ 的SHAP值 $f(x_{ij})$ ,本

① 限于篇幅,相关结果留存备索。

② Chen Tianqi and Carlos Guestrin, "Xgboost: A Scalable Tree Boosting System," *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp.785-794.

③ 变量替换法基于变量间相互独立的假设,当变量之间存在相关性时,变量替换法的结果将存在一定偏差。LIME方法使用线性回归拟合机器学习在样本局部的预测值,从而在局部上解释机器学习模型的预测机制,但由于使用不同样本的回归模型参数不具有可比性,使得LIME方法无法用于比较不同样本上变量重要性的差异,因此不适用于本文研究。

文参考 Aas 等的改进做法予以测算。<sup>①</sup>测算过程如下:

假定  $\Omega$  为由  $p$  个房价影响因素组成的变量全集,  $U$  为不包含第  $j$  个房价影响因素的变量集, 从全集  $\Omega$  中抽取  $|U|$  个不包含  $\{j\}$  的因素, 一共有  $\frac{|\Omega|!}{|U|!(|\Omega|-|U|-1)!}$  种组合方式, 基于此构造出用于模拟计算的样本  $x_i^u$ 。具体而言, 样本  $x_i$  中的变量按照变量是否在  $U$  中被分成两部分: 一部分是属于变量集  $U$  的房价影响因素的观测值, 保留了其在样本  $x_i$  中的观测值; 另一部分是不属于集合  $U$  的其他房价影响因素, 它们的样本观测值是从样本  $x_i$  之外的训练样本中随机抽取而得, 所以  $f(x_i^{U \cup \{j\}}) - f(x_i^U)$  表示样本  $x_i$  中第  $j$  个特征的观测值被随机替换后, 样本  $x_i$  的预测结果的变化。通过构造不同的变量组合  $U$  以及对应的样本  $x_i^{U \cup \{j\}}$  和  $x_i^U$ , 便可以近似地模拟估计出第  $j$  个影响因素对房价增速预测值  $y_i$  的贡献大小。最后, 将第  $j$  个影响因素在所有变量组合  $U$  下的变化值进行加权平均, 即可得到第  $j$  个因素对一线城市房价增速的影响大小:

$$f(x_{ij}) = \frac{|U|!(|\Omega|-|U|-1)!}{|\Omega|!} \frac{1}{N} \sum_{n=1}^N (f(x_i^{U \cup \{j\}}) - f(x_i^U)) \quad (2)$$

由此可见, 相较于 XGBoost 方法输出的特征重要性排序, SHAP 值清晰地展现了每个样本中各个影响因素的贡献率大小, 并且反映了每个样本中各个因素对房价的作用方向。

### 3. 研究思路与测算过程

基于上述建模基础, 本文将 XGBoost 方法和改进后的 SHAP 值可解释性方法相结合, 首先计算所有一线城市房价上涨的主要影响因素, 并在此基础上挖掘每一个一线城市房价上涨背后的主要驱动因素。具体的研究思路如下。

第一, 使用预处理后四座一线城市的全部样本训练出最优超参数下的 XGBoost 模型, 通过 Aas 等改进的 SHAP 值可解释性方法计算一线城市样本中各个房价影响因素的 SHAP 值大小, 这些 SHAP 值反映了不同时期各个因素对一线城市房价增速的影响大小。

第二, 根据一线城市的平均房价增速是否大于 5%, 划分出一线城市整体房价上涨的区间。本文选取 5% 为房价上涨阈值主要基于以下两点考虑: 一是 SHAP 值可解释性方法一般将基准值设定为变量均值, 上涨阈值与 SHAP 值基准值一致能确保结果的解释性。二是参考陈小亮和李诚浩<sup>②</sup>对房价上涨轮次划分标准, 房价涨幅超过 5% 的城市具有明显的持续上涨迹象, 从以往的房价数据也印证了该标准。在一线城市房价超过 5% 的每一轮时期, 其均价都处于 10% 的高位, 并且持续时间都在半年以上。此外, 需要说明的是, 本文在训练 XGBoost 模型中使用了所有的一线城市样本, 而在分析中着重于分析房价上涨区间的影响因素, 一方面是因为在样本期内的大多数时间里, 一线城市的房价均处于上涨状态, 上涨区间的分析具有较强的代表性; 另一方面房价下跌时期的观测样本较少, 且大部分样本的房价增速也在 0 附近, 会因此导致蒙特卡洛计算的结果不稳定和贡献率计算值的异常。

第三, 借鉴 VAR 模型中方差分解的思想, 本文对房价上涨时期各因素的 SHAP 值进行平均和加总, 以此分析一线城市房价上涨的主要影响因素。类似地, 可以分别计算四座一线城市房价影响因素的平均 SHAP 值, 从而计算各个影响因素对四座一线城市房价上涨的贡献大小。由此, 可以对比分析整体上一线城市房价上涨的影响因素和各个一线城市房价上涨的影响因素的异同。

<sup>①</sup> Aas 等提出在模拟抽样时将高斯核函数作为抽样样本与样本  $x_i$  之间的相似性度量, 以此为权重加权各个模拟样本的结果, 进而考虑不同变量相关性对结果的影响, 打破了 Štrumbelj 和 Kononenko 要求 SHAP 值可解释性方法影响因素之间相互独立的假设带来的限制, 因此本文采用 Aas 等改进的 SHAP 值可解释性方法。

