分类号_____　　密级 _____
UDC _____　　编号 10741

硕士学位论文

(专业学位)

论文题目 基于优化随机森林模型的兰州市住宅类房产税税基评估研究

研究生姓名：袁晓芳

指导教师姓名、职称：石志恒 教授　邢铭刚 注册资产评估师

学科、专业名称：资产评估硕士

研究方向：房地产估价师

提交日期：2022 年 6 月 1 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：袁晓芳　　签字日期：2022.6.2

导师签名：石志恒　　签字日期：2022.6.5

导师(校外)签名：邢红研　　签字日期：2022.6.7

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：袁晓芳　　签字日期：2022.6.2

导师签名：石志恒　　签字日期：2022.6.5

导师(校外)签名：邢红研　　签字日期：2022.6.7

Research on the Tax Base Assessment of Residential Property Tax in Lanzhou City based on the Optimized Random Forest Model

Candidate : Yuan Xiaofang

Supervisor: Shi Zhiheng Xing Minggang

摘 要

近年来，房地产行业飞速发展，房价市场也在不断改变。希望进一步促进我国房地产市场的发展，中国对房产税税收作出了大量调整，并积极推进了房产税改革。党的十九届五中全会报告指出，要完善现在税收制度，完善地方税、直接税的体制。中央政府将积极稳妥推动房产税立法与改革。可见，征收房产税的大趋势所在。而未来征收房产税的征收过程中，由于可能存在因买卖“阴阳合同”而产生的巨大税务风险，通过对住宅的计税价格评估，将能够为房产税税基评估工作提供理论、技术支撑，进而促进房地产行业的健康发展，并保障我国财政收入的平稳进行。因此，本文探索兰州市房产税税基价格的评估方法，评估出兰州市房产税税基价格，希望能为房产税税基评估提供理论依据，同时为税基评估、房地产评估提供价值参考。

本文重点介绍了房产税、税基评估、批量评估、税基评估理论、随机森林模型、随机森林优化方法的有关概念和理论基础，并根据国内外部分国家和地区对房产税税基批量评估的实际经验，探讨了随机森林模型在税基批量评估中的适用性分析，以及兰州市房产税税基评估应用优化随机森林模型的可行性分析。接着以特征价格理论、地租论、区位理论为理论基础,并借鉴了以前学术界关于影响房地产价值特征变量的研究成果,在此基础上建立了税基估价指标体系,并对 22 个指标进行量化,最后建立了随机森林模型、优化随机森林模型、多元回归模型，利用 670 套住宅类房屋样本，对三个模型进行检验，利用最佳模型预测房屋市场价值，并计算出兰州市住宅类房产税税基。案例结果表明：①优化随机森林模型可应用于兰州市房产税税基评估中。②turnRF()函数能有效提高随机森林模型的预测准确性。③预测结果 1.327 万元/平方米可用于最近年度兰州市征收房产税的计税依据。基于此，可以考虑将评估出的均价和优化随机森林模型纳入到房产税征收与评估体系中。

关键词：兰州市 房产税 税基评估 批量评估 优化随机森林模型

Abstract

In recent years, the real estate industry has developed rapidly, and the housing market is constantly changing. Hoping to further promote the development of my country's real estate market, China has made a lot of adjustments to the property tax and actively promoted the reform of the property tax. The report of the Fifth Plenary Session of the 19th Central Committee of the Communist Party of China pointed out that it is necessary to improve the current tax system, and improve the system of local tax and direct tax. The central government will actively and steadily promote the legislation and reform of property tax. It can be seen that the general trend of property tax collection lies. In the process of collecting property tax in the future, since there may be huge tax risks arising from buying and selling "yin and yang contracts", the assessment of the taxable price of residences will be able to provide theoretical and technical support for the assessment of the tax base of property tax. In order to promote the healthy development of the real estate industry, and ensure the smooth progress of my country's fiscal revenue. Therefore, this paper explores the evaluation method of the real estate tax base price in Lanzhou, and evaluates the real estate tax base price in Lanzhou, hoping to provide a theoretical basis for the real estate tax base assessment, and at the same time provide a value reference for the tax base evaluation and

real estate evaluation.

This paper focuses on the related concepts and theoretical basis of property tax, tax base assessment, batch assessment, tax base assessment theory, random forest model, and random forest optimization method. Practical experience, discusses the applicability analysis of random forest model in batch assessment of tax base, and the feasibility analysis of the application of optimization random forest model in Lanzhou property tax base assessment. Then, based on hedonic price theory, land rent theory, and location theory, and drawing on previous academic research results on characteristic variables affecting real estate value, a tax-based valuation index system was established, and22indicators were quantified. , Finally, a random forest model, an optimized random forest model, and a multiple regression model were established. 670 residential housing samples were used to test the three models. The best model was used to predict the housing market value, and the residential property tax in Lanzhou City was calculated. tax base. The case results show that: ①The optimized random forest model can be applied to the assessment of the property tax base in Lanzhou. ②The turnRF() function can effectively improve the prediction accuracy of the random forest model. ③The forecast result of 13,270 RMB/square meter can be used as the tax basis for the property tax levied by Lanzhou in the recent year. Based on this, it can be considered to incorporate the estimated average price and the

optimized random forest model into the property tax collection and evaluation system.

Keywords：Lanzhou; Real Estate Tax; Tax base assessment; Batch evaluation; Optimize the random forest model

目 录

1 绪论

1.1 研究背景

近年来，房地产市场日趋活跃，伴随着房产价格的不断上涨，政府出台相关政策控制房价上涨，房产税作为近些年来控制房价的税收工具，对房产税的相关研究变得更加紧迫。我国从首次提出房产税到随后的不断试点改革，政府对房产税征收投入了大量精力。国务院办公厅在 2013 年 2 月，提出要推进房产税改革工作，扩大房产税试点区域的计划；接着房产税试点区域扩大，试点城市也在不断增加，房产税已明确成为控制增量住房房价的有效工具。在 2014 年 4 月份的博鳌亚洲论坛年会上，讨论了楼市的改革和调整，财政部经济科学研究院的处长贾康，提出了房屋变革不仅是要注重保障，也要完善房产税制度立法的新观念。2020 年的《政府工作报告》中明确提出了，在 2021 年，我国在财税、金融有关的体制中需要进行深入改革，并且这些改革是 2021 年政府的重点工作。到 2020 年，试点城市上海颁布了关于缴纳房产税试点问题的公告，该公告对征收面积、计税价格核定以及免征房产税等问题作出了具体通知，并要求于 2021 年 1 月 28 日实施。2021 年 11 月,上海市出台《关于缴纳 2021 年度个人住房房产税有关事项提示》,明确规定了个人房产税征收方式以及缴纳流程。中央政府每年都会对税种作一些调整甚至撤销之类的工作,但却未有任何一项税种的调整像房产税改革这种改革周期之长,且程序繁琐,由此可见,房产税改革困难度颇大,对房产税改革的深入研究已迫在眉睫。

在房产税征收管理过程中，最关键的就是对房屋计税价值进行评估，即房产税税基评估（张辉，2021）。在房产税现实的交易过程中，部分纳税人故意报少价值而实现避税的目的，各地税务机关为了保障纳税人的合法权益，会有与其联合的专门评估组织辅助税收机关开展评价工作（陈钊，2015），这种方式在现实征收程序中面临如下三方面主要问题：合规性差，该问题一般是由于单宗评估方法本身的特性所决定的，单宗评估方法更多地依靠估价者本身对专业知识的把握，评估人员的主观臆断对税基结果的科学客观和准确度有较大的负面影响；评估成本高，单宗评估的表现形式确定了评价的有效性相对比较低，因而生产成本也就

相对比较高，无论是消费者本身抑或税务机关都承受了高额的评价成本费（玄永生等，2011）；依据欠缺，目前我国还没有制定官方规范房产税纳税评估的组织，该估价是否由第三者完成，而作为征收方，地方政府也在其中面临着相应的司法风险（傅樵，2017）。另外在房产税征收的过程中，可能存在“阴阳合同”，从而造成税收不公平的情况（张辉，2021）。在这些背景下，怎样通过房产市场的变化和实际交易的记录对征税对象作出迅速、精确的估价就变得格外关键和紧迫。

目前的房产税制改革，主要对房产征收结构中流通环节的征收方式提供了技术保障，而忽视了保有环节。这就形成了居民在持有住宅的成本低，部分居民购买多套房产，把房产当作投机工具，最终形成了房价升高，房屋空置率严重的局面（崔志坤等，2020）。考虑到上述情况，我国政府必须对房产交易市场实施有效控制，所以政府出台了房产税的相关政策，将房产税作为控制房屋保有环节房价的有力工具。征收房产税要求对当前大量的房屋作出快速、准确的估价，从目前的调研状况来看，批量评估能够适应这样的要求（陈艳，2019）。

根据上述背景，本文采用了机器学习中的随机森林算法，以确定适合于兰州市的房产税课税基准，并研究出适合兰州市房产税基的评估模型，为兰州市征收房产税前期准备奠定基础。

1.2 研究目的与意义

1.2.1 研究目的

（1）由于传统的房产税税基批量评估中需要耗费较多的成本及精力，并且步骤冗杂，准确性也难以保证，故以随机森林算法为评估工具，目的在于模拟验证随机森林算法适用于房产税税基批量评估的科学性，为兰州市评估出准确的房产税税基价格。

（2）我国的房产税税制改革不断深入，房产税相关规范不断完善，征收房产税已是必须趋势。兰州市作为我国西北地区重要的省会城市之一，房价一直呈现出增长的趋势，近些年来也不断有学者对兰州市房产税征收问题进行研究，未来兰州市房产税税基价格评估成为了亟待解决的问题，故本文以兰州市房产税征收为前提，目的在于探索出适合兰州市房产税税基批量评估的合适方法。

1.2.2 研究意义

从理论意义上来讲，房产税税基的批量评估其关键问题是房地产价格评估问题，国内外学者已经对房地产价格批量评估和房产税税基评估进行了相关的大量研究，这两方面理论已经相对成熟，而将随机森林模型应用到房产税税基评估中的研究相对较少，这些研究还都是处在介绍和探索阶段。我国房地产发展不平衡，对房产税税基评估研究都处在大型城市，如北京、上海、广州、重庆等，本文选择中等城市兰州作为研究对象，有利于拓展深化我国房产税税基评估理论。

从现实意义上来说，今年上海市将成为全国第一个房产税征收的城市，从长期上来讲，希望我国实现社会主义共同富裕，缴纳个人在保有环节上的房产税是社会发展趋势。在全国大量征收房产税的前提下，税基批量评估起到了重要作用，不仅可以提高评估效率，还可以降低人为干扰，得到符合客观规律的市场价格，为征税工作提供前期准备，从而保障我国的财政收入。本文从提高随机森林算法精确度出发，将优化模型应用于住宅类房产税税基批量评估中，可以提高评估结果的准确性，增强说服力，得到更加客观的市场价格，防止了由于“阴阳合同”带来的税收风险，降低房产税的征收难度，也为兰州市后续征收房产税提供借鉴和参考。

1.3 研究内容与研究方法

1.3.1 研究内容

我国不断出台房产税相关的改革政策，并且部分城市正有序推进房产税试点工作，关于房产税的研究成为了目前大多数居民、政府关注的内容，在征收房产税之前需要有真实的税基价格。基于此，确定了本文的研究思路，以下介绍了文章的研究内容和研究框架，文章的写作安排如下：

（1）第一部分是税基批量评估问题的提出。在我国不断改革房产税的前提背景下，分析了保有环节进行科学评估税基的必要性，随后提出本文如何正确、高效评估税基的研究问题。本文从房产税税基相关概念及理论、评估方法展开论述，讨论了房产税税基评估相关理论、批量评估特点和随机森林算法原理，并结

合前三者的研究推论出随机森林模型适合于批量评估的规律。

（2）第二部分是房产税税基评估模型分析。以税基评估作为本文研究的核心，研究主要围绕如何准确批量评估房产税税基展开。基于上述理论与方法适用性分析，提出使用默认参数可能导致模型存在局限，需建立准确性更好的优化随机森林模型。本文通过多途径搜集数据并量化，首先利用 R 语言中 randomForest 包中的 RF()函数训练出模型，同时利用该包中的 tuneRF()函数对随机森林模型中的参数 mtry 进行优化，最后再次通过 RF()函数构建调参后的随机森林模型。

（3）第三部分是房产税税基评估案例模拟应用。首先以兰州市七里河区、西固区、城关区、安宁区作为案例分析区域，建立关于该区域房产税税基批量评估的随机森林模型，调查研究兰州市目前房产税税基现状，分析兰州市整体住房价格情况，接着对兰州市房产税税基的模型的应用进行可行性分析。在明确税基评估可使用该模型进行评估后，通过在住宅销售平台搜集相关数据、实地调研和地图测量的方式获取原始数据，经过对兰州市 2021 年 6 月到 2021 年 11 月半年期间的住宅数据搜集工作后，采用了 4 种量化方式处理数据，并通过 SPSS 软件对量化后的数据进行描述性统计分析。然后，在进行大数据分析之后利用经过优化后的随机森林模型对数据进行必要性分析，分析特征变量在模型中的必要程度，并通过误差分析对结论加以检验。最后，选择了传统多元回归模型评估结果与优化随机森林模型评价结果进行比较，检验模型评估准确性，差距大则模型效果差。

本文的技术路线如图 1.1 所示：

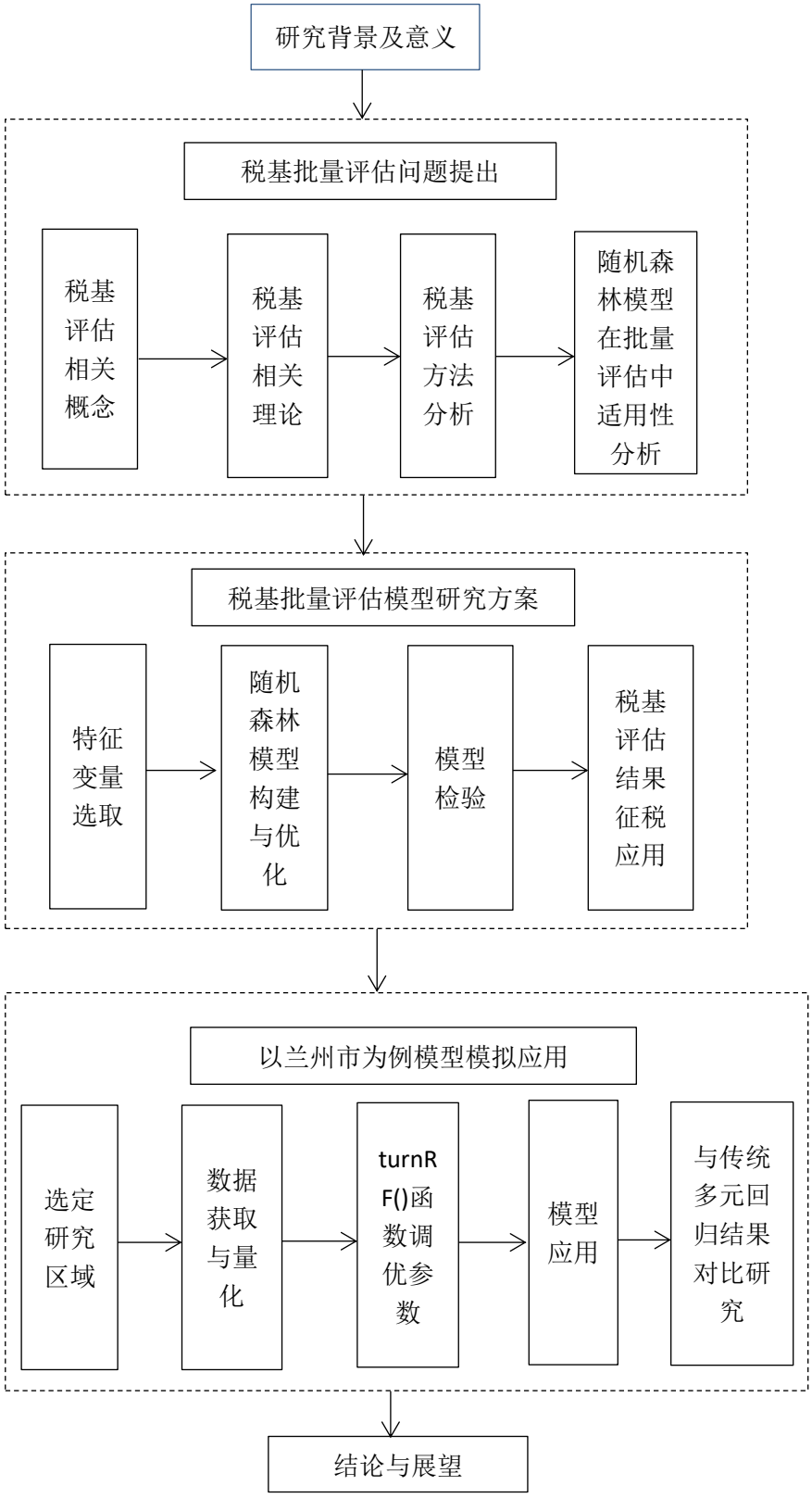

图 1.1 技术路线图

1.3.2 研究方法

案例研究法：在甘肃省兰州市城关区、七里河区、安宁区、西固区四个区域的搜集住宅类房地产的案例样本，应用数学模型进行客观评估。

文献研究法：分析与研究国内外税基的相关理论、文献和实证，也对全国硕士论文、相关期刊进行研究。结合国内的相关问题研究作为对比与参考，对所研究的问题进行系统的学习。

机器学习法：本文采用的是机器学习法中的随机森林算法。随机森林算法需要使用计算机编程软件来建立，论文使用 R 语言来进行模型的建立。R 软件是指一种完整的数据分析、运算和绘图功能的软件工具，R 语言的基本原理是，通过给出多个集合的数据模型，以及多个数学计算、数据运算的函数，是操作者可以自主方便地进行数据分析，或者自行开发出符合需要的新数据运算方式。

定量分析法：本文搜集了评估对象的多个变量，对各变量采用统一的量化方式，通过分析软件进行数据分析，再根据数据结果得出客观规律。

1.4 本文可能的创新之处

本文基于大量学者研究的基础上，将本文的创新点归纳出两方面：

（1）国内学者对我国的房产税税基研究主要是基于发达城市，如北京、上海、广州等，本等选择的研究对象新颖，对中等城市、西部城市的研究较少，本文选择了我国中等城市中的兰州市作为研究对象，此研究可为我国的中等城市房产税征收提供参考的意义。

（2）本文将随机森林模型应用到房产税税基评估中，方法较新颖。在随机森林模型的构建过程中，大部分学者对模型内部的参数选择是基于 A.liaw（2002）理论中的 mtry=p/3 作为最优参数，或者选择手动调优。基于学者研究，本文应用了薛震、孙玉林（2020）提到在 randomFoerst 包中，提供了一个优化随机森林模型参数 mtry 的 turnRF()函数，通过该函数对模型进行优化，这样的优化方式具有创新预测结果精确度的意义。

2 国内外研究综述

2.1 国外研究综述

（1）房产税税基概念及价值内涵研究现状

房产税在国外研究中有较早发展，Marvin Anderson，Robert L 等（2013）研究发现房产税起源于中世纪的美国、英格兰等发达国家。对于房产税税基的价值内涵，征收房产税的各国制度已经相对完善。一些发展中国家的房产税征税对象只涉及土地，如越南；部分发达国家房产税征收是以房屋租金为税基，如英格兰；而日本则是以住房建造成本为房产税税基。Mc Cluskey W. L（1999）将房产税的税基界定为建筑本身使用价值、土地使用价值、租金收益及房屋的一些特点，作者还比较剖析了各个发达国家的税基。Sabaliauska、Kestutis 和 Albina Aleksiene（2002）在《Progress Toward Value-Based Taxation of Real Property in Lithuania》中指出，由于立陶宛需要建设全国房屋信息系统，对各种房屋统计资料都要进行更新，并以房屋的市场价格作为估值基础。Combs,S（2011）在《Appraisal review board manual》中指出，美国联邦德克萨斯州的缴纳房产税以市场价格为主要依据，一旦评估价格高于市场实际价值，对纳税人的合法权益便可提出异议；Alison J.Iavarone（2014）在《New York State Property Tax Assessments and the Homestead Option》中认为，市场价值与通过市场法求得的价值是不同的，只是在自由经济市场上，自愿购买者和卖家之间实现公平处理的经济价值。房产税，国外有着多种定义，学者们针对房产税税基的价值大多赞同以市场价值反映其内在价值。在商品经济中，商品价格总是围绕价值运动，所以在房地产中，房地产的价值主要可以通过价格反映。房产税作为房地产税收中的一类税种，其价值也可通过市场价格反映。

（2）房产税税基评估理论与方法研究现状

房地产估价的有关研究与应用早于房产税税基评估，所以房产税税基评估相关的理论与评估技术是在房地产估价中不断发展而来的。对于房产估价，最先引进并发展起来的是市场法、收益法、成本法，目前这三个方法是最基本的房产估价方法。随着国外研究对评估的研究深入，房地产评估相关的方法理论体系逐步

完善。Rober J.Gloudem ans（1999）系统阐述了房地产评估理论。Rosen（1874）首先把特征价值理论引进了房地产评估范畴，深入研究了房屋价值和各种因素间的关联，为今后的学术界深入研究房屋价值与建筑物、邻里环境和区位等各种因素的关联发挥了引导意义。除了上述提到了房地产评估理论之外，部分学者也逐渐关注其他关于房地产估价中存在的理论，MaurizioAmato 教授认为，批量评估中可能存在数据不完整的情况，那么在这种情况下可以利用粗糙集理论，该理论用于估价中可以不受数据不完整的限制，最后同样得出比较准确的估价结果。国外将多种多样的方法应用于房地产估价中，最基本的方法还是回归分析。

随着房地产评估的范围不断扩大，国外税基评估逐渐从单一评估转向批量评估。起初的税基都是单宗评估，在 2013 年国际税收评估协会发布了批量评估的标准，明确了批量评估规范，并且还在规范中明确了技术要求，如计算机辅助估价的使用等。KontrimAs,V，Verikas,A（2011）提出税基评估的最终目的是为了房产税征收工作，评估结果不可用于其他用途。作者还对房产税批量评估构建了模型，将支持向量机、最小二乘法与计算机结合，提高了评估结论的可靠性。William M.和 Jensen David L（1990）共同讨论了人工智能技术和计算技术软件系统相结合使用的可行性，并指出将多重线性回归分析融入智能信息技术实现了批量评估的有效性。

在国外的税基批量评估方法应用上，Robert J Gloudemans（2002）在《Comparison of three residential regression models: Additive, multiplicative, and nonlinear》分析了关于住宅估价的三种回归模型，并且使用同一数据对三种模型进行了应用，同时对各模型的优劣进行了讨论。John D.Benjamin,Randall S.Guttery,C.F.Sirmans（2004）将传统回归模型与市场法进行对比，并将两种方法运用到房产税税基批量评估中，分析了两者的优缺点。由于房地产起步早于房产税征收，故国外房地产批量评估早于税基批量评估，且税基批量评估大多基于特征价格理论并利用传统多元回归评估，该方法存在一定的局限性，为了提高批量评估的效率、准确性，研究学者不断提出高效模型应用于税基批量评估中。

（3）随机森林算法与应用研究现状

由于传统批量估价方法存在局限，国外研究学者逐渐将机器学习应用到评估中。随机森林模型作为机器学习的一种算法，在国外关于机器学习较多的前提下，

随机森林算法的研究也较多。上世纪八十年代，Breiman（1884）等首次引入了分类树算法，因其具有分类决策功能，后被叫做决策树算法。该算法相比神经网络这样的深度学习模型具有更高的计算效率。Breiman（2001）最早提出随机森林模型，采用的是决策树组合的算法，通过计算机技术将大量的具有分类和回归功能决策树的预测结果进行汇总，所以称为随机森林模型，作者认为随机森林模型的预测效果要优于决策树。该模型不会因数据缺失、存在异常数据而导致预测效果下降，是一种理想的统计方法（Iverson et al. 2008）。关于随机森林模型中的参数选择，A.Liaw(2002)认为所有建立从随机森林模型中返回的参数中，取值为特征变量的三分之一，为最佳模型参数。EA Antipov，EB Pokryshevskaya（2012）第一次在房地产价值评价中介绍了随机森林模型，利用随机森林模型对房屋价值作出了评估。通过实证研究得出了，数据在随机森林算法技术中的表现比在其他技术，比如回归决策树、近邻算法、多重线性回归分析、人工神经网络中的表现要好；并提出以上方法可能是以后批量评估常用方法。关于国外随机森林模型的一些理论应用，如 Nicolai（2006）提出了分位数回归森林法。Sexton（2009）使用了 Bagging 和 randomForest 技术开展了财务预警评估，比较了二者的评估效果，并发现了基于 Bootstrap 的 randomForest 技术预测效果更好。Bart（2005）在客户关系管理的研究中引入了随机森林方法，并且通过与传统线性回归模型和 Logistic 模型的对比发现，随机森林方法的预测效果更加准确。

2.2 国内研究综述

随着中国房地产行业的蓬勃发展，以及国内外目前关于批量估价技术方面的学术成果逐步趋向完善，因此批量估价问题也引起了许多研究者的关注，关于批量估价方面相关的学术研究也日渐增多。

（1）房产税税基概念及价值内涵研究现状

税基是税收征收的计税依据，陈小悦和孙力强（2007）较早提出国外研究物业税、房产税的规律，他们发现国外主要以租金价格和市场价格反映税基。以租金价格和市场价值作为税基的实质上相同，但以市场价格为税基则充分反映了征税周期内房地产市场存在的价格差别，在一定程度上也能够保障政府征税的合理、公平。随着学者对房产税税基概念的提出，逐渐有学者对房产税税基价值内涵进

行了深入分析。姜楠（2008）探讨了在我国不同地域下征收物业税的可能性，并且提出对物业税税基评估的过程中，要以批量、高效、公平的原则进行评估，还提出了关于我国物业税税基相关的估价体系和物业税适合的评估模型。在其研究出的各种模型中，不仅适合房地产市场价格的评估，也适合税基的评估，对我国房地产相关的评估具有重要的指导意义。孙健夫（2010）认为我国房产税征收应该覆盖全国的房屋，任何区域，任何类型的住宅都要征收房产税。从征税范围和评估客体上来看，叶发强，陈西婵（2014）认为由于房产税是区域财政税收，所以针对一些地方问题深入研究房产税对地方经济社会发展的影响比宏观经济层面的有效分析更有意义。由此可见，国内学者认为租金和市场价值都能体现税基，但是市场价值更能客观体现。有的学者认为可以在市场价值反映税基价值的过程中，应该加入地域差异性的因素调整。

（2）房产税税基评估理论与方法研究现状

国内研究学者在税基估价理论方面大多从评估中的问题及其评估体系的建立方面入手。纪益成（2005）在研究税基评估方法和税基评估问题时提出了我国税基批量评估的相关理论，同时还对评估对象、评估机构等作了探讨。耿星（2004）提出在在税基评估开始之前，需要确定评估主体和评估方法，并且需要有科学的评估流程，主要包括：理论政策准备、数据收集、数据量化最后进行价格计算。郭文华（2005）阐述了立陶宛使用批量估价技术开展评估工作的方法。杭州市作为我国试点城市之一，该地区的房产税研究组（2009）对房产税税基征收涉及批量评估的理论问题做了总结，并就杭州市的税基批量评估工作进行了研究，形成了属于该地区的税基批量评估系统。房产税税基与物业税税基本质上属于同一类研究，周琳娜（2007）在开征物业税研究中，对物业税税基评估的相关理论与评估流程做了比较详细的介绍，并且讨论了数据处理最合理的方式。国内这些较早的理论研究为我国后续税基评估方法的应用奠定了基础。

我国对于税基估价方式的研究，大多聚焦在对同一个城市进行量化评估，是具有可行性的应用研究，数据来源多是二手房交易市场。张思雪等（2016）研究了哈尔滨的房地产，搜集了该城市的部分房地产数据，并阐述了评估模型的选择、检验、确定过程，为我国住宅房地产的批量评估提供理论依据和方法指导。张旭（2008）选取了厦门的住宅数据开展研究，并重点分析了特征价格法在厦门房产

税税基批量评价中的适用性。郭庆（2010）在对河北省某市的房产税税基研究中，预测了未来批量评估的良好前景，同时开展了回归方法对该区域的房产税税基的应用研究，最后探索出适合该区域的税基评估方法是分区回归法。程亚鹏（2010）分别使用三种对数回归模型和 Box-Cox 函数对保定市的二手房进行批量估价，最后验证了 Box-Cox 函数构建的模型估价效果最佳。对房产交易市场不活跃的城市，可通过其他方式来确认房产的市场价值，如交易不活跃的城镇土地可以用分等定级的成本法。学者们对税基评估的研究较全面，不仅研究了评估主体、评估流程、评估理论、评估方法，也针对不同城市、不同类型的住宅进行了应用研究。