<sup>②</sup> 陈小亮、李诚浩:《防范房价泡沫化风险的对策研究:“疏堵结合”的新思路》,《人文杂志》2022 年第 2 期, 第 26—38 页。

## 四、指标体系与数据说明

### (一)房价增速的测度指标

关于城市房价增速的测度指标大体可以分为三类:一是以商品房销售额除以各市商品房销售面积测度城市房价,如倪鹏飞,褚敏等,邹琳华等。<sup>①</sup>二是直接使用国家统计局公布的大中城市的月度住宅销售价格指数测度城市房价增速,如洪勇,张明等。<sup>②</sup>三是通过其他网络信息平台挖掘的微观房价数据测度城市房价增速,如朱恺容等,方慧芬等。<sup>③</sup>考虑到数据的可得性和本文研究问题的需要,本文选择国家统计局每月公布的城市二手房住宅销售价格同比指数作为四大一线城市房价增速的测度指标。<sup>④</sup>综合考虑所有指标数据可得性,本文重点关注的2006年以来一线城市房价上涨的影响因素,并将样本期确定为2006年1月至2021年12月,数据频率为月度。在此基础上得到一线城市房价走势,详见图1。

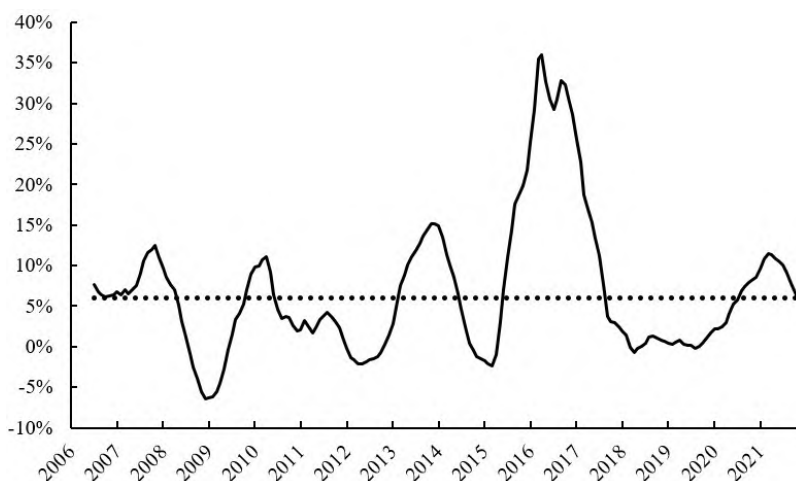


图1 一线城市平均房价增速变化

数据来源:国家统计局。

### (二)房价影响因素指标体系构建

基于宏观经济理论以及国内外相关研究,本文从供给因素、需求因素、政策因素和预期因素四大类别出发,构建了一线城市房价及其影响因素的指标体系。其中,供给因素和需求因素是影响房价走势的基本面因素,而且基本面因素和政策因素都会影响房价预期,进而影响各个城市的房价走势。

1. 供给因素。从供给视角出发,土地价格、房地产投资额、住宅用地的供应面积、竣工面积等

① 倪鹏飞:《货币政策宽松、供需空间错配与房价持续分化》,《经济研究》2019年第8期,第87—102页;褚敏、常钰:《城市高房价导致少儿抚养比下降了?》,《中国软科学》2022年第8期,第58—66页;邹琳华、钟春平:《饥饿供地,还是售地冲动——基于地级以上城市土地出让及房价数据的实证分析》,《财贸经济》2022年第3期,第82—97页。

② 洪勇:《中国区域房价联动度测度及其影响因素分析——基于35个大中城市面板数据的研究》,《管理评论》2020年第6期,第62—71页;张明、刘瑶:《中国城市房地产价格走势与波动的驱动因素探析——来自全国31个省份与70个大中城市的经验证据》,《南京社会科学》2021年第6期,第26—38页。

③ 朱恺容、李培、谢贞发:《房地产限购政策的有效性及其外部性评估》,《财贸经济》2019年第2期,第147—160页;方慧芬等:《中国城市房价对生育率的影响——基于长三角地区41个城市的计量分析》,《地理研究》2021年第9期,第2426—2441页。

④ 相较于新建住宅价格指数,二手房住宅价格指数的编制过程中综合考虑了多组数据源,并且兼顾住宅的地理位置编制价格指数,同时价格的环比指数比同比指数更易受到季节因素的影响,因此二手房住宅价格同比指数更能反映每个城市房价的变化。



供给因素都会对房价产生影响。综合考虑开发商和政府住宅用地的供给行为,本文选取单位土地价格同比增速、商品住宅房屋造价同比增速、住宅用地供应面积和商品房竣工面积衡量供给因素对一线城市房价的影响。

2. 需求因素。对一线城市而言,城市居民数量和年龄结构、收入水平以及城市的公共服务水平都会影响住房需求,进而对房价产生影响。其一,大量实证研究发现,人口规模和人口年龄结构都是房价的重要影响因素,如李嘉楠等,徐建炜等。<sup>①</sup>综合已有文献的做法,本文分别从人口密度、人口结构和人口流动等维度选取指标衡量一线城市人口因素对房价的影响。具体地,使用一线城市常住人口与城市建成区面积的比值表示常住人口密度,用人口自然增长率衡量人口年龄结构,用(常住人口—户籍人口)/常住人口衡量移民占比的涨幅表示人口流动的变化,如韩立彬和陆铭。<sup>②</sup>其二,居民收入水平的高低会影响住房需求进而影响房价,本文分别选取人均GDP增速和职工平均工资增速测度居民收入对一线城市房价的影响。其三,由于住房兼具商品和投资的双重属性,城市公共服务作为房地产附带价值的重要组成部分对房价具有一定影响。参考相关研究选取人均床位数、人均财政教育支出同比增速、人均道路面积和绿化覆盖率衡量四个指标测度公共服务水平对一线城市房价的影响,如汤玉刚等,刘诚等。<sup>③</sup>