（3）随机森林算法与应用研究现状

在税基批量评估的理论与方法不断完善的情况下，国内学者不断将统计分析与人工智能结合，解决评估乃至其他领域的现实问题。方匡南（2012）不仅分析了非参数随机森林模型的相关理论，还将随机森林模型应用于经济金融中，其研究在股票市场、房屋抵押等方面的风险预测中有着重要价值。他认为关于金融数据的预测方法主要有机器学习和深度学习两类智能预测方式，bagging、boosting 这样的综合预测方式也有助于金融数据的预测。

我国不仅对统计预测方法进行总结，也有部分学者将该算法应用于实证中。杨沐晞（2012）使用随机森林模型对广州市天河区的 49 个居民小区进行价格预测，从区位、建筑、邻里三个宏观角度选取变量，并分析了各变量对模型的重要性程度，该研究为居民购买广州市天河区域内的二手房提供了参考，也为研究随机森林模型提供了参考。陈奕佳（2012）使用随机森林模型建立了北京市二手房的批量估价模型，并且使用了五折交叉的验证方式对多种机器学习模型展开了对比，结果证明随机森林模型预测效果良好。徐戈、张科（2014）根据房价影响因素，选择了广州市某区域二手房的 21 个特征变量进行构建关于住房价格预测的随机森林模型，在上万个有效样本的基础上，验证了随机森林模型的优越性。上述三位学者在随机森林模型确定随机参数时，选择默认 mtry 的值，或者手动滚动确定参数值。陈诗沁、王洪伟（2020）根据特征价格原理，利用从链家网和 GIS 提取的数据，构建三种模型，并对上海市二手房市场开展应用研究。他们的研究成果表明：随机森林模型的绝对误差为 7.4%，评估效果理想，且优于神经

网络，更适合于现实中的二手房的批量估价。张望舒、马立平（2020）用 lasso 法、行政加权法、随机森林模型分别对房地产总体均价、不同行政区域权重、特定经济环境下二手房特征价格的三类估值，建立了关于经济环境、地域、房价的组合模型，最后证明了该种评估方式效果良好。

随机森林模型在其他领域的应用，学者发现模型回归预测的缺陷，并提出优化方法。艾裕（2020）认为传统随机森林在金融应用中存在过拟合的问题，因此以随机森林模型为基础，结合遗传算法，建立综合模型预测沪深 300 指数，研究发现优化后的模型能够有一定程度的解决过拟合的问题。随机森林模型的优化思路主要是改进模型使用的数据和改进参数，并结合其他算法综合预测结果，如遗传算法、网格搜索、集成学习等。

2.3 研究述评

从上述的国内外研究中可以看出，国外对房产税、批量评估以及随机森林模型和改进方法的研究都比较早，由于我们房产税的试点改革逐渐成为热点，国内的相关研究才不断开始。梳理相关文献，本文关于房产税、评估方法、随机森林模型的优化研究有了以下几点总结：

（1）在初期学者开展房产税研究工作时，学者们主要是对征收管理体系的研究，即对房产税的征收对象征税依据的研究。有的学者认为以房产的市场价值为税基，而有的学者则主张以房屋的租金收入为税基。

（2）由于目前房地产市场交易活跃，房地产估价逐渐从单一评估转向到批量评估中，为了提高评估效率许多学者将计算机评估系统纳入到批量评估中。房产税税基的评估实质上是对税基的市场价格进行评估，且只要符合征收条件的征税对象都需要有一个市场价值为参照进行征税，批量评估技术这时候就可以很好的应用于房产税税基评估中，再依据房地产评估中的计算机辅助系统，税基评估就可能变得高效容易，所以房地产批量为研究以市场价值作为房产税税基批量评估奠定了基础。

（3）目前与房地产价格相关的研究中，主要是以房产影响因素的研究较多，并且是以特征价格理论、区位理论等为基础构建模型指标，方法使用上多以传统方法较多，研究机器学习的批量评估较少，且从已有学者研究来看，随机森林对

房产评估效果优于其他人工智能评估。故本文选择随机森林模型作为对兰州市房产税税基的评估工具。

（4）关于随机森林模型应用到房地产价格预测中，国内大多数学者赞同随机森林模型的预测准确性好于其他方法，预测结果更能接近真实的市场价值；以及该模型能够很好的提高评估效率，降低人工耗费，是未来重要的批量评估方法。另外，目前国内学者对随机森林算法的参数选择都是基于非参数构建或者系统默认构建，对优化模型的应用研究较少。

3 房产税税基评估相关概念、理论、方法分析

3.1 房产税税基相关概念

（1）房产税税基概念

关于税基、房产税税基的相关概念。纪益成（1999）在税基评估研究方面，对税基做了解释，税基是指税收征收的客观依据，房产税征收工作不仅要求计税的实施对象，同时也要求计算征税对象应该缴纳的税额。房产税的税基则是指房产税的税收前提，也可以叫做计税依据。另外，本文也对税基估价给出了一个具体化的概念：税基估价是指具备资质的评估机构和评估工作人员，以征纳税的目的，以一些没有明确价值的客体为估价对象，按规定的流程，使用合适的估值标准和办法，独立客观公允的进行确定价格的流程。要开始征收房产税，就首先必须对房产税计税依据做出估价，也可以认为是对房产税税基的评估。

一般来说，税基估价是房产税征收的前提，为了保证公平的税收工作，需要提供科学的税基价格，为房产税征收提供依据。按照税收计征标准的不同，分为从量征收和从价征收。从量计征是指按照征税对象的计量单位征收，比如以升为单位或者以吨位单位的商品。从价计征是根据计税对象的市场价值为标准，例如需要缴纳消费税的应税消费品，缴纳企业所得税的应纳税额。

（2）税基批量评估概念

根据国际估价协会、国际评估准则、专业评估准则对批量评估的界定，批量评估是在一定的时期内，通过采用完全统一的标准方法和统一数据对大量的资产进行估价，并采用计算机辅助技术评估，以及对结果进行检验的技术手段。批量评估并不仅仅是最后形成某一综合结果，也可以是利用某一规律评估出某一批次的各个结果，这些结果客观、科学，可反映自身市场价格。

3.2 房产税税基评估相关理论与方法

3.2.1 地租理论

马克思认为，任何形态土地租赁的形成都要以土地所有权为前提，而由于随着农产品成为价值产品而发展要求和它的价值实现要求的进一步发展，土地所有权在这种由于没有它参与而创造起来的价值中不断增加，部分的权利也就发展出来，将剩余价值中一个日益增长的部分转变为了地租。马克思地租理论中，地租在一定程度上决定了土地价格，而土地用途以及所处区位主要是影响地租的因素。一般情况下土地价格是由土地用途决定的，用途越佳则土地价格越高；反之则越低。例如，不同土地上面的建筑物用途不同，位于城市中心的土地，由于其上的建筑物商业繁华、人口较多等，能带来更高的收益，使得该区域的土地价格较高，而远离中心区域的土地因其上的房屋带来的收益少，所以土地价格较低。地租理论也包括了级差地租和绝对地租，级差地租理论为土地、房地产分等定级提供了理论基础，房产税在不同区域的征收标准是不同的，在对税基评估时，需要进行区域划分，级差地租理论就为该种划分方式提供了理论支撑。兰州市居民住宅主要分为八个区域，不同区域的地租不同，从理论上解释了不同区域住宅价格差异的根本原因，有助于评估出兰州市房产税税基。

3.2.2 区位理论

区位论一般来说是指某一经营活动中所占据的经营地点及其附近事物间的经济地域关系。从某种程度上来说，区位也是影响房屋使用价格的重要因素。最早，巴基斯曾提出过同心圆住宅区位理论，该理论指出城市内部空间结构是由各种用途的小地块包围了某一中央区域，有规律的由内向外扩展，构成了同心圆形式的圈层构造，在该构造中，离中央区位越远，土地的使用便利性愈差，其租金也就越低。但是该理论却忽略了交通运输道路、河流、湖泊、影响土地使用的社会文化因素以及区位偏好等方面的影响因素。

我国各地区城市不断发展，城市中的基础设施、经济活动、市场需求呈现出多样化的趋势，不同城市间的设施空间分布有所差异，不同区位上设施对房价的

影响也表现出不同，居民对综合设施的考虑成为选房购房时常见的标准，这就形成了空间差异对房产税税基的影响因素。在房地产市场中，购买行为成为了影响房价的主要因素，居民在空间分布中有着重要考虑，许多人更愿意选择离中心区域较近的房地产。影响房地产的区位因素同样是房产税税基中的重要因素，房地产的区位价格能反映税基市场价格，所以在影响房价的区位指标，如距 CBD 距离、交通、基础设施等，可以作为税基在区位因素上的估价指标。

3.2.3 特征价格理论

特征价格理论最早出现于 1939 年，由 Court 提出的。该理论主要是指假设商品的价格与其各种类型的影响因素之间存在某种关系。任何一种关键因素的改变则会使商品的价格发生变化。此时可以利用这种变化来制定价格与影响因素之间的作用规律。基于此理论，学者们认为多元回归模型能很好的解释价格与因素之间的这种规律，所以采用了线性回归的计算分析。

特征价格理论最初并不是运用在房地产估价中，而只是用来分析农场用地的土地价值。此后，研究者们认为只要能找寻到商品的影响因素，并且成某种规律都能使用此规律，因此扩大了理论的研究范围，将其用在除了不动产之外的其他商品中，如金融资产、固定资产等。在 1974 年，是 Rosen 教授首次将特征价格理论应用到房地产领域。

特征价格理论是从经济学中效用价值论发展而来的，考虑了房屋对居民效用的大小。目前城市不断发展，人民生活水平提高，居住的目的已经不再是只满足基本居住，更加注重生活的质量，居住的各种附加服务更能影响房价，比如：交通、教育、环境、生活设施等，这些影响因素带来的效用或许比建筑物本身的建筑成本更高。如果能够量化这些效用，那么就能计算出房屋的市场价值。住宅和人类的日常工作生活息息相关，可以解决人类住宿、娱乐和办公等多方面的需要，是一种复杂的商品。因为房地产的不可移动性和唯一性，因此其异质性非常明显。房屋的不同因素体现着不同的效用，税基估价实质上是房地产的市场价格评估，

正是因为影响房价的因素可以用效用价值体现，所以特征价格理论成了税基评估中比较重要的理论基础，为后续指标体系的构建提供理论依据。

3.2.4 随机森林算法

在 2001 年，数学家布雷曼教授首次将随机森林算法带入到大众的视野。该算法中有大量的决策树，每一棵决策树具有分类预测功能，这些决策树共同组成了一个巨大分类器。随机森林算法能够克服决策树过拟合和局部最优的问题。

随机森林的随机性主要表现为以下两点：

（1）随机的训练集。布雷曼教授认为随机森林模型中的训练集是原始数据通过有放回的随机抽样，得到了与原始数据有差异的训练集，差异产生主要是由于抽样后的训练集只是抽取了部分数据，反映的是抽样前的部分样本，所以能使构建的模型不会太过发散。抽样后的样本也具有随机性，就很好的克服了局部最优的情况。

（2）随机的特征变量。随机森林模型使用特征变量构建模型，特征变量的选择是根据森林内部的每棵决策树上的分枝节点，决策树可以不受限制的分枝，模型构建过程中，不是随意使用分枝节点，而是通过挑选最优节点构建模型，从而达到准确的预测效果。随机森林算法的分类原理是基于决策树的分类功能，在原始数据中，通过随机分类，随机形成多个样本，再由每个样本形成相应的决策树，每棵决策树都会形成一个分类结果，根据每个分类结果投票选出最终结果，从而实现了数据分类的功能。随机森林示意图如下图 3.1 所示：

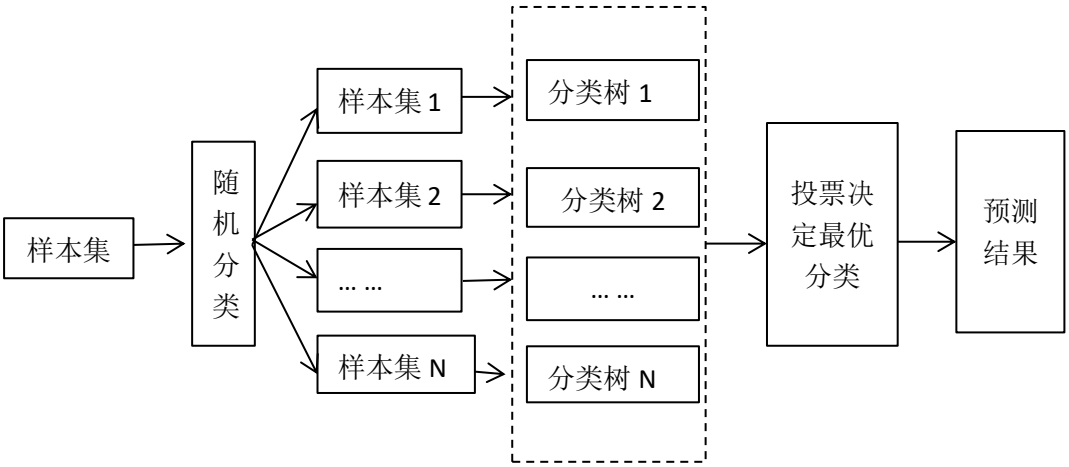

图 3.1 随机森林模型预测原理

随机森林算法所需要的参数较少，模型分类过程中，较少出现过拟合的情况，并且该算法可以对大量的数据进行分类，当模型中出现了噪音时，随机森林算法

可以不受干扰进行准确分类，不仅能提高分类效率、节省成本，而且分类精确度也较高，是一种较为理想的分类模型。

随机森林算法对房产税税基批量评估时，除了具有单宗评估的特点，同时也具有其他特点：

（1）批量评估效率高。本文中房产税作为控制城市房价的一个重要工具，征收的过程中需要以税基价格为前提，而只有将大量的同类房产价格进行加权汇总时，才能反映真实的税基价格，那么批量评估就能很好的应用于房产税基估价中。在税基估价前，估价人员可对影响住房的因素进行信息搜集，只需要把房屋的所有自变量都纳入算法中，再通过估价平台进行批量评估，这样减少了评估的时间，也减少了人工成本的耗费。

（2）批量评估体现客观性。传统的单宗评估资产时，评估价格可能因为评估人员的主观臆断，导致价格与价值之间存在差异，所以有时候不同的评估人员有不同的估价结论。但是批量评估法是从大量的估价对象中找到带有规律性的价格特征，评估出来的价格不仅能反映某一估价对象的真实价格，甚至可以反映出某一时间段、某一区域内的综合价格。该方法评估出来的价格具有统一性，能较好地体现出评估对象的客观价值。

（3）批量评估基准日具有变动性。在税基估价过程中，影响税基的因素随着时间的推移不断变动，如城市交通的修建、城市规划、小区绿化环境改善等原因，导致房价出现偏高或偏低的情况，那么批量评估就需要定期对以往构建好的模型进行修正。随着房地产市场的发展，房产交易量也在不断增加，保有环节的房产数量变化也需要对数据库进行补充，使批量评估成为一个动态的估价系统，从而保证能够评估出最具权威的市场房价，为城市房产税征收提供技术支撑，保障税收的公平。

以上的随机森林算法在批量评估中的特点能很好的契合税基评估的要求，税基作为需要大批量评估的对象，批量评估一定是必然趋势，研究好随机森林模型批量评估能为税基评估奠定基础。

3.3 随机森林模型在税基批量评估的适用性分析

随机森林模型属于机器学习模型中比较新的一种模型，在面对不完整的数据

也能够保证预测结果的准确性。随机森林模型是一种以决策树分类算法为前提，多棵决策树共同组成了随机森林模型，每个决策树由随机组成的随机向量组成，并且每个随机向量都是独立的随机森林模型主要是解决分类和回归问题，该模型可以对数量庞大的高维数据进行分类和回归。在我国已有部分学者对随机森林模型应用在批量评估中做了研究。陈钊（2015）以哈尔滨南岗区为例，将随机森林模型应用于房产税税基批量评估，实证研究得出随机森林模型预测效果良好的结论，且为模型应用在批量评估中提供了一种新思路。李侠男（2017）将随机森林模型应用到房地产项目风险评估中，批量评估各项目的风险，并对影响房地产项目的因素进行了重要性排名，为未来的风险评估工作提供了参考。在进行批量评估时，数据量庞大，随机森林模型的随机性能很好的从大量的住宅数据中分析影响房价的重要因素，并且基于这些重要因素得出较为准确的预测。庞枫（2016）在研究了多种算法模型后，证明出随机森林模型的预测准确性最好，并且在样本量足够大时，也同样能有准确性好、效率高的优势。

近些年来，兰州市的房地产快速发展，交易量也增多。政府对房产税征收前期工作中，需要有合适的评估技术对交易日益增多的住房进行税基估价，从梳理文献中可以看出，随机森林模型能对大量的数据进行分类处理，因其自身特性能满足批量评估的要求，所以该模型能够批量评估出税基价格，这为运用随机森林模型在批量评估房产税税基中奠定了基础。

4 基于优化随机森林模型的房产税税基评估模型构建

4.1 传统随机森林模型的不足

随机森林模型通过 bagging 过程进行改进，该模型（曹正凤，2014）在分类准确度、泛化误差、算法强度等性能方面与决策树相比而言，单分类器的分类准确性有了较大提升，随机森林模型的方便性、精确性使其应用于各行各业，但是随着各专家学者的深入分析与研究，模型逐渐显露了一些问题，主要表现在：

（1）随机森林模型在构建过程中，数据集的属性选择会导致 bagging 随机训练时遗漏掉重要属性，另外数据集的属性变量过少会使决策树分类过程中分类规则过于简单。因此，在模型构建中尽可能的多选择影响因素，通过增加分类规则提高分类准确性。

（2）参数是直接影响评估效果的关键数值，参数的选择一般是人工进行逐一挑选，或者使用模型默认的参数，缺乏一个基于优化理论生成的最优参数，这种参数的确定方式加大了模型构建的工作量，导致评估效率降低，并容易出现参数挑选错误的情况。

（3）随着数据集的增加，模型的噪音随之增加，当噪音过大时，该模型容易出现过拟合的情况，因此，在每次结果预测前都需要对模型中重要变量进行分析，防止噪音过多，误差过大。

4.2 随机森林模型优化过程

4.2.1 特征变量的选择

国外对影响住宅因素的研究相对较早，并且已经总结了合适的特征变量。但因为地域特点、人文习俗及经营水平的不同，国内的住宅影响因素与国外存在差异，所以在对特征变量进行选取时，国内特征变量的研究才更具参考价值。国内学者大多数学者依据特征价格理论、区位理论、地租理论等，从建筑、邻里、区位三个方面选择特征价格变量，本文参考了国内学者的研究，对变量的选择进行了总结，具体如表 4.1 所示。

表 4.1 国内特征价格理论中使用过的的特征变量

研究地区	学者	研究时间	建筑因素	邻里因素	区位因素
北京市	陈亦佳	2015 年	建筑面积、朝向、装修、楼层、建筑年龄、居室布局、建筑类别	绿化率、停车位、物业管理质量、环境卫生、治安、消防	公交、地铁、学区房、配套设施服务
广州市天河区	徐戈、张科	2014 年	面积、卧室、楼层、总层数、朝向、装修、建筑年龄	绿化、卫生、空气、小区周边自然环境、物业、生活配套、停车位	到 CBD 距离、公交、地铁
上海市	陈诗泌、王洪伟	2020 年	建成年代、总楼层、建筑面积、建筑类型、建筑结构、朝向、电梯、房屋用途、产权年限、房屋年限、交易权属	物业、小区年代、小区栋数、小区户数、学区房、三甲医院	板块、环线、与地铁站距离
哈尔滨市	陈钊	2015 年	建筑类型、建筑结构、房屋质量、建筑物的外观、户型、装修、设施与设备、建成年份、小区内部环境、物业管理	朝向、楼层、外部配套、教育配套、环境景观	距商业中心距离、交通
重庆市沙坪坝区	王筱欣，何晓斐	2017 年	建筑面积、朝向、内部装修、卧室数、客厅数、楼层、房龄、总楼层	绿化率、生活配套、教育配套、健康配套	交通

资料来源：作者整理

本文通过对上述几个有代表性区域的研究中，进一步总结出国内学者常用的特征变量，这样不仅可以减少模型构建误差，也防止遗漏重要变量。总结出来的常用特征变量如下表 4.2 所示：

表 4.2 国内特征价格理论常用解释变量

变量分类	常用解释变量
建筑因素	面积、朝向、所在楼层、总楼层、房间数、内部装修、房龄
区位因素	地铁、公交、距 CBD 距离
邻里因素	生活配套设施、小区环境、教育、绿化、容积率、商业配套、物业费、附近环境、运动设施配套（健康配套）、小区户数

资料来源：作者整理

从区位因素来看，根据以往特征变量的总结中，选择最多的是公共交通和住宅到 CBD 的距离。本文对兰州市房产税税基研究中参照以往学者研究，主要从住宅到 CBD 的距离和兰州市住宅附近公共交通情况分析。

本文具体到兰州的 CBD，根据城市结构，偏东的区域选取西关十字，偏西的区域选择兰州中心，这两个商圈集中了大量的写字楼、餐饮、商场等，交通十分便利，商业气氛浓厚，因此选择这两个地方作为兰州市的 CBD。

公交、城市轨道交通是目前兰州市民主要的公共交通出行方式，具体来说，经过住宅附近的公交线路越多，住宅交通越方便，居民出行成本也越低，住宅价格也相对高于交通不方便的住宅；目前兰州市有一条 1 号线轨道交通，连接了城市东部和西部地区，对于需要出行的居民来说十分方便，离地铁站越近的住宅便利程度较高，所以房价较高。

从建筑因素来看，主要涉及到的是每套住宅本身的情况，在选择变量时，尽量将更可能多的变量纳入进来，包括了住宅的面积、房间数、楼层、总层数、朝向、绿化率、容积率、装修、房龄。

从邻里因素来看，根据以往总结，本文选取了生活配套设施、小区环境、学区房、绿化、容积率、商业配套、物业费、附近公园、运动设施配套（健康配套）、小区户数 10 个指标。这些指标能够反映出居民的居住质量和环境，容积率反映了小区居住人口密度，绿化率、小区环境、附近公园能体现出空气质量，一些配

套设施则方便居民日常生活。

综上所述，本文根据以往国内学者的相关研究，从建筑、邻里、区位三个方面总结了多个变量，考虑到兰州市住房的具体情况，选择了不包括总价单价在内的 21 个特征变量，利用这些重要变量构建估价模型，具体的特征变量如表 4.3 所示：

表 4.3 本文选择的特征变量及其意义

特征分类	变量名	变量含义
区位因素	到 CBD 的距离	从住宅小区开车到 CBD 的最短距离
	公交线路条数	住宅小区周围公共交通的便利程度
	地铁站个数	住宅小区周围地铁站的个数
建筑因素	建筑面积	住宅的建筑面积
	房间数	住宅的卧室个数
	所在楼层	住宅所在楼层
	总层数	住宅总楼层
	朝向	住宅朝向
	容积率	住宅小区的容积率
	装修	住宅的装修程度
	房龄	住宅的建筑年龄
	基础配套设施	住宅的基本配套情况
邻里因素	小区户数	住宅小区总户数
	绿化率	住宅小区的实际绿化率
	小区环境	住宅小区的环境质量
	附近公园	周边自然环境质量
	物业费	住宅小区的物业管理费率
	运动设施	住宅小区内运动设施配套
	商业配套设施	住宅小区附近生活配套设施
	附近大学	住宅小区是否临近大学
	学区房	住宅小区是否有学位资格

4.2.2 基于随机森林模型的房产税税基评估模型构建

实际问题的复杂与多变，通常并不能存在一个具体而通用的函数，可以表示各种事物之间的数值变化规律。与传统的多元线性回归模型并不相同，随机森林模型无须对函数形式先加以假设，减少了假设偏差。利用随机森林模型开展的税基估价研究，实际是建立在其研究样本内精确地拟合学习规律，在研究样本之外高置信水平下推广规律的能力。

随机森林模型包括了分类和回归两类技术，但由于本文研究的是房产税税基评估，即对不动产价格进行评估，因此属于回归估计，随机森林模型的回归算法的基本思路是：

（1）原始样本数据的数量为 N，用 bootstrap 有放回地选择了 N 个随机样本集，并据此建立 N 棵决策树，这 n 个结果形成了随机森林模型的测试集。随机森林模型是 N 棵树，{T_1(X),…,T_n(X)的集合，其中，X={x_1,…,x_n}是住宅的 n 维特征向量。集合将会产生个 N 结果，{$\widehat{y_1}$=T_1(X),…, $\widehat{y_N}$=T_N(X)}，其中，$\widehat{y_N}$为第 N 棵树对房产价值的预测值。对于回归问题, $\hat{y}$是所有棵树预测的平均值。

（2）原始样本数据的特征变量为 P 个，则在每棵决策树的各个节点随机选择 mtry 个变量，然后按照最优分枝选择 mtry。mtry 是随机森林回归模型中的唯一模型调整参数，随机森林中的每棵决策树能够最优化的生长发育，无须剪枝。

（3）重复（1）（2）步骤后，直到 N 棵树全部建好。完成上述两步，随机森林模型的训练集就构建完成。最后，通过将房产税税基测试集的特征变量代入到设置好的随机森林模型中，可以得出测试集的结果，利用该结果可以分析测试集的预测情况，并且和住宅的实际价格进行比较，检验随机森林模型的预测效果。

4.2.3 turnRF()函数参数优化

随机森林算法的两个重要参数 mtry 和 ntree 的功能主要是使模型更易于训练，本文主要是对 mtry 进行优化，即寻找模型中单棵树中最优的特征数，希望提高模型预测效果，提高模型性能。

模型中两项参数的引入提升了模型对噪声数据的处理能力，较好地克服了过拟合问题。mtry 可确定每一次迭代的特征变量抽取个数，一般 mtry 的默认值为

2，A.liaw（2002）提出以 mtry=p/3 作为最优参数，或者选择 for 命令找最优参数，但随着人工智能的发展，部分学者在寻找参数优化中，提出了新的方法，基于学者研究，本文运用薛震（2020）提到在 randomFoerst 包中，给出了一个优化随机森林模型参数 mtry 的 turnRF()函数，该函数通过对变量从 1 到最大变量进行内部训练，直至训练出误差最小的 mtry，找到最优参数以后，再通过 randomForest 训练随机森林模型

4.2.4 模型优化前后误差对比

优化前后的随机森林模型的预测误差都使用包外误差 OOB Error 反映，这里使用 ntree 默认值 500。随着决策树数量的增加，模型训练误差也随之变化，为了更直观的展示，本文选择训练集数据，对优化模型前后进行误差进行可视化对比。

```
> rfregerr<-as.data.frame(plot(rfreg))
>colnames(rfregerr)<-"rfregerr"
>rfregbesterr<-as.data.frame(plot(rfregbest))
>colnames(rfregbesterr)<-"rfregbesterr"
>plotrfdata<-cbind.data.frame(rfregerr,rfregbesterr)
>plotrfdata$ntree<-1:nrow(plotrfdata)
>plotrfdata<-gather(plotrfdata,key="type",value="Error",1:2)
ggplot(plotrfdata,aes(x=ntree,y=Error))+geom_line(aes(linetype=type,colour=type),size=0.9)+theme(legend.position="top")+ggtitle("随机森林回归模型")+theme(plot.title=element_text(hjust=0.5))  #可视化优化模型随着树的增加，误差 OOB 的变化
```

rfregerr 表示优化前随机森林模型误差变化，rfergbesterr 表示优化后随机森林模型误差变化，从可视化图中可以看出，ntree 接近默认值 500 时，误差趋于稳定，故本文使用默认值作为 ntree 值，另外优化模型后产生的误差低于默认值，由此可见，该种优化方式大大减少了随机森林回归中的误差，增加了预测的精确性，具体模型变化如图 4.1 所示：