3. 政策因素。货币政策会从住房贷款的可得性以及贷款成本高低等方面影响住房需求。由于货币政策由央行统一实施,因此各个城市的货币政策与全国相同。参考张清源等、张红和李洋的做法,分别选取全国个人住房贷款同比增速和全国住房贷款利率衡量央行的数量型和价格型货币政策。<sup>④</sup>此外,财政政策也是影响房价的重要因素,土地出让金在地方政府财政收入中的占比越高,地方政府越依赖房地产市场,那么财政政策对房价的影响也就越大。参考李英利的做法,选取各城市土地财政依赖度(地方政府土地出让金收入与一般公共预算收入的比例)和一般土地公共财政赤字率((财政支出—财政收入)/财政收入×100%)这两个指标衡量财政政策。<sup>⑤</sup>

4. 预期因素。预期会通过影响企业和居民行为,进而对经济运行产生影响。房地产因其生产周期长、流动性低等特征,更容易受到预期的影响。通过梳理相关文献可知,房价预期的测算方法主要有两类:一是基于适应性预期测算房价预期,如高波等,王先柱等<sup>⑥</sup>;二是基于问卷调查获取房价预期信息,如董纪昌等,孙伟增等<sup>⑦</sup>。考虑到数据可得性和研究需要,本文在基准分析中采用第一类做法,并参照高波等选取滞后一期的房价增速作为房价预期的测度指标。为了保证研究结果

① 李嘉楠、游伟翔、孙浦阳:《外来人口是否促进了城市房价上涨?——基于中国城市数据的实证研究》,《南开经济研究》2017年第1期,第58—76页;徐建炜、徐奇渊、何帆:《房价上涨背后的人口结构因素:国际经验与中国证据》,《世界经济》2012年第1期,第24—42页。

② 韩立彬、陆铭:《供需错配:解开中国房价分化之谜》,《世界经济》2018年第10期,第126—149页。

③ 汤玉刚、陈强、满利苹:《资本化、财政激励与地方公共服务提供——基于我国35个大中城市的实证分析》,《经济学(季刊)》2016年第1期,第217—240页;刘诚、杨继东:《土地策略性供给与房价分化》,《财经研究》2019年第4期,第68—82页。

④ 张清源、梁若冰、朱峰:《货币政策加剧城市房价的冷热不均吗》,《统计研究》2018年第5期,第75—87页;张红、李洋:《房地产市场对货币政策传导效应的区域差异研究——基于GVAR模型的实证分析》,《金融研究》2013年第2期,第114—128页。

⑤ 李英利:《财政压力、土地财政与区域房价的时空演化——基于GTWR模型的实证研究》,《财政研究》2020年第5期,第78—89页。城市层面的政策变量代表性会更强,但在实证中面临两大约束:一是城市层面房地产信贷情况难以获得;二是地方政府限购、公积金等政策披露的透明度、频率相对较低,政策力度大小难以量化。所以在评估政策因素影响时本文仅考虑了易于量化的财政政策,未考虑其他地方层面政策因素。

⑥ 高波、王辉龙、李伟军:《预期、投机与中国城市房价泡沫》,《金融研究》2014年第2期,第44—58页;王先柱、杨义武:《差异化预期、政策调控与房价波动——基于中国35个大中城市的实证研究》,《财经研究》2015年第12期,第51—61页。

⑦ 董纪昌等:《公众房价预期形成机理分析与实证研究——基于社会学习视角》,《管理评论》2020年第10期,第34—46页;孙伟增、林嘉瑜:《教育资源供给能够降低学区房溢价吗?——来自北京市新建小学的证据》,《经济学(季刊)》2020年第2期,第51—59页。

的稳健性,本文还将在稳健性检验中对房价适应性预期的测度方法进行更详细、更深入的探讨。

综上,为了研究一线城市房价上涨的影响因素,本文构建了包含四类因素共计18个指标的中等规模指标体系。值得说明的是,相较于多数使用机器学习方法研究所用的上百个指标的文献,本文的指标选择相对较少。之所以没有纳入更多指标,是因为过多的指标不仅会影响XGBoost方法和SHAP值可解释性方法的准确性(Aas等),同时变量过多会使得模型的可解释难度增大。因此,本文构建的指标体系既利用了机器学习的预测精度和SHAP值可解释性方法的优良性质,也尽可能全面地涵盖了房价的多种影响因素。

### (三)数据来源和预处理

本文所构建的一线城市房价影响因素的指标体系和数据来源参见表1。本文的数据预处理需要重点解决两个问题:第一,统一各个指标的数据频率。由于房价增速数据是月度数据,而指标体系中存在部分年度或季度指标,因此需要将相关指标转换为月度数据。参考张春华等的做法<sup>①</sup>,使用三次样条法对相关指标进行插值处理<sup>②</sup>。第二,补全数据缺失值。参考Stekhoven等的做法<sup>③</sup>,使用MissForest方法填补指标体系中人均GDP等指标的缺失值<sup>④</sup>。

表1 一线城市房价影响因素指标体系

因素类别	具体指标	频率	数据来源
供给因素	单位土地价格同比增速	月度	Wind中国土地大全
	商品住宅房屋造价同比增速	年度	国信房地产信息网
	住宅用地供应面积同比增速	月度	Wind中国土地大全
	商品房竣工面积同比增速	年度	国信房地产信息网
需求因素	常住人口密度增速	年度	《中国城市统计年鉴》
	人口自然增长率		
	移民占比增速		
	人均GDP增速		
	职工平均工资增速		
	人均床位数		
	人均财政教育支出同比增速		
	人均道路面积		
政策因素	全国个人住房贷款同比增速	季度	中国人民银行
	全国住房贷款利率	季度	中国人民银行
	各城市土地财政依赖度	年度	《中国城市统计年鉴》
	一般公共预算赤字率(滞后一期)	季度	财政部
预期因素	各城市的房价上涨预期	月度	笔者自行测算

经处理后,所有数据频率一致且无明显异常值,为了更全面地了解一线城市房价影响因素的

① 张春华、高铁梅、陈飞:《经济时间序列频率转换方法的研究与应用》,《统计研究》2017年第2期,第92—100页。

② 本文进行三次样条插值处理时考虑了存量和流量指标处理方式的差异,如移民占比等存量指标直接进行三次样条插值,而人均GDP、职工平均工资等流量指标先除以相应单位再进行三次样条插值。

③ Daniel J. Stekhoven & Peter Bühlmann, "MissForest—Non-parametric Missing Value Imputation for Mixed-type Data," *Bioinformatics*, Vol. 28, No. 1, 2012, pp.112-118.

④ MissForest方法借鉴机器学习的优势,运用随机森林回归方法,充分利用已知的样本信息来插补缺失值,因此具有较好的预测精度。例如,填补某城市人均GDP的缺失值时, MissForest会将所有城市人均GDP的观测值作为训练的因变量,将对应的其他城市特征作为自变量,训练模型对缺失值进行预测。

整体情况,表2给出了预处理后所有样本中各个房价影响因素指标的描述性统计分析结果。

表2 一线城市房价影响因素指标体系的描述性统计

指标	最小值	最大值	均值	标准差
单位土地价格同比增速(%)	-23.93	84.49	13.47	11.09
商品住宅房屋造价同比增速(%)	-23.00	54.00	7.47	10.70
住宅用地供应面积同比增速(%)	-48.00	91.00	0.45	27.50
商品房竣工面积同比增速(%)	-52.00	143.00	-1.70	22.14
常住人口密度增速(%)	-1.30	1.26	-0.86	0.36
人口自然增长率(‰)	-3.26	21.89	7.23	6.92
移民占比增速(%)	-2.08	5.83	0.28	0.81
人均GDP增速(%)	1.20	16.60	9.15	2.86
职工平均工资增速(%)	-5.00	32.00	10.96	4.37
人均床位数(张/人)	0.18	0.66	0.42	0.12
人均财政教育支出同比增速(%)	-13.00	58.00	17.60	12.02
人均道路面积(平方米/万人)	4.05	10.42	6.41	1.93
绿化覆盖率(%)	33.36	49.13	42.76	3.92
全国个人住房贷款同比增速(%)	10.10	44.30	20.75	6.93
全国住房贷款利率(%)	4.65	7.83	5.97	0.99
各城市土地财政依赖度(%)	7.00	59.00	25.66	9.95
各城市一般公共财政赤字率(滞后一期)(%)	-0.23	5.47	2.27	1.21
各城市的房价上涨预期(%)	-18.07	55.52	6.73	10.11

## 五、主要实证结果分析

本文的研究目标是,探寻一线城市房价上涨的主要影响因素,为此将开展两方面主要研究工作。一是分析完整样本期内(2006—2021)一线城市房价上涨的主要影响因素,以及四大一线城市房价影响因素的异同,从而为“因城施策”提供决策参考。二是从动态视角分析2012年前后一线城市房价上涨的影响因素是否存在明显差别,从而为不同经济环境下的房价调控提供针对性的决策参考。

### (一)完整样本期内的测算结果和主要结论

在基准分析中,本文使用的是XGBoost方法,相关结果参见表3的Panel A。为了提高结果的稳健性,本文还使用Random Forest方法进行了分析,相关结果参见表3的Panel B。通过对比可知,使用XGBoost方法得到的测算结果与使用Random Forest方法得到的测算结果较为相似,无论是四大类影响因素的贡献率还是四大类影响因素的相对重要性排序,两种方法的测算结果均呈现出高度一致性。为了简化分析,下文主要基于XGBoost方法的测算结果展开研究,并且得到了三点主要结论。从表3的基准测算结果可知,对一线城市整体而言,预期因素是一线城市房价上涨的首要影响因素,在四大类因素中排名第一,而且其贡献率达到了45.0%。具体到每一个城市来看,北京、上海、深圳房价上涨的首要影响因素都是预期因素,其贡献率分别达到了53.6%、44.0%和45.7%。对广州而言,预期因素对房价的影响略有逊色,在四大类因素中排在第二位,不过其贡献率也达到了36.6%的相对较高水平。

第一,预期因素是一线城市房价上涨的首要影响因素,其贡献率达到了45%左右。

为了确保研究结论的稳健性和可靠性,按照高波等研究的做法,对房价适应性预期的测度进行了更丰富的尝试。具体而言,将房价适应性预期的计算公式设定为: $\dot{P}_{t+1}^e = a \cdot \dot{P}_t^e + (1 - a) \cdot \dot{P}_t$ ,其中 $\dot{P}_t^e$ 为 $t$ 时刻对房价增速的预期, $\dot{P}_t$ 为 $t$ 时刻的房价增速。本文将四大一线城市在2006年1月份的房价同比增速作为计算城市房价适应性预期的初始值,按照适应性预期的计算公式,迭代计算每个月四大一线城市房价适应性预期的具体数值。将权重 $a$ 进行不同赋值,即可测算不同适应性预期方式下房价的形成机制,进而分析预期因素对一线城市房价的影响。从表4呈现的结果来看,本文的实证结果对权重 $a$ 并不敏感,当权重 $a$ 分别赋值0.2、0.3和0.4之后得到的结果,与表3的基准结果较为一致,不管是从四大类因素对一线城市房价的贡献率还是重要性排序来看,预期因素都是一线城市房价上涨的首要影响因素,这在北京、上海和深圳体现得尤为突出。