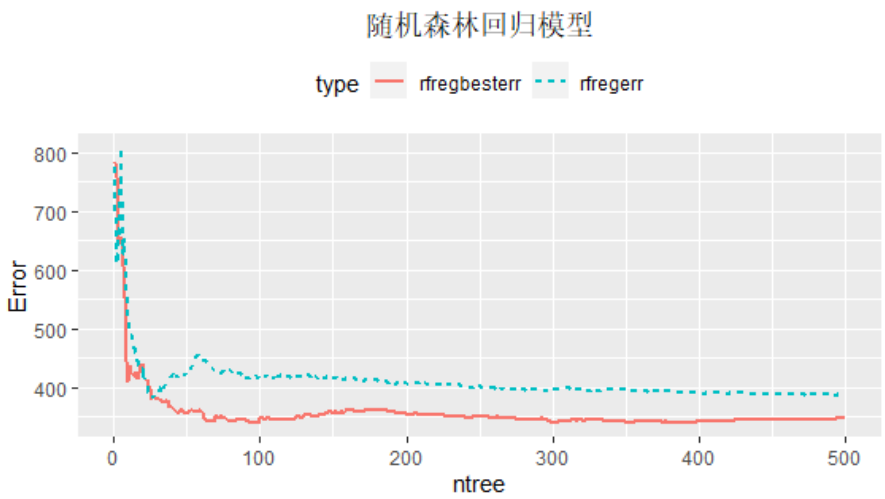

图 4.1 优化前后随机森林回归误差对比

4.3 测试模型评估准确性及征税应用

构建随机森林模型目的是提高评估效率和评估精确度，模型的准确性越高，在未来实际应用中预测的价格才能更加贴近市场价格，这样的预测价格才能作为确定房产税税基计算的评判标准，因此我们需要对评估结果的准确性进行检验。

4.3.1 测试模型评估准确性

构建随机森林模型目的是提高评估效率和评估精确度，模型的准确性越高，在未来实际应用中预测的价格才能更加贴近市场价格，这样的预测价格才能作为确定房产税税基计算的评判标准，因此我们需要对评估结果的准确性进行检验。本文选取了匹配度、绝对误差、相对误差、平均平方根误差四个指标来对模型评估结果进行测试。

（1）匹配度

$$\frac{\hat{P}_i}{P_i} \qquad (1)$$

上述公式中，$\hat{P}_i$是通过优化随机森林模型预测出来的数值，P_i表示房产税税基的真实数值。两项的商表示预测的住宅价格与真实价格之间接近程度。如果越接近 1，那么预测效果就越理想。

（2）绝对误差

$$\left|\left(\hat{P}_i - P_i\right)\right| \qquad (2)$$

该数值主要是反映预测值与实际值间的实际偏差，数值越小，模型评估效果越理想，具体的实际评估情况要重新预测。

（3）平均相对误差

$$1/n \times \sum \frac{\left|\left(\hat{P}_i - P_i\right)\right|}{P_i} \qquad (3)$$

该指标反映的是绝对误差与真实值之间比例大小，根据国际评估经验，数值在 20%以内，表明预测结果理想。

（4）平均平方根误差

$$1/n \times \sqrt{\sum_1^n \left(\hat{P}_i - P_i\right)^2} \qquad (4)$$

平均平方根误差也是评价模型预测结果的指标之一，该指标反映了真实值与预测值之间的离散程度，结果不超过 2 结果则在合理范围，数值越小，预测效果越好。

4.3.2 税基评估结果征税应用

参考上海市房产税征税规定，兰州房产税税基可依据评估的应税住宅的市场均价，同时市场价值还需定期估算。由优化后的随机森林模型预测出的各房屋市场价值求取平均数后，最后价值作为房产税税基。

$$1/n \times \sum_1^n p \qquad (5)$$

5 优化随机森林模型应用案例分析——以兰州市为例

5.1 兰州市房产税税基现状分析

兰州市位处于中国中西部地带，市辖区面积较小，全城共包括五区三县，即城关区、七里河区、红古区、西固区、安宁区和永登县、皋兰县、榆中县。其中城关区地处兰州市东南方位，该区域面积较大，是住宅类房屋集中的地区；西固区作为兰州市的工业区，工厂比较多，空气质量较差，所以居住类住房相对较少；安宁区是高校聚集区域，兰州大部分高校都集聚此处；笔者在搜集数据时未查询到红古区的相关数据，兰州市的三个县属于兰州的远郊，住房交易量相比城区较少。所以本文选择的样本范围放在城关、七里河、西固、安宁四区。

在 2020 年 3 月到 2021 年 3 月，兰州市批准上市了 10 万套房地产，批准上市面积为批准上市面积为 1218.27 万㎡，共成交约 8.2 万套房地产，共计 1068.77 万㎡的销售面积，销售额达到了 1023 亿元，兰州楼市 2020 年 3 月到 2021 年 3 月因受疫情影响，销售数量有所减少，全年销售均价为 9102 元/㎡，这一年间兰州市住宅类房地产上市情况如图 5.1 所示：

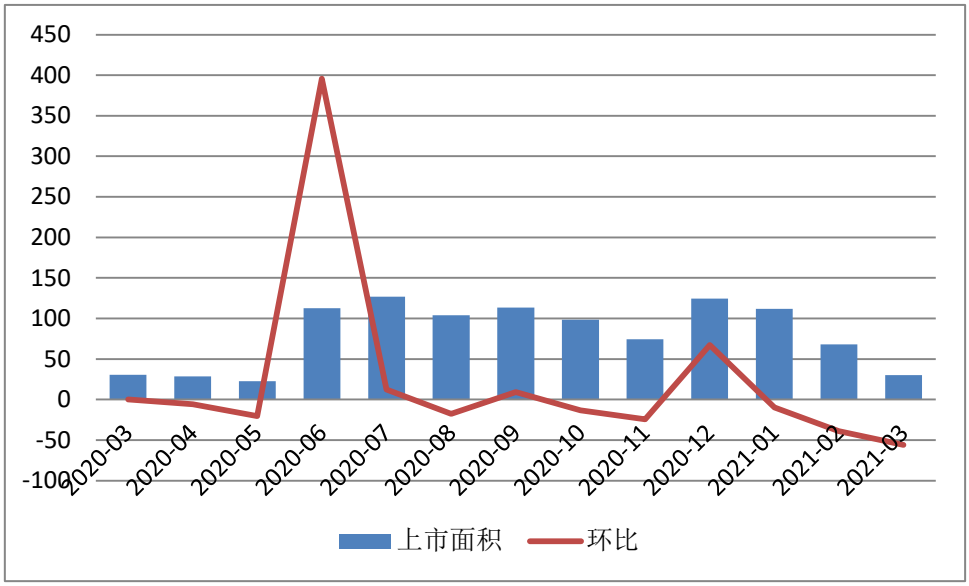

图 5.1 2020 年 3 月-2021 年 3 月兰州住宅供应面积指数变动

资源来源：房天下官网

2020 年 3 月到 2021 年 3 月之间，兰州市商品房成交总套数 8 万套，其中住宅类房地产成交总套数约为 7.5 万套，占总成交的 90%以上，销售面积 908.43 万平方米，销售额 826.83 亿元，销售均价维持在 9574 元/平方米。在 2021 年 2 月份兰州市住宅的销售面积减少，销售数量也较少，主要是由于 2 月份正是国内春节假期，居民更多的是回家准备休假，所以住宅销售面积出现了下降，但从整个周期来看，波动幅度不大，兰州市整体住房均价成交都较为平稳，如下图 5.2 所示：

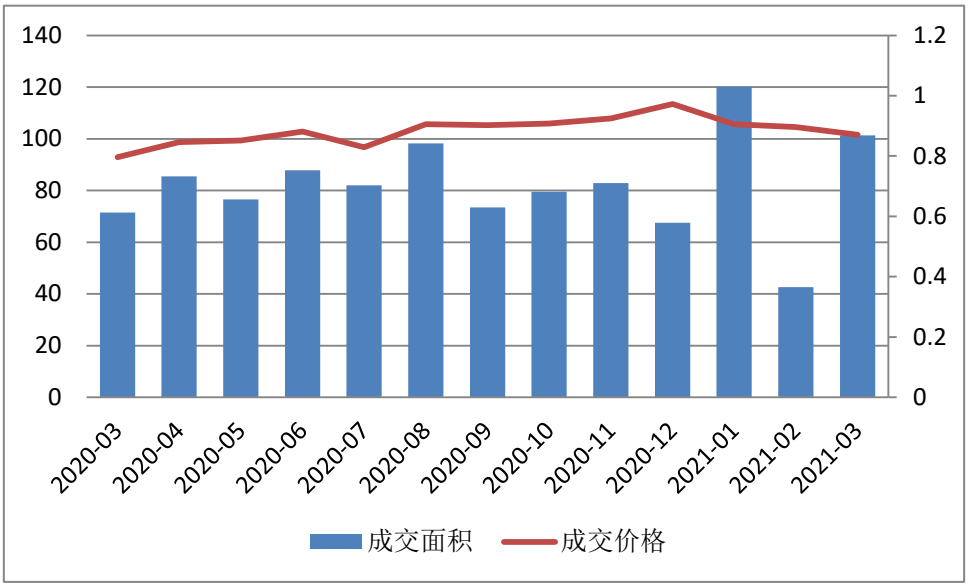

图 5.2 2020 年 3 月-2021 年 3 月兰州市住宅面积和价格交易指数变动

数据来源：房天下官网

5.2 兰州市房产税税基评估应用优化随机森林模型可行性分析

通过对兰州市住宅类房产税税基市场特点的分析可知，目前兰州市住房税基已达到了批量评估的市场条件基本要求。要完成税基批量评估还需要考虑房产登记制度、数据获取难度、搜集数据工具，以下分别从这三方面对兰州市税基应用随机森林模型进行批量评估的可行性分析：

（1）兰州市房屋信息登记制度相对完善

目前，随机森林模型在批量评估时，要求该地区需要有比较完善的产权登记制度。我国所有城市包括兰州市内都有完善的产权登记制度，我国的土地所有权

属于国家和集体所有，但是土地上的房屋建筑物居民个人有处置权。为了更好的管理城市房屋，我国成立了专门的住房主管部门专门对房屋进行了产权登记变更管理。税基批量评估人员可在管理部门统一获取兰州市住宅的房屋信息和产权信息，该种制度不仅满足了税基批量评估要求，也满足随机森林模型的应用要求。

（2）地理信息系统的发展

地理信息系统的发展为批量评估带来了数据收集的技术支撑。地理信息系统是获取房产地理相关的空间数据。该技术利用卫星定位房产和周边的空间数据，从而使地理信息系统融入到税基批量评估体系中，从中获取房产的区域定位对批量评估构建模型也是很有意义的。而且基于地理信息系统覆盖性广和精确度高的特性，对该技术的使用上还有助于减少在数据遗漏地域划分的错误。把评估建模流程和估价结果传输到该系统，纳税人就能够随时随地查询到相应的房产税信息，提高了评估的透明度。

（3）数据库系统的不断完善

兰州住房数据库是从上世纪开始建设的，目前逐渐发展起来。最初的数据库容量较小，随着计算机的发展，目前兰州市建立了巨大的数据库系统，并且形成了富有特点的大数据库系统产品。虽然兰州市建立的数据库还存在生产规模小、重复建库、兼容性、标准化低等问题，但是随机森林模型进行批量评估时所需要的住房信息已经在数据库中较完善，数据库完全可以满足模型构建的要求。此外，兰州市的房屋各种信息还存在于不同的地方政府部门，如果能将各部门数据库内容资源进行整合，那么更能提高评估工作效率。

5.3 数据获取与处理

5.3.1 数据获取

由于实际交易资料的存于政府，出于对购房者的隐私考虑，交易信息只是登记留存，不会对外公布，所以关于房屋的信息难以获取。房屋的中介挂牌价存在于网上交易平台，对于有购房需求的购买者获取较容易，如果挂牌价过高，需求者转而购买其他住房，无人购买住房；如果价格过低，难以应付建房、装修、物业等费用，所以挂牌价在一定程度上能够反映房地产市场供需情况。另外，通过

网上交易平台不但可以掌握房屋的挂牌价格，还可以掌握房屋的一些特征数据。所以，本文用房屋的挂牌售价为本文研究的税基价格，以房屋的总价作为因变量。

本文在房天下、链家、贝壳找房、搜房网等网站查询样本的资料信息，不能通过交易平台获取的信息，采用现场走访的方式进行记录。本文搜集了兰州市2021 年 6 月到 2021 年 11 月的住宅房价，以 2021 年 11 月 18 日为评估基准日，样本经过整理汇总，并进行了数据预处理，在除去个别缺失数据以及特别异常房价的住宅后，共搜集了 670 个样本数据信息，从其中随机选择 638 个作为训练集，以构建随机森林模型，将剩余的 32 个样本数据作为模拟最后模型预测结果的预测集。当随机森林模型构建之后，再利用 638 个样本数据进行特征价格模型回归，分析传统回归下的模型准确性，再次利用 32 个样本数据进行模拟预测。利用 spss 对训练集和测试集数据进行描述性统计分析，并且发现：两个数据集分布位置相似，测试集能够表现出训练集相同的市场情形，可以较显著显示出随机森林模型的各项影响指标对住宅总价的影响程度。

5.3.2 特征变量量化

本文对房产税税基价格的研究主要是从建筑因素、邻里因素、区位因素三个方面。针对这三个方面的各项特征变量采用了四种方式进行量化。

（1）采用真实值

该种方式是直接选择特征变量指标的实际值。这种方式简单易懂、容易操作也能客观清晰的反映出特征变量的真实情况。本文采用这种量化方式的特征变量有，建筑面积、单价、总价、总层数、房间数、房龄、绿化率、容积率、小区户数、距 CBD 距离、公交线路条数、地铁站个数、附近公园，共计 13 个指标，各项指标的具体量化方式如表 5.1 所示：

表 5.1 住宅特征数值量化变量

特征变量	指标量化	来源
建筑面积	住宅建筑面积（m^2）	挂牌数据
总价	住宅总价（万元）	挂牌数据
单价	住宅单价（元）	挂牌数据

续表 5.1

特征变量	指标量化	来源
建筑房龄	住宅房龄（以 2021 年的建筑房龄为 1 计算）	挂牌数据
总层数	住宅总层数	挂牌数据
容积率	小区容积率	挂牌数据
绿化率	小区绿化率	挂牌数据
小区户数	住宅小区户数	挂牌数据
公交线路条数	住宅小区 500 米之内公交线路条数	电子地图
地铁站个数	住宅小区 1000 米之内地铁站个数	电子地图
距 CBD 距离	住宅小区到市中心的最短驱车距离（单位：千米）	电子地图
附近公园	住宅小区 1000 米之内公园个数	电子地图
房间数	住宅卧室房间数	挂牌数据

资料来源：作者整理

（2）量化评分

第二种是对各项特征变量进行量化评分，在搜集原始数据是，利用 Likert 量化表进行预先分级，数据完成搜集后，再对数据进行量化评估。采用这种量化方式的指标有：小区环境、生活配套两个指标，情况如表 5.2 所示：

表 5.2 Likert 量化评分表

特征变量	指标量化	来源
小区环境	Likert 量化	实地调研
生活配套	Likert 量化	实地调研

数据来源：作者整理

（3）综合性评分

该种量化方式是将住宅的某一特征变量分解成不同的评分标准，符合标准即可得分，最后再将得分汇总得出这一变量的的综合评分。本文中采用这种量化方式的指标有：商业配套一项变量，如表 5.3 所示：