表3 各类因素对一线城市房价上涨的贡献率及排序

因素类别	一线城市整体	北京	上海	广州	深圳
Panel A: 基于XGBoost方法的测算结果					
供给因素	24.6% (2)	21.6% (3)	38.2% (2)	8.5% (4)	30.0% (2)
需求因素	19.3% (3)	2.1% (4)	13.3% (3)	38.6% (1)	23.4% (3)
政策因素	11.1% (4)	22.6% (2)	4.6% (4)	16.3% (3)	0.9% (4)
预期因素	45.0% (1)	53.6% (1)	44.0% (1)	36.6% (2)	45.7% (1)
Panel B: 基于Random Forest方法的测算结果					
供给因素	27.5% (2)	20.1% (3)	38.3% (2)	17.4% (3)	33.8% (2)
需求因素	21.4% (3)	12.2% (4)	11.6% (3)	41.6% (1)	19.3% (3)
政策因素	7.6% (4)	26.2% (2)	9.4% (4)	1.9% (4)	-5.7% (4)
预期因素	43.5% (1)	41.6% (1)	40.7% (1)	39.0% (2)	52.6% (1)

注:表格中的百分比数字代表各类因素对房价上涨的贡献率,括号里的数字代表相应因素贡献率的排序。下表同。

第二,供给因素和需求因素这两大类基本面因素对一线城市房价上涨也起到了较为重要的作用,二者贡献率分别达到了25%左右和20%左右,合计达到了45%左右。

从表3可知,基于XGBoost方法的测算结果,供给因素和需求因素这两大类基本面因素对一线城市房价上涨的贡献率之和达到了43.9%。其中,基本面因素对上海和深圳房价上涨的贡献率分别达到了51.5%和53.4%,对广州房价上涨的贡献率也达到了47.1%。相比之下,基本面因素对北京房价上涨的贡献率偏低,只有23.7%。究其原因,预期因素对北京房价上涨的贡献率更高,从而使得基本面因素的贡献率偏低。

就供给因素而言,根据表3中XGBoost方法的测算结果可知,对一线城市整体而言,供给因素

的贡献率达到了24.6%,在四大类因素中排名第二,仅次于预期因素。具体到每一个城市来看,供给因素对上海、深圳和北京房价的贡献率分别达到了38.2%、30.0%和21.6%,贡献率相对较高。相比之下,供给因素对广州房价的贡献率相对偏低,只有8.5%,而且在四大类因素中排名最后。

表4 预期因素对一线城市房价上涨的贡献率及排序

因素类别	一线城市整体	北京	上海	广州	深圳
Panel A: 权重 $\alpha=0.4$					
预期因素	50.7% (1)	50.3% (1)	72.3% (1)	36.5% (1)	43.9% (1)
Panel B: 权重 $\alpha=0.3$					
预期因素	44.5% (1)	52.3% (1)	44.1% (1)	37.5% (1)	43.9% (1)
Panel C: 权重 $\alpha=0.2$					
预期因素	41.3% (1)	41.9% (1)	42.3% (1)	36.1% (2)	44.9% (1)

就需求因素而言,通过表3中XGBoost方法的测算结果可以看出,对一线城市整体而言,需求因素的贡献率为19.3%,低于供给因素和预期因素的贡献率,位列第三。从四个城市的具体情况来看,需求因素对深圳、上海和北京房价上涨的贡献率分别为23.4%、13.3%和2.1%,在四大类因素中分别排名第二、第三和第四。与之明显不同的是,需求因素对广州房价上涨的贡献率达到了38.6%,位列四大类因素之首。

不难发现,基本面因素对四大一线城市房价上涨的影响既有一定的共性,又呈现出一些不同之处。以广州为例,广州房价上涨期间,更多是因为需求因素所导致的,供给因素的贡献率则相对偏低。要想进一步探寻其背后的原因,需要对本文指标体系中每一个供给因素和每一个需求因素的贡献率加以测算和分析,详见表5和表6。

从表5可以看出,四个城市的住宅用地供应面积增速存在较大差异,从而导致单位土地价格同比增速存在较大差异,最终使得供给因素对房价增速的影响存在差异。具体而言,广州的住宅用地供应面积增速最高,相应地广州的单位土地价格增速处于偏低水平,二者共同作用使得供给因素对房价上涨的贡献率处于低位。从样本期内(房价上涨时期,下同)的实际数据看,广州住宅用地供应面积同比增速平均值达到了15.0%,深圳为3.4%、北京和上海分别为-4.4%和-7.6%。在较高的住宅用地供应面积增速的支撑下,广州单位土地价格同比增速处于13.8%的相对低位,明显低于深圳的28.1%,也低于北京的15.4%,与上海基本持平。总体而言,广州的土地和住房供给相对充裕,进而使得房价上涨趋势相对平缓,这是供给因素对广州房价上涨贡献率相对偏低的主要原因。

表5 供给因素对一线城市房价上涨的贡献率

影响因素	一线城市整体	北京	上海	广州	深圳
供给因素(合计)	24.6%	21.6%	38.2%	8.5%	30.0%
单位土地价格同比增速	9.4%	9.1%	15.4%	-3.4%	16.3%
商品住宅房屋造价同比增速	4.5%	1.8%	9.6%	3.6%	3.3%
住宅用地供应面积	6.2%	4.5%	11.7%	1.6%	6.8%
商品房竣工面积	4.5%	6.2%	1.5%	6.7%	3.5%

从表6可知,需求因素当中的人口因素和公共服务因素对广州房价增速的贡献率较高,从而使需求因素的整体贡献率较高。其中,人口因素和公共服务因素对广州房价上涨的贡献率分别达到了16.0%和19.4%,比四大一线城市的整体水平高出了10.2个和10.9个百分点。进一步分析数据可知,就人口因素而言,样本期内广州的常驻人口密度是最高的,平均值达到了0.75万人/平方公里,此外广州的人口自然增长率和移民占比增速也处于较高水平,高于北京和上海,仅次于深圳,从而使得人口因素对广州房价上涨的贡献率较高。就公共服务因素而言,样本期内广州的公共服务水平得到显著提升,绿化覆盖率从36.6%提升到45.5%,提升了9个百分点,是四个城市中涨幅最大的;此外,广州的人均道路面积从8.72万平方米提升到10万平方米以上,而北京和上海稳定在5万平方米左右,深圳从9万平方米左右降到了7万平方米以下。公共服务水平的改善提升了广州住房的吸引力,更显著地促进了广东房价的上涨趋势。

表6 需求因素对一线城市房价上涨的贡献率

影响因素	一线城市整体	北京	上海	广州	深圳
需求因素(合计)	19.3%	2.1%	13.3%	38.6%	23.4%
人口因素	5.8%	-4.3%	6.8%	16.0%	5.0%
收入因素	4.9%	5.3%	9.1%	3.3%	1.8%
公共服务因素	8.5%	1.1%	-2.7%	19.4%	16.6%

注:人口因素包括常住人口密度、人口自然增长率和移民占比增速三个指标,收入因素包括人均GDP增速和职工平均工资增速两个指标,公共服务因素包括人均床位数、人均财政教育支出同比增速、人均道路面积和绿化覆盖率四个指标。

第三,货币政策等政策因素对一线城市房价上涨的影响相对偏弱,其贡献率仅为10%左右。

通过表3所展示的基准结果可知,对一线城市整体而言,基于XGBoost方法的测算结果显示政策因素的贡献率为11.1%,基于Random Forest方法的测算结果显示政策因素的贡献率仅为7.6%,在四类因素中排名末尾。具体到四个一线城市来看,以XGBoost方法的测算结果为例,政策因素对上海和深圳房价上涨的贡献率分别仅为4.6%和0.9%,在四大类因素中均排名末尾。政策因素对广州房价上涨的贡献率为16.3%,在四大类因素中排名第三。由此可以初步发现,政策因素对一线城市房价上涨的整体影响相对偏弱。

考虑到宏观政策从出台到产生作用效果存在一定的时滞,因此本文进一步分析了将政策因素滞后3个月和6个月的结果,详见表7的Panel B和Panel C。通过将表7的Panel B、Panel C的结果与Panel A对比可知,在考虑了政策时滞之后,政策因素对一线城市房价上涨的贡献率同样处于较低水平,由此表明本文基准结果是较为稳健的。需要注意的是,虽然政策因素对四大一线城市房价上涨的直接影响相对偏低,但是可能会通过影响居民对房价的预期进而间接影响房价,因此对于政策因素尤其是货币政策因素对房价的影响仍要重视。

2012年前后中国的经济发展态势和房地产市场发展态势均发生了较为显著的变化。就经济发展态势而言,2012年之后中国经济进入新常态,经济从高速增长阶段向中高速增长阶段换挡。就房地产市场而言,2012年之后一线城市与二三线城市房价分化态势愈发明显。<sup>①</sup>在此背景下,一线城市房价上涨的影响因素也很可能发生了重要变化。本部分将从动态视角进行分析,以探寻2012年前后一线城市房价上涨影响因素的动态变化。SHAP值方法等机器学习解释性方法在动态分析方面具有显著优势,从而为本文分析一线城市房价影响因素在不同时期的动态变化提供了方

① 陈小亮、刘哲希、陈彦斌:《“三重压力”下宏观政策的效应评估——宏观政策评价报告2023》,《改革》2023年第1期,第51—71页。

法上的可行性,具体分析结果见表8。

表7 政策因素对一线城市房价上涨的贡献率

影响因素	一线城市整体	北京	上海	广州	深圳
Panel A:不考虑政策时滞					
政策因素(合计)	11.1%	22.6%	4.6%	16.3%	0.9%
货币政策	4.2%	11.9%	4.3%	4.8%	-4.3%
财政政策	6.9%	10.7%	0.3%	11.5%	5.1%
Panel B:考虑政策时滞(滞后3个月)					
政策因素(合计)	9.8%	19.3%	6.0%	13.5%	-0.1%
货币政策	3.8%	9.4%	5.9%	4.1%	-4.4%
财政政策	6.0%	9.9%	0.1%	9.4%	4.4%
Panel C:考虑政策时滞(滞后6个月)					
政策因素(合计)	15.3%	31.2%	11.9%	15.1%	3.0%
货币政策	10.3%	22.7%	10.9%	6.7%	0.8%
财政政策	5.0%	8.5%	1.0%	8.4%	2.2%

表8 2012年前后各类因素对一线城市房价上涨的贡献率对比

	一线城市整体	北京	上海	广州	深圳
Panel A:2012年之前的测算结果					
供给因素	23.9% (3)	31.5% (3)	28.4% (2)	-6.9% (4)	41.8% (1)
需求因素	25.7% (2)	-8.9% (4)	22.3% (3)	62.8% (1)	27.5% (2)
政策因素	17.8% (4)	33.6% (2)	2.5% (4)	17.5% (3)	17.1% (3)
预期因素	32.6% (1)	43.8% (1)	46.9% (1)	26.5% (2)	13.6% (4)
Panel B:2012年之后的测算结果					
供给因素	25.0% (2)	15.5% (3)	44.2% (1)	18.0% (3)	22.4% (2)
需求因素	15.3% (3)	9.0% (4)	7.8% (3)	23.8% (2)	20.8% (3)
政策因素	7.0% (4)	15.7% (2)	5.9% (4)	15.5% (4)	-9.7% (4)
预期因素	52.8% (1)	59.8% (1)	42.2% (2)	42.7% (1)	66.4% (1)