表 5.3 综合性量化表

特征变量	指标量化	来源
商业配套	住宅 1 千米范围内是否有医院、商场、超市、菜市场、银行五项设施，每项 1 分	电子地图

资料来源：作者整理

（4）虚拟变量量化

该种方式是在量化的时候采用虚拟变量，每项变量的具体情况都是独立的，采用这种量化方式的指标有：是否学区房、附近大学、所在楼层、朝向、装修五项，如表 5.4 所示：

表 5.4 虚拟变量量化表

特征变量	指标量化	来源
是否学区房	住宅有学区资格为 1，不是则为 0	挂牌数据
附近大学	住宅 1 千米内有大学为 1，不是则为 0	电子地图
所在楼层	高楼层为 3，中楼层为 2，低楼层为 1	挂牌数据
朝向	南北、南、西南、东南为 1，否则为 0	挂牌数据
装修	精装修为 3，简装修为 2，毛坯为 1	挂牌数据

资料来源：作者整理

综上所述，本文采用四种特征变量量化方式，由于住宅总价需要用来建模，故对余下的 21 项特征变量进行量化，用于优化随机森林模型的构建。

5.4 兰州市房产税税基估价模型

5.4.1 描述性分析

随机选择 638 个样本数据作为训练集，用来构建随机森林模型，将所有变量进行描述性统计分析，具体的分析情况如表 5.5 所示：

表 5.5 训练集样本数据的描述性统计

	样本个数	极小值	极大值	均值	标准差	方差
总价（万元）	638	49.00	468.00	126.2442	4.660306	2.171846
单价（元）	638	7407	25159	13239.55	2579.592	6654293.827
建筑面积	638	36.00	338.00	95.4500	25.03723	626.863
房间数	638	1	6	2.16	.478	.229
总层数	638	5	42	21.18	10.805	116.748
所在楼层	638	1	3	1.97	.810	.655
朝向	638	0	1	.81	.392	.154
装修	638	1	3	2.50	.510	.260
房龄	638	-1[①]	23	11.95	5.274	27.817
是否学区房	638	0	1	.51	.5003	.250
距 CBD 距离	638	.89	15.00	7.4543	3.28739	10.807
公交线路条数	638	0	29	5.40	6.057	36.683
地铁站个数	638	0	2	.70	.781	.609
容积率	638	.90	7.12	3.0942	1.02810	1.057
小区户数	638	208	7072	1802.52	1366.327	1866850.426
绿化率	638	.04	.72	.3307	.14018	.020
小区环境	638	3	5	4.34	.751	.564
附近公园	638	3	5	4.37	.719	.516
商业配套	638	1	5	3.85	1.122	1.259
物业费	638	0.2	2.1	1.30081	0.483083	0.233
运动设施	638	3	5	4.62	0.656	0.43
基础设施	638	3	5	4.75	0.5	0.25
附近大学	638	0	1	0.56	0.497	0.247

数据来源：由 SPSS 运行得出

将余下的 32 个样本数据用于测试优化后模型的评估效果，预测集中特征变

[①] 是指 2021 年现房基期年份与 2022 年期房年份差值

量的描述性如下表 5.6 所示：

表 5.6 预测集描述性统计

	样本数	最小值	最大值	均值	标准差	方差
总价	32	83.8	193.5	129.528	28.7750	828.001
单价	32	8226	17729	13502.09	2164.453	4684855.378
建筑面积	32	73.48	173.03	97.0506	17.64018	311.176
房间数	32	2	4	2.09	.390	.152
总层数	32	6	34	25.91	9.610	92.346
所在层数	32	1	3	2.13	.833	.694
朝向	32	0	1	.91	.296	.088
装修	32	1	3	2.41	.615	.378
房龄	32	3	21	10.75	4.235	17.935
是否学区房	32	0	1	.50	.508	.258
距 CBD 距离	32	2.1	15.0	6.913	3.1329	9.815
公交线路	32	0	9	3.59	3.140	9.862
地铁个数	32	0	2	.69	.693	.480
容积率	32	1.30	7.12	3.4447	1.27269	1.620
小区户数	32	248	4626	1966.53	1053.813	1110522.515
绿化率	32	.10	.72	.3769	.14421	.021
小区环境	32	3	5	4.50	.718	.516
附近公园	32	3	5	4.38	.751	.565
商业配套	32	2	5	4.22	0.87	0.757
物业费	32	0.35	2	1.48766	0.338481	0.115
运动设施	32	3	5	4.75	0.568	0.323
基础设施	32	4	5	4.91	0.296	0.088
附近大学	32	0	1	0.47	0.507	0.257

数据来源：由 SPSS 软件运行得出

从以上两表中极大值、极小值可以看出，训练集的总价变量范围在 49~468

之间，预测集的范围在 83.8~193.5，可以看出住宅总价所在范围包含了预测集所在范围。

在建筑因素上，训练集房间数范围在 1~6 之间，预测集房间数在 2~4 之间，训练集总楼层和所在楼层的范围在 5~42 和 1~3 之间，测试集住宅总楼层和所在楼层的范围在 6~34 和 1~3 之间。训练集住宅房龄在-1~23 年，测试集住宅房龄在 3~21 之间。

在区位因素上，训练集中地铁站个数范围在 0~2 之间，公交线路条数范围在 0~29 之间，距 CBD 距离范围在 0.89~15 之间；预测集中地铁站个数范围在 0~2 之间，公交线路条数范围在 0~9 之间，距 CBD 距离范围在 2.1~15 之间。

在邻里因素上，训练集中小区环境、附近公园、运动设施、商业配套设施的范围都在 1~5 之间，以上特征变量的预测集变动范围也在 1~5 之间，变动区间一致。

在以上指标中，由于本文采用样本较多，训练集数据较多，反映出来的范围更大，更能够反映实际征税中的房地产市场，此外，训练集建筑因素上的各范围都包含了预测集范围，且两者的变动方向一致，能在一定程度上反映预测集情况，说明随机选择的测试样本能够反映出建筑因素在税基价格中的影响。

5.4.2 参数优化

随机森林模型有两个重要参数，依次为森林中的决策树棵树、单棵决策树的分枝节点数，两个参数的加入不仅提高了模型的噪音处理能力，而且可以克服过拟合的情况。使用随机森林模型进行回归估计时，由于估计工作量大，所以还需借助相应软件加以运算。本文使用的 R 编程语言及其附带的软件包，主要进行了建立随机森林模型、优化模型参数、训练集和测试集结果预测的研究工作。R 语言有着巨大的数据处理环境，并且提供了大量的程序包。

本文所使用的是 R 语言中的 randomForest 程序包来完成建模。构建模型时需要对参数 mtry 加以选择。基于 Andy Liaw（2002）理论，建立随机森林回归的参数 mtry=p/3 时表现较好，其中 p 表示特征变量个数。根据该理论其参数的确定需要逐一确定最佳参数，在 randomForest 包中，提供了一个优化随机森林模型参数 mtry 的 tuneRF()函数，该函数可以通过多次训练可以寻找更优的随机森

林模型的参数。故本文首先采取 Andy Liaw 理论下的参数方法，模拟出随机森林模型，同时运用 tuneRF()寻找最佳参数，将两者进行对比研究，最终构建出最佳的随机森林模型。

5.4.3 模型构建

随机森林模型不同于其他模型的是，该模型构建完成后无法直观的反映模型情况，如同一个黑箱，把完成处理的数据放入系统中，系统通过模型的构建在黑箱中形成规律，研究者无法总结出该规律，只能重新放入预测集，对比差异推算出模型的预测效果，所以本文模型构建通过模型代码与拟合优度反映，预测结果通过误差分析反映。

模型构建前，在本文第四部分误差对比中，已经确定 ntree 采用默认值 500 误差较小，故在此部分 ntree=500，根据学者理论使用 mtry=7 作为构建模型的优化前参数，模型随机森林模型的代码和拟合优度为：

```
>library(randomForest)                    #加载 randomForest 包#
>dat<-read.csv("XLJ.csv",header = T)       #导入训练集#
>dat                                      #输出样本导入情况#
>set.seed(1234)                           #设定随机数种子#
>rf<-randomForest(prices~.,data=dat,mtry=7,na.action=na.omit,ntree=500)
#构建随机森林模型#
>print(rf)                                #结果输出#
Call:
randomForest(formula = prices ~ ., data = dat, mtry = 7, ntree = 500,
na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 500
                           mtry: 7
         Mean of squared residuals: 193.4565
                    % Var explained: 91.08
```

由上述运行结果可以看出，模型运算过程中的几个重要参数。其中，

regression 表示此模型的种类属于随机森林回归模型。由结果可知，本次模型构建中有 500 棵决策树，任一棵决策树节点处确定的变量数目 mtry 是 7。结果 193.4565 表示随机森林模型的残差平方和。上述模型的构建中，na.omit 表示当原始数据存在缺失或者异常时，能够忽略此类数据，利用其余数据构建模型。以上随机森林模型的拟合优度达到了 91.08%，拟合优度越接近于 1 说明预测效果越好，随机森林回归模型已经获得了很小的误差，那么该回归模型的预测精度还可以再提高。如图 5.3 所示：

>rftune<-tuneRF(x=dat[,-c(1:3)],y=dat$prices,stepFactor=1.5,ntreeTry=500)
#参数搜索，寻找合适的 mtry，训练更好的模型#

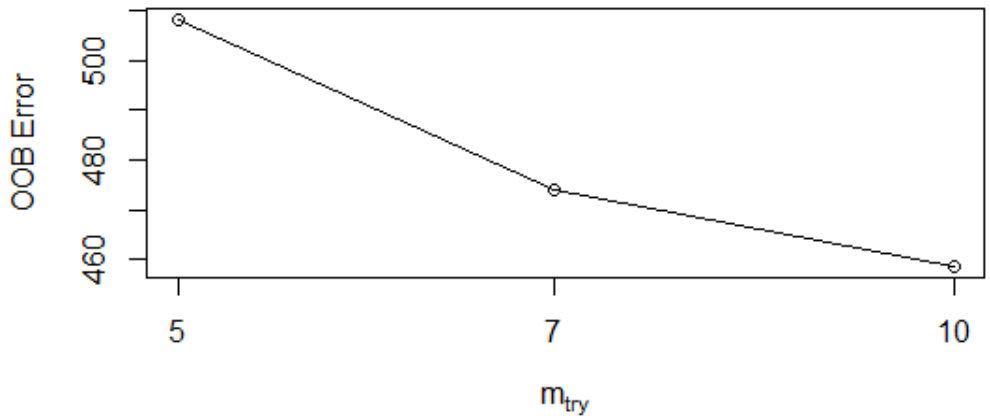

图 5.3 寻找随机森林模型的最优参数

从图 5.1 可以发现，随着参数 mtry 取值的增加，OOB Error 的取值逐渐变小，当 mtry=10 的时候，OOB Error 误差取值最小，说明使用参数 mtry=10 可以建立预测精准度更准确的随机森林模型。

>rf<-randomForest(prices~.,data=dat,mtry=10,na.action=na.omit,ntree=500)
#优化随机森林模型#

>print(rf) #输出结果#

Call:

randomForest(formula = prices ~ ., data = dat, mtry = 10, ntree = 500,

na.action = na.omit)

Type of random forest: regression

Number of trees: 500

mtry: 10

Mean of squared residuals: 156.1826

% Var explained: 92.8

经过两次模拟随机森林模型的拟合度发现，当 mtry=7 时，训练集的拟合度是 91.08%，通过 turnRF()函数寻找到模型参数 mtry=10 时，拟合度为 92.8%，拟合优度提高，该种优化方式能直观的看出优化后的模型效果，并且 turnRF()也方便参数的寻找,具体的预测效果还需要进一步对各个预测集中的样本进行分析。

5.5 优化随机森林模型后的结果分析

随机森林模型在构建过程中，当数据中存在干扰变量时，也能完成分类和回归功能，从而在最终结果预测中表现良好。随机地为各特征变量指标添加了噪音干扰，然后通过观察随机森林模型预测准确率的改变，接着观察预测准确率减少的程度，对特征变量的必要程度作出了可视化分析，IncNodePurity 指的是节点纯度，该值越大，则说明在添加噪音干扰时，所增加的方差系数也越大，表示该特征变量越必要。将前面构建模型用到的特征变量当作输入变量，反映出的特征变量必要性如下图 5.4 所示：

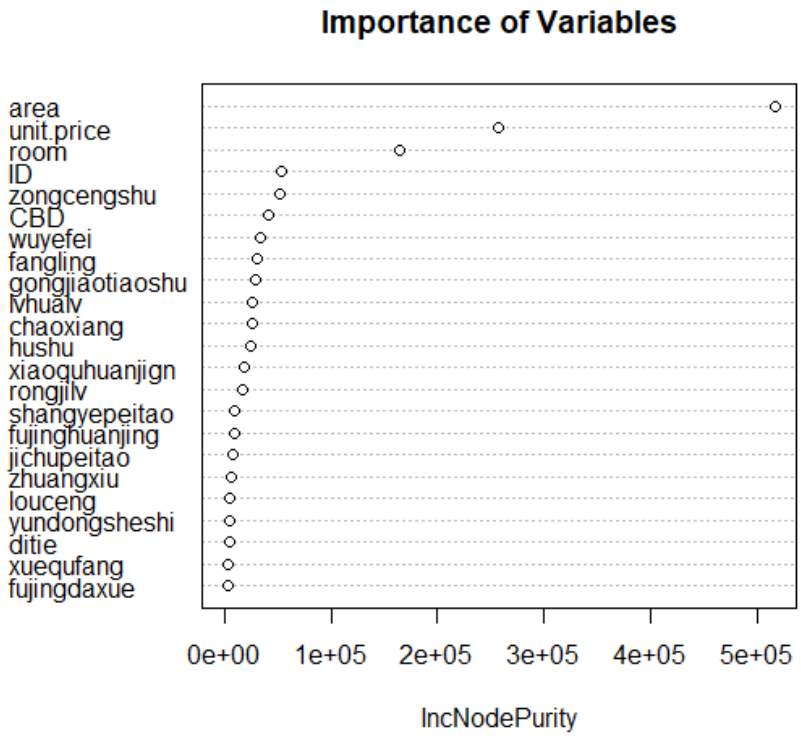

图 5.4 回归模型的变量必要性程度

从特征变量的必要性可视化图中，可以看出房屋面积、单价、房间数、总层数、距 CBD 的距离这几个变量的重要性是靠前的。其中，住宅的面积、单价直接关系到因变量总价的大小，即相同面积下，单价越高，总价越大；相同单价下，面积越大，总价越大。所以对总价的影响是最大的。其次数量也是住宅面积的另外一种体现方式，房间数量越多，所占的面积就越大，与住房的价格有很大的关系。总层数体现了住宅的居住密度，一般来说总层数越低居民户数可能越少，人口密度更少，总层数合适的住宅更能吸引居民购买。在距 CBD 距离指标中，由于 CBD 有发达的商业氛围、有方便的交通工具，对于很多工作的居民来说，相比通勤时间长和休息时间相对较长来说，更愿意选择后者，所以距 CBD 距离与住宅的价格有很大关系。