第一,预期因素对一线城市房价上涨的贡献率在2012年之前就达到了30%以上并且排名首位,2012年之后其贡献率进一步大幅升高至50%以上。

表8的结果显示,对一线城市整体而言,2012年之前预期因素对房价上涨的贡献率为32.6%,2012年之后则大幅提升到52.8%。其中,预期因素对深圳房价上涨的贡献率从13.6%提高到了66.4%,提升幅度达到了52.8个百分点,在四大类因素中的排名也从第四名提升到了第一名。此

外,预期因素对北京和广州房价上涨的贡献率也分别提高了16个和16.2个百分点。对上海而言,2012年之后经济结构调整促使上海市产业升级和创新发展,引入了更多高附加值的产业和技术创新,进而提高了供给的质量和数量。这种供给因素的增加对于房地产市场的供应带来积极的影响,从而减轻了房价上涨的预期压力。从结果来看,虽然预期因素对上海房价的贡献率略有下降,从2012年之前的46.9%小幅下降到了2012年之后的42.2%,但是仍然处于相对高位。

总体来看,预期因素对四大一线城市房价上涨的影响在2012年之后整体上呈现上升趋势。究其原因,从历史经验来看,当经济面临下行压力时,地方政府通常“将房地产作为稳增长的工具”,1998年亚洲金融危机和2008年全球金融危机等时期均为如此,2012年之后中国经济增速持续放缓,地方政府很可能再次“将房地产作为稳增长的工具”,由此便强化了一线城市房价上涨的预期。值得注意的是,2012年之前广州的需求因素和深圳的供给因素对房价增速影响更明显,使得2012年之前广州和深圳的预期因素贡献率相对较低,但是北京和上海预期因素方面的影响较大,综合排名后预期因素仍然占据了重要位置。

第二,供给因素与需求因素这两大类基本面因素对一线城市房价上涨的贡献率之和有所下降,从2012年之前的50%左右下降到了2012年之后的40%左右。

从表8可以看出,供给因素与需求因素这两大类基本面因素对一线城市房价上涨的贡献率从2012年之前的49.6%下降到了2012年之后的40.3%。需要注意的是,对于不同城市而言,基本面因素贡献率的下降有不同的表现。对北京而言,主要表现为供给因素贡献率的下降,2012年之前供给因素对北京房价上涨的贡献率为31.5%,但是到了2012年之后供给因素的贡献率降至15.5%。对于上海和广州而言,主要表现为需求因素贡献率的下降。2012年之前,需求因素对上海和广州房价上涨的贡献率分别为22.3%和62.8%,2012年之后需求因素的贡献率则分别降至7.8%和23.8%。广州是中国南方重要的经济中心之一,吸引了大量的人口流入和外来投资。在2012年之前,广州的人口规模快速增长,城市化进程迅猛,这导致了强劲的房地产需求。然而,在2012年之后,当地政府加强了调控措施,限制了需求因素对房价的影响,从而降低了需求因素的贡献率。对于深圳而言,则表现为供给因素和需求因素贡献率的双双下降。2012年之前,深圳受到特区经济政策的影响,吸引了大量的外来投资和企业进驻,促进了深圳的经济快速发展和城市化进程,但也导致供给相对紧张。然而,随着城市化进程逐渐深入,人口流入和经济增长速度相对减缓,房地产需求也相对减少。其中,供给因素的贡献率从2012年之前的41.8%下降到了2012年之后的22.4%,需求因素的贡献率从2012年之前的27.5%下降到了2012年之后的20.8%。

第三,政策因素对一线城市房价上涨的贡献率在2012年之前就相对偏弱,2012年之后进一步降至10%以下。从表8可知,2012年之前政策因素对一线城市房价上涨的贡献率为17.8%,在四大类因素中排名末尾。2012年之后,政策因素对一线城市房价上涨的贡献率进一步下降到了7.0%,在四大类因素中继续排名垫底。表9进一步展示了货币政策和财政政策各自的贡献率,从结果可知,不管是货币政策还是财政政策,它们对一线城市房价上涨的贡献率在2012年之后均出现了下降趋势,而且货币政策贡献率的下降幅度更大。具体而言,货币政策的贡献率从9.3%下降到了1.1%,财政政策的贡献率从8.6%下降到了5.9%。从四个城市的详细情况来看,货币政策对北京、上海和深圳房价上涨的贡献率均出现了较大幅度的下滑,广州则主要是因为财政政策的贡献率出现了大幅下降进而使得政策因素对房价上涨的贡献率有所下滑。需要再次说明的是,本文测算的结果主要呈现的是政策因素对房价的直接影响。从现实情况来看,政策因素的变化还会通过影响居民预期进而影响房价走势,因此不应忽视政策因素对房价上涨的影响。



表9 2012年前后政策因素对一线城市房价上涨的贡献率对比

	一线城市整体	北京	上海	广州	深圳
Panel A: 2012年之前的测算结果					
政策因素(合计)	17.8%	33.6%	2.5%	17.5%	17.1%
货币政策	9.3%	16.6%	8.1%	-1.7%	13.5%
财政政策	8.6%	17.0%	-5.6%	19.2%	3.6%
Panel B: 2012年之后的测算结果					
政策因素(合计)	7.0%	15.7%	5.9%	15.5%	-9.7%
货币政策	1.1%	9.0%	1.9%	8.8%	-15.8%
财政政策	5.9%	6.7%	3.9%	6.8%	6.2%

## 六、结论与政策建议

本文基于2006—2021年四大一线城市的月度数据,综合使用SHAP值可解释性方法和XG-Boost等机器学习方法,深入剖析了四大一线城市房价的主要影响因素及其动态变化。研究结果表明,四个一线城市房价的影响因素既存在一定的差异之处,也存在一定的共同特征,总体而言可以得到如下三点研究结论。第一,预期因素是一线城市房价上涨的主要影响因素,并且其影响在不断增强。预期因素对一线城市房价上涨的贡献率在2012年前就达到了30%以上并且排名首位,2012年之后其贡献率进一步大幅升高至50%以上并且继续排名首位。第二,供给因素和需求因素这两大类基本面因素对一线城市房价上涨也起到了较为重要的作用,不过其作用呈现出减弱态势。两大类基本面因素对一线城市房价上涨的贡献率在2012年之前达到了49.6%,但是到2012年之后下降到了40.3%。第三,货币政策等政策因素对一线城市房价上涨的贡献率相对偏弱,并且近年来其影响进一步下降。2012年之前政策因素对一线城市房价上涨的贡献率为17.8%,在四大类因素中排名末尾。2012年之后,政策因素对一线城市房价上涨的贡献率进一步下降到了7.0%,在四大类因素中继续排名垫底。需要注意的是,本文测算的主要是政策因素对房价的直接影响,但同时也应该认识到政策因素还会影响居民预期进而影响房价走势,因此不应忽视政策因素对房价的影响。

本文的研究结论具有较为重要的政策参考价值。考虑到预期因素是一线城市房价上涨的最主要因素而且近年来其贡献率有所增强,因此对一线城市而言,稳房价的关键在于稳预期。进一步地,结合本文实证结果可知,要想切实稳定一线城市房地产市场的预期,需要让房价更多地由基本面因素来决定,并且通过稳定房地产调控政策来稳预期,具体如下。

第一,从供给端发力,构建起一线城市土地供给与房价以及土地供给与常住人口之间的联动机制。由于一线城市房价上涨较快、常住人口较多,要相应增加住宅用地的供给,保证一线城市的住宅用地供应量能够与人口增长相适应。需要强调的是,在构建土地供应与常住人口之间的联动机制时,不能简单地认为一个城市常住人口数量增长放缓时就应该同步减少住宅用地供应量,还应该深入分析和测算该城市已有的常住人口中是否存在较大的住房供需缺口。如果尚且存在较大的住房供需缺口,那么即便新增常住人口数量减少,也仍然应该按实际需求增加住宅用地供应量,从而有效解决历史上形成的住房供需缺口。

第二,从需求端发力,降低一线城市的外来人口压力所引发的住房需求。进一步增强不同城市在教育、医疗、养老、环境保护等方面的基本公共服务均等化,尤其要缩小一线与其周边临近城

市以及其他三四线城市之间的公共服务差距。这样既有助于构建城市群,又有助于提高三四线城市对人口的吸引力,从而切实减轻一线城市的外来人口压力以及由此引发的住房需求增长。

第三,保持房地产金融政策的连续性、一致性、稳定性,通过稳定政策来稳定房地产市场的预期。近年来,中央多次在重要会议和文件中提及要“保持房地产金融政策的连续性、一致性、稳定性”,并且强调要“稳地价、稳房价、稳预期”,其本质都是希望通过稳定政策来稳定房价预期。需要强调的是,近年来在经济增速持续放缓、经济下行压力持续存在的大环境下,尤其要“保持房地产金融政策的连续性、一致性、稳定性”,避免地方政府再度将房地产作为“稳增长”的手段和工具,这样才能切实稳定一线城市的房价预期。

## Research on the Factors Affecting Housing Prices in First-tier Cities Based on Machine Learning Methods

Chen Xiaoliang, Cheng Shuo, Chen Kan, Xiao Zhengyan

**Abstract:** The real estate market in first-tier cities holds a crucial position in the national real estate market. To promote stable nationwide housing price growth and maintain the stability of housing prices in first-tier cities is of paramount importance. This study combines machine learning methods such as XGBoost with interpretability methods like SHAP values to calculate and analyze the main factors affecting housing prices in the four major first-tier cities and their dynamic changes before and after 2012. Three key conclusions have been drawn: 1) Anticipatory factors are the primary drivers of rising housing prices in first-tier cities, and their influence continues to strengthen; 2) Supply and demand factors, the two major categories of fundamental factors, also play a significant role in driving up housing prices in first-tier cities, although their impact is diminishing; and 3) Policy factors such as monetary policy have a relatively weak impact on the rising housing prices in first-tier cities, and this impact has further declined in recent years. Considering that anticipatory factors are the most significant drivers of housing price increases in first-tier cities, the key to stabilizing housing prices in these cities lies in stabilizing expectations. Further analysis, in conjunction with the empirical results in this study, suggests the need for housing prices to be determined more by fundamental factors. This can be achieved by stabilizing real estate regulation policies to manage expectations. Three key recommendations are: 1) Focus on the supply side by establishing a mechanism linking land supply and housing prices, as well as land supply and the resident population in first-tier cities; 2) Address the demand side by reducing the gap in public services between first-tier cities and their surrounding cities and other tier-three and tier-four cities. This can alleviate the pressure on first-tier cities from incoming populations and the resulting increase in housing demand; and 3) Maintain the continuity, consistency, and stability of real estate regulation policies to stabilize expectations in the real estate market.

**Keywords:** First-tier Cities; House Price; Stabilizing Expectations; Machine Learning Methods

【责任编辑:张栋豪】