以上指标对房屋市场价格起着至关重要的作用，说明消费者在购买房屋时更加看中以上指标，那么在对兰州市房产税税基评估时，税基采用的价格是房屋的市场价格，故在税基评估的过程中可以综合考虑上述指标。

地铁站个数、公交线路条数、商业配套服务设施、附近大学以及朝向在本次

模型应用中的必要性比较靠后。这主要与兰州市地域样本的区域有关，本文选择的四个区域，居民都比较集中，公共交通密集，公交覆盖了大部分区域，所以公交对税基价格的影响相较于其他指标较小。对于地铁站，实证中将范围取到了1000 米，是公交量化标准的 2 倍。当取到 500 米时，大部分住宅都没有地铁；当取到 1000 米以上时，大部分住宅都能有地铁设施，所以地铁对住宅价格的贡献程度偏小，随着兰州地铁的不断修建，如果扩大评估范围，地铁站这个变量的影响程度应该会更高。

本文案例选取的是兰州市四个城区，住宅较多，人员也较多，并且选取的都为成熟小区，配套设施比较完善，只是部分小区的配套设施有些损坏，但总体而言，住宅小区的基础配套设施都是良好以上的；另外，兰州市属于北方城市，供热等设施齐全，区域性的差异不大。因此，基础配套设施对兰州市房产税税基价格的影响程度较小。

本文选取了学区房这一特征变量，考虑到目前教学资源有限，家长们更加注重孩子教育问题，在我国城市教育中，有住宅的居民一般具有学位资格，其孩子就能到住宅所在区域学校就读，关注孩子教育的家长在购买房屋时考虑了学区资格，所以本文构建模型选择了该项指标。但模型的必要程度显示，是否学区房这一变量在模型中表现不好。分析发现，由于受国家政策影响，防止学区房投机，造成房产市场混乱，学区房价格得到有效控制，其价格与正常居住房屋市场价格都差异不大，因此该指标对税基价格的影响程度较小。

从大范围来看，学校周边的房屋价值应该是周边环境较好、租房人士较多，也较其他的房屋价值好，所以本文构建模型选择了该项指标，最后模型的结果显示该指标表现不好。分析发现，大学附近住宅除了环境好以外，还存在人员流动大，社区管理困难的情况，另外像甘肃农业大学、西北师范大学这些大学的位置距离 CBD 的距离比较远，上班购物距离较远，因此该指标未能表现出期望效果，如果位于市区近的高校较多运用此指标，该指标对税基价格的影响体现的更好。

当对朝向加入干扰因素时，对比其他变量，方差系数较小，表明该变量在对兰州市房产税税基的评估中并不关键，这一点也和其他研究者的研究结论一致。

利用新建立的优化后的随机森林模型对测试集 32 个样本进行住宅价格评估，具体的结果和模型评估结果如表 5.7 所示，检验如表 5.8：

表 5.7 随机森林评估预测结果

序号	销售价（万）	预测价格（万）	差异度	绝对误差
1	120	124.26	1.04	4.26
2	153.6	147.85	0.96	5.74
3	93	104.46	1.12	11.46
4	160	154.22	0.96	5.78
5	158	152.10	0.96	5.9
6	160	157.59	0.98	2.41
7	169	163.44	0.97	5.56
8	160	162.07	1.01	2.07
9	139	138.45	1	0.55
10	124	126.46	1.02	2.46
11	144	146.01	1.01	2.01
12	150	147.63	0.98	2.37
13	120	126.28	1.05	6.28
14	125	125.86	1.01	0.86
15	152	152.16	1	0.16
16	153	148.34	0.97	4.67
17	113	117.08	1.04	4.078
18	128	124.68	0.97	3.32
19	123	123	1	0
20	142	140.53	1	1.47
21	166	151.17	0.91	14.83
22	125	123.88	0.99	1.12
23	122	122.68	1.01	0.68
24	83.8	94.45	1.13	10.65
25	89.5	94.26	1.05	4.76
26	92	94.34	1.03	2.34
27	87	92.92	1.07	5.92

续表 5.7

序号	销售价（万）	预测价格（万）	差异度	绝对误差
28	91.5	95.04	1.04	3.54
29	110	112.48	1.02	2.48
30	193.5	162.77	0.84	30.73
31	96	112.73	1.18	16.73
32	102	117.06	1.15	15.05

使用优化后的随机森林模型可以得出准确的预测结果，本文选择了 32 个样本作为预测集，主要用来检验模型评估结果与实际值的差异情况。在表 5.7 中，测试集预测出来的差异度都十分接近 1，说明预测价格与实际价格差异较小；绝对误差也表现出较小的差异，但是预测检验第 30/31/32 号住房的绝对误差较大，在原始数据中他们的房屋面积分别为 173/96/124m^2，房屋类型为复式住宅，房屋面积较大，与其他房产存在建筑结构差异，不过总误差分摊到单价误差就较小。由于优化的随机森林模型本身具有不会因为异常数据干扰影响最终结果精确度的特性，所以最后预测出的综合价格误差也是合理的。为了减小误差，建议在未来评估房产税税基中，按照不同类型的房屋对其评估，保障同类型的房地产公平征收房产税。

表 5.8 应用优化后的随机森林模型评估房产税税基测试样本的检验描述

拟合优度	0.928
平均平方根误差	1.487926
平均绝对误差（万元）	5.632652
平均相对误差	0.045836

数据来源：作者整理

从上表中可以得出，在优化前的随机森林模型拟合优度为 91.8%，而优化后模型的拟合优度则达到 92.8%，说明了 turnRF()函数可以较好的增强模型预测的准确性，从而完成良好的预测效果。本文为了减少后续的计算工作量，直接选择住宅的总价作为模型的因变量，平均绝对误差约为 5.63，误差值略大，但是折算

到单价中就明显减小，该误差为正常误差。优化后模型的平均平方根误差小于 2，结果在合理范围内。从上文中可知，预测集能够在一定程度上反映训练集，所以训练集的预测误差也在合理范围内。上表中部分住宅价格预测准确性较低，主要是由于本文选择的样本数据是随机抽取的兰州市住宅数据，其中包含了面积较大的别墅住宅、公寓住宅、着急出售的住宅等，其与正常价格存在一定差异，所以导致预测时存在误差较大的情况。据此，可以细分各市场，将不同类型的住宅进行区分，按类别构建模型；针对急售房屋，经过筛选后，可适当剔除，以此提高模型的准确性，这也为后续房产税税基评估提供研究思路。

在以上检测预测结果的准确性良好后，通过优化后的随机森林模型预测样本的每套住宅价格，通过公式（5）计算出预测样本 638 套住宅的均值为 1.327 万元/平方米，以该值作为兰州市房产税基价格，建议将该价格应用于兰州市房产税税基价格征收标准中。

5.6 与传统多元回归评估对比讨论

通过文献梳理发现，学者们对多元回归模型研究较多，并且已有大量学者验证了多元回归模型可以应用到估价中，且评估效果良好，故本文选择了该模型作为对比模型，分析其与随机森林模型优化前后的结果差异，多元线性回归模型的函数表达式如下：

$$P = a_0 + \sum a_i z_i + \varepsilon \qquad (6)$$

其中：P 为住宅税基的市场总价，Z_i为住宅的特征变量。

采用 SPSS 26，逐一对随机森林模型选取的变量进行筛选，具体如下表 5.9 所示：

表 5.9 多元回归模型确定模型特征变量

模型	R	R 方	调整 R 方	标准估计的误差	Durbin-Watson
1	.881	.777	.772	22.07947	1.428

a. 预测变量: (常量)，是否有大学，公交线路条数，总层数，附近公园，房间数，距 CBD 距离驱车时间，装修，绿化率，小区环境，房龄，建筑面积，物业费。

b. 因变量: 住宅总价

数据来源：由 SPSS 运行得出

最后选取了 12 个特征变量作为多元回归模型的特征变量，即附近大学、公交线路条数、总层数、附近公园、房间数、距 CBD 距离、装修、绿化率、小区环境、房龄、建筑面积、物业费。从上表中可以看出,该模型的拟合优度为 0.777，低于优化后的随机森林模型。本文对多元回归模型的显著性检验结果如表 5.10 所示：

表 5.10 回归方程的显著性检验

模型	平方和	df	均方	F	Sig.
回归	1059250.056	12	88270.838	181.067	.000
残差	304689.447	625	487.503		
总计	1363939.504	637			

a. 因变量: 住宅总价

b. 预测变量: (常量), 是否有大学, 公交线路条数, 总层数, 附近公园, 房间数, 距 CBD 距离, 装修, 绿化率, 小区环境, 房龄, 建筑面积, 物业费。

数据来源：由SPSS运行得出

方程的研究有助于检验变量之间线性关系的显著性，如果F的sig值小于0.01，就可以说建立的模型线性显著，即税基的自变量与因变量之间存在显著的线性关系。如上表 5.10 所示，Sig 值为 0，那么数据中有自变量与因变量显著相关，所以数据集可以使用线性模型。在可以采用线性回归的基础上，本文还继续对模型的相关系数进行了检验，各特征变量的显著性情况如表 5.11 所示：

表 5.11 回归系数检验

模型	B	标准误差	标准系数	t	Sig	容差	VIF
(常量)	-111.693	12.460	.688	-8.964	.000		
建筑面积	1.272	.055	.078	23.171	.000	.405	2.467
房间数	7.531	2.779	.074	2.710	.007	.433	2.307
总层数	.317	.128	.051	2.478	.013	.400	2.498
装修	4.589	1.821	.117	2.520	.012	.888	1.126
房龄	1.023	.267	-.187	3.829	.000	.385	2.594
距CBD距离	-2.638	.293	.100	-9.016	.000	.827	1.209
公交线路条数	.767	.161	.062	4.775	.000	.809	1.236
绿化率	20.439	7.743	.117	2.640	.009	.649	1.540

续表 5.11

模型	B	标准误差	标准系数	t	Sig	容差	VIF
小区环境	7.202	1.684	.084	4.277	.000	.479	2.090
附近公园	5.428	1.302	.143	4.168	.000	.874	1.144
物业费	13.719	2.904	.112	4.724	.000	.389	2.572
附近大学	10.473	1.877	.112	5.579	.000	.880	1.136

数据来源：由 SPSS 运行得出

由上表可知，建筑面积、房间数、总层数、装修、房龄、距 CBD 距离、公交线路条数、绿化率、小区环境、附近公园、物业费、附近大学共计 12 个自变量回归系数的 t 检验显著水平都低于 0.05，所以这些自变量通过了显著性检验。从上述运行结果可以看出，除开距 CBD 距离这一变量外，其余变量与总价呈现出正相关，即特征变量数值越大，总价则越大；而距 CBD 距离与房价呈负相关，是因为在对该特征变量量化时采用的是住宅到 CBD 最短的距离，当这个值越大时，说明住宅到 CBD 越远，根据区位理论可以知道，到中心区域的距离越远价格则越低，所以表现出相反的关系。依据各项系数情况本文建立了关于多元回归模型的的回归方程，该回归方程为：

总价=-111.693+建筑面积×1.272+房间数×7.531+总层数×0.317+装修×4.589+房龄×1.023-距 CBD 距离×2.638+公交线路条数×0.767+绿化率×20.439+小区环境×7.202+附近公园×5.428+物业费×13.719+附近大学×10.473 （7）

运用该模型对预测集进行预测，得到的预测结果和模型拟合效果分别如下表 5.12 所示：

表 5.12 传统多元回归模型预测结果

序号	销售价（万）	预测价格	差异度	绝对误差
1	120	118.82	0.990167	1.18
2	153.6	132.83	0.864779	20.77
3	93	93.79	1.008495	0.79
4	160	144.82	0.905125	15.18
5	158	126.6	0.801266	31.4

续表 5.12

序号	销售价（万）	预测价格	差异度	绝对误差
6	160	168.68	1.05425	8.68
7	169	158.1	0.935503	10.9
8	160	158.1	0.988125	1.9
9	139	137.28	0.987626	1.72
10	124	136.68	1.102258	12.68
11	144	151.5	1.052083	7.5
12	150	152.77	1.018467	2.77
13	120	112.03	0.933583	7.97
14	125	147.05	1.1764	22.05
15	152	140.34	0.923289	11.66
16	153	141.92	0.927582	11.08
17	113	88.76	0.785487	24.24
18	128	130.78	1.021719	2.78
19	123	129.28	1.051057	6.28
20	142	136.26	0.959577	5.74
21	166	121.29	0.730663	44.71
22	125	136.3	1.0904	11.3
23	122	139.56	1.143934	17.56
24	83.8	90.91	1.084845	7.11
25	89.5	104.26	1.164916	14.76
26	92	125.86	1.368043	33.86
27	87	89.93	1.033678	2.93
28	91.5	81.56	0.891366	9.94
29	110	111.88	1.017091	1.88
30	193.5	226.91	1.172661	33.41
31	96	124.18	1.293542	28.18
32	102	147.23	1.443431	45.23

数据来源：作者整理

根据表 5.12 模型预测结果，分别计算平均平方根误差、平均绝对误差、平均相对误差，以各项误差反映出整个模型的预测效果，可以看出多元回归模型的平均绝对误差值较大，平均平方根误差约为 3.35，超过了合理范围 2,，故多元回归模型预测结果不理想，误差较大，具体误差值如下表 5.13 所示：

表 5.13 应用多元线性回归评估兰州市房产税税基评估拟合描述

拟合优度	0.777
平均平方根误差	3.34657
平均绝对误差（万元）	14.31688
平均相对误差	0.114785

数据来源：作者整理

通过将优化前后随机森林模型预测拟合度与传统多元回归模型预测拟合度进行对比，可以直观的看出，传统多元回归模型的拟合度是最低的，优化后的随机森林模型拟合度是最理想的，具体的预测结果拟合度如下表 5.14 所示：

表 5.14 模型评估拟合度对比

标准	传统多元回归模型结果	随机森林模型评估结果	优化随机森林模型评估结果
拟合优度	77.7%	91.08%	92.8%

数据来源：作者整理

随机森林模型、优化后随机森林模型、传统多元回归模型三个模型都能达到税基评估的目的，并且拟合度和各项误差都在合理范围内，在评估兰州市房产税税基的过程中，优化后的随机森林模型预测主要表现为以下几点：

首先，与优化前的随机森林模型相比，优化后的随机森林模型误差更小，拟合度更高，匹配度表现更好，评估结果总体上表现得更加稳定。

其次，turnRF()能很好的提高模型预测的准确性，解决了手动挑选模型参数的问题。该种优化方法能够极好的运用于随机森林模型批量评估中。

最后，与传统多元回归模型相比而言，优化随机森林模型无需剔除不相关变量，评估效率大大提高，能够适合房产税税基的批量评估要求；另外，传统的多元回归模型在本文的模拟预测中表现出较大的误差，降低了模型预测的准确性，多元回归模型不适合兰州市关于住宅类的房产税税基评估。

5.7 对兰州市房产税税基评估结果应用建议

为了防止在开展房产税征收的过程中出现的“阴阳合同”带来的税收风险，本文在优化了随机森林模型后评估出了兰州市房产税税基的市场价值，针对该评估结果，作了如下应用建议：

（1）应用于房产税征税范围和对象

为了保障兰州市的刚需住房，同时遏制住宅投机行为，期望政府能把家庭视为一个单位，同时还针对部分家庭对购买的首套住房实行减免房产税制度，对于购买了两套或以上的住宅家庭按比例征收房产税。在征税对象上，也可设定减免面积基数，按居民户口计算减免面积后，余下的面积按照一定住房市场价格和税率进行房产税征收。

（2）确定税基市场价值

利用第四章节公式（5），计算出优化后随机森林模型预测的 638 套房屋样本的单价均值 1.327 万元，通过前面章节的分析与验证，该价值能客观的反映出兰州市房产税税基的市场价值。依据该市场价格，进一步征收兰州市住宅类房产税，《甘肃省房产税暂行条例实施细则》对住宅类存量住房房产税征收具体为：住宅房产税税应纳税额=（房屋面积-减免面积）*住房平均价格*税率，而此处的住房平均价格建议本文通过优化随机森林模型预测的平均住宅市场价格 1.327 万元/平方米。

（3）税基市场价格定期重估

按照房屋买卖过程中计税价格评估征管的相关规定，如果本地房产价格指数在某个时间内环比变化很大，那就必须及时去调整评估模型。本文建议税务局建立健全的房产税税基评估周期，以避免评估价值偏离市场价格，从而导致税负不公。

6 结论与展望

6.1 研究结论

本文对房产税税基及随机森林算法的相关理论和研究进行了梳理，分析总结了随机森林模型评估房产税税基的现状及问题，以房产税税基理论和随机森林算法理论为本文的研究基础，对比了优化前后随机森林模型与传统多元回归模型，对影响房产税税基的特征因素进行了分析，提出基于随机森林模型的优化，文中重点介绍了随机森林模型，并基于优化后的随机森林算法构建了兰州市房产税税基估值模型，并对近期兰州市的税基情况预测估计。具体的相关结论如下：

（1）优化随机森林模型可应用于兰州市房产税税基评估中。经验证，本文评估预测价值与真实价值相差不大，研究结论与实际结论一致，由此可以表明运用优化随机森林模型可以提高评估结果的准确性。

（2）turnRF()函数能有效提高随机森林模型的预测准确性。在随机森林模型预测效果好的基础上，本文继续对提高模型精确性进行研究，该函数运用于模型构建过程中，能使最终的拟合度提升，误差减小。从而确定该函数可以作为优化随机森林模型的一个方法。

（3）预测结果 1.327 万元/平方米可用于最近年度兰州市征收房产税的计税依据。本文搜集了 670 个数据，将 638 套住房作为本文研究训练集，32 个样本作为检验样本，并且最终检验结果表现良好。根据兰州市房产税征收政策，在预测了 638 套住房价格后按均值计算出 1.327 万元/平方米作为征收标准，为兰州市房产税征收工作提供参考。

6.2 研究不足与展望

本文虽通过优化的随机森林模型提高了应用的客观性与准确性，但论文中还是存在很多不足之处，并提出了未来展望：

首先，在整个样本搜集过程中，由于受限于样本搜集的难度很大，所以样本容量相对较小，另外在特征变量的设置上，依据的是几个典型的区域研究中常用的特征变量，我国的城市与区域之间房价的影响因素存在差异，所以体现的住宅

影响因素不够全面，希望在未来的研究中，实地调研房价的影响因素后再对模型的特征变量进行修正。另外，本文选择的是兰州市四区的住宅类房价，其分类的不同可能也对房价产生影响，故房地产相关的数据资料也有待细化。其次，本文是对兰州市的住宅类房产税税基市场价格的评估方法，其他城市和其他类型的房地产是否适用此法，需进一步论证。最后，只选择了优化前模型和多元回归模型对优化后模型进行预测结果对比，体现了优化后随机森林模型的预测准确性和处理干扰变量的能力，研究验证过程中应该将大量的估价模型进行对比分析，选择最优预测模型，故选取的参照模型较少，在进一步的研究中，希望可以将其他模型进行多对比研究。

参考文献

[1] Andy Liaw. Classification and Regression by Random Forest[J]. R News, 2002, 2(3):18-23.

[2] Antipov, EA, pokryshevskaya, EB. Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics[J].Expert Systems With Applications,2012,39(2): 1772-1778.

[3] Alison J.Iavarone. New York State Property Tax Assessments and the Homestead Option. The CPA Journal, state & local taxation, 2014, 5.

[4] Breiman, L.J.H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees[M]. New York: Chapman and Hall, 884.

[5] Breiman, L. Random forests[J]. Machine Learning, 2001, 45: 5-32.

[6] Breiman, L. Statistical modeling: The two cultures[J]. Statistical Science, 2001, 16:189-215.

[7] Carbone Robert, Longini Richard. A Feedback model for Automated Real Estate Assessment[J]. International Valuation Standards Seventh Edition, 2005: 11-13.

[8] IAAO. Standard on Mass Appraisal of Real Property[S]. 2013:1-24.

[9] Iverson, L.R., A.M. Prasad, S. N. Matthews, and M. Peters. Estimating potential habitat for 134 eastern US tree species under six climate scenarios[J]. Forest Ecology and Managemen. 2008, 254:390-406.

[10] John D. Benjamin, Randall S. Guttery and C. F. Sirmans. Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation[J]. Journal of Real Estate Practice and Education,2004, 32: 65-77.

[11] Kontrimas, V, Verikas, A. The mass appraisal of the real estate by computational intelligence[J]. Applied Soft Computing, 2011, 11(1): 443-448.

[12] Lancaster, K.J. A new approach to consumer theory[J]. Journal of Political Economy, 1866, 74: 132-157.

[13] La riviere B, Poel D V D. Predicting Customer Retention and Profitability by Using RandomForests and Regression Forests Techniques[J]. Expert Systems

with Applications, 2005, 29(2)：11-15.

[14] Lee Junsoo, Kwak Seung-Jun List, John A. Average Derivative Esti-mation of Hedonic Price Models[J]. Environmental and Resource Eco-nomics, 2000, 3(16):81-91.

[15] Mc Cluskey, W.L, Wiliams, B. “Introduction: A Comparative Evaluation” in Mc Cluskey W.L.(ed)[J]. Property Tax: An international Comparative Review, A shgater Publishing Ltd. 1999.

[16] Marvin Anderson, Robert L. Brown, Marion R Johnson, William J. Mc Cluskey. Property Tax in Malaysia and South Africa: A Question of Assessment Capacity and Quality Assurance. Journal of Property Tax Assessment & Administration. 2013,Volume 10,Issue 4.

[17] Nicolai, Meinshausen. Quantile Regression Forests[J]. Journal of Machine Learning Research, 2006, 7(6), 983-999.

[18] Robert J. Gloudemans. Comparison of three residential regression models: Additive, multiplicative, and nonlinear[J]. Assessment Journal, 2002(6).

[19] Robert J Gloudemans. Comparison of three residential regression models: Additive, multiplicative, and nonlinear[J]. Assessment Journal, 2002, 8(7), 25-35.

[20] Rosen, S. Hedonic prices and implicit markets: Product differentiation in pure competiton[J]. Journal of Political Economy, 1874, 82(1): 35-55.

[21] Sexton J, Laake P. Standard Errors for Bagged and Random Forest Estimators[J]. Computational Stastistics, 2008, 2(3): 841 -860.

[22] Sabaliauskas, Kestutis and Albina Aleksiene. Progress Toward Value-Based Taxation of Real Property in Lithuania, Land Lines Article, October 2002, Volume 14, Number 4.

[23] Willian G Hardin III, Marvin L Wolverton. An introduction to the analysis of covariance model using an empirical texst of foreclose status on sale price[J]. Assessment Journal, 1999(6): 31-33.

[24]艾钰.基于随机森林优化的沪深 300 指数走势预测研究[D].济南:山东大学,2020.

[25]陈诗沁,王洪伟.基于机器学习的房地产批量评估模型[J].统计与决策,2020,36(09):181-185.

[26]陈奕佳.基于随机森林理论的北京市二手房估价模型研究[D].北京:北京交通大学,2015.

[27]程亚鹏.我国城市住房价格测度:Hedonic 方法与实证[D].重庆:重庆大学，2010:3-7.

[28]陈小悦,孙力强.关于建立中国房地产税批量评估系统的几点思考[J].财政研究,2007(12):48-51.

[29]陈艳.批量评估技术在房产税税基评估中的应用分析[J].住宅与房地产,2019(27):4.

[30]曹正凤.随机森林算法优化研究[D].北京:首都经济贸易大学，2014.

[31]陈钊.基于随机森林模型的房产税税基批量评估研究[D].哈尔滨:哈尔滨工业大学,2015.

[32]崔志坤,吴迪,刘冰.关于推进房地产税改革的思考[J].税务研究,2020(05):62-65.

[33]董倩,孙娜娜,李伟.基于网络搜索数据的房地产价格预测[J].统计研究,2014,31(10):81-88.

[34]邓元东.基于马克思地租理论的中国城市房价形成研究[J].中国经济问题.2018.(6):3-12.

[35]方匡南.随机森林组合预测理论及其在金融中的应用[M].厦门:厦门大学出版社,2012：7-11.

[36]傅樵.房产税税基评估问题研究[J].财经界,2017(06):115-116+119.

[37]郭文华.为合理计算不动产税立陶宛引入计算机化批量评估系统[J].国土资源情报，2005(6)：9-12.

[38]杭州市财政局直属征收管理局课题组.房地产批量评税技术的理论探索与实践创新[M].北京：经济科学出版社，2009.

[39]纪益成,傅传说.批量评估：从价税的税基评估方法[J].中国资产评估,2005(11):5-9.

[40]李欣海.随机森林模型在分类与回归分析中的应用[J].应用昆虫学报,2013,50(04):1190-1197.

[41]邱少明,杨雯升,杜秀丽,王雪珂.优化随机森林模型的网络故障预测[J].计算机应用与软件,2021,38(02):103-109+170.

[42]孙健夫.关于物业税负担水平设计的几个问题[J].税务研究,2010(04):39-40.

[43]时文静.基于 Lasso 与数据挖掘方法的影响北京二手房价格的因素分析[D].北京:北京工业大学,2017.

[44]王晶晶.支持向量回归在房地产税税基批量评估中的探究[D].厦门:厦门大学,2018.

[45]吴喜之.复杂数据统计方法——基于 R 的应用[M].北京:中国人民大学出版社，2012.

[46]邢彪.基于粗糙集的随机森林算法优化研究[D].成都:成都理工大学,2019.

[47]徐戈,张科.基于随机森林模型的房产价格评估[J].统计与决策,2014(17):22-25.

[48]玄永生,王建忠,王余丁.我国房产税税基评估问题研究[J].环渤海经济瞭望,2011(04):57-59.

[49]薛震，孙玉林.R 语言统计分析与机器学习[M].北京:中国水利水电出版社，2020.

[50]叶发强,陈西婵.重庆房产税试点改革的实效分析[J].西部论坛，2014,24(1):46-52.

[51]杨沐晞.基于随机森林模型的二手房价格评估研究[D].长沙:中南大学,2012.

[52]周彩英,周艳秋.房地产税税基评估区位理论综述及启发[J].内蒙古财经大学学报,2019,17(03):6-9.

[53]周琳娜.基于物业税开征目的的批量评估系统研究[D].厦门:厦门大学,2007.

[54]周淑敏.积极推动不动产税基评估工作实施税收制度的重大改革[J].中国资产评估，2004(3):41-45.

[55]周淑敏.房地产税收征收中的税基评估方法[J].中国资产评估,2005(7)：39-44.

[56]张望舒,马立平.城市二手房价格评估方法研究——基于 Lasso-GM-RF 组合模型对北京市二手房价格的分析[J].价格理论与实践,2020(09):172-175+180.

[57]张辉.房地产税基评估研究——以甘肃省 86 个县为例[J].中国物价，2021（06）:87-90.

附录

Likert 评分表

较差	差	良	中	优
1	2	3	4	5

致 谢

行文至此，意味着我的研究生生活，我的学生生涯，即将落幕。始于二零一九年金秋，终于二零二二年盛夏。

桃李不言，下自成蹊。感谢我的导师选择了我，入师门三年，老师总是悉心指导学习和论文中遇到的问题，并在毕业论文从选题到撰写再到修改都给予我极大帮助，学生感恩于心。祝福我的导师及家人身体健康，工作顺利！

山水一程，三生有幸。感谢师兄在论文上的指导和帮助；感谢同门在学习和生活中的帮助，跟你们一起学习，一起旅游，我的读研生活丰富多彩；感谢好友的鼓励与陪伴，一起经历、一起成长，在我需要你们时你们都在，跟你们做过的每一件事都值得纪念一生；感谢一直以来帮助和关心我的人；祝我们保持热爱，高处相见。

借此机会我要特别感谢父母、爷爷奶奶对我的无微不至的照顾，在背后默默地支持我的学业。还要感谢我的弟弟和堂弟，感谢你们时常的挂念与关心。祝我的家人身体健康，平安顺遂。

最后感谢答辩组的老师，感谢各位专家学者能够抽出宝贵的时间对我的论文进行查阅，也感谢各位老师们提出的宝贵意见和建议。

感激之情，溢于言表，凡此种种，皆铭记于心。