DOI:10.13546/j.cnki.tjyjc.2014.17.061

基于随机森林模型的房产价格评估

徐　戈[1,2],张　科[1]

(1.中南大学 商学院,长沙 410083;2.湖南商学院 国际教育学院,长沙 410205)

摘　要:文章以住宅性房产为切入点,基于特征价格理论,选择了区位特征、建筑特征和邻里环境三大类共计21项特征变量,构建了房价格评估的随机森林模型。并以广州市天河区某区域的二手住宅性房产为例,应用所构建的模型对该区域内的二手房产进行了价格评估。

关键词:房产价格评估;随机森林模型;特征价格理论;广州市天河区

中图分类号:F293.3　　**文献标识码**:A　　**文章编号**:1002-6487(2014)17-0022-04

0　引言

经济发展和快速城镇化导致城市可供开发利用的土地越来越少,加之房地产市场成熟度的日益提高,二手房产交易日渐活跃。在成熟市场国家,二手房成交量往往是一手房成交量的数倍。当前我国大部分城市新房成交仍占主导,但在部分发展较快的一线城市(如上海、广州等),已出现二手房成交量接近或超过一手房成交量的现象,二手房市场处于由量变到质变的临界点。二手房交易活动的日益活跃,催生出了如何科学合理评估房产价格的重要问题。同时,在房地产市场宏观调控的背景下,试点于重庆和上海的房产税政策有望在全国各大中小型城市中全面起征。房产税的征收过程中,涉及到一个基本问题——房地产税基的确定,也就是房产价值的评估问题。由于不同的房产在区位、朝向、面积等方面可能存在相似之处,而无论是二手房交易价格的确定还是房地产税基的计算,对每一套房产进行单独估价的人力成本和时间都是非常高昂的,因此,房产价格的批量评估成为解决问题的可行手段。

由于房地产评估中涉及的特征变量众多,特征变量与价格的关系复杂,导致神经网络算法、支持向量机算法等常用智能算法存在较强的样本数据依赖,而且评估精度也存在一定波动。这就需要更新更合理的机器学习方法来更好地挖掘特征变量与价格的关系,从而取得更好的评估效果。因此,需求样本数据少、分类速度快、能显著提高预测精度的随机森林模型开始被学者尝试引入房地产价值评估领域。E.A. Antipov, E.B. Pokryshevskaya(2012)首次尝试将随机森林模型应用到住宅价格的评估中,实证研究发现,与回归树、多元回归和神经网络算法相比,随机森林模型估价的表现是最好的[1]。

作为一种在房地产估值研究领域的全新方法,随机森林模型方法在中国房地产市场应用的合理性和优劣特征都亟待考证。因此,本文拟采用随机森林模型研究中国的房产估值问题,并与传统评估模型进行比较,为二手房产价值评估、房产税税基确定以及开发商房屋开发建设决策提供理论依据和实践指导。

1　房产价值评估的随机森林模型

1.1　基于特征价格理论的变量选取

按照用途的差异,房地产被划分为住宅、商业地产、住宅、旅游酒店等类型,为探索方便起见,本文选择住宅性地产为估值对象展开研究。国内外学者对房地产特征价格方面的研究已经形成了较为丰富的研究成果[2,3],他们在实证研究中提取了丰富的住房特征变量。然而,由于地域、文化、经济水平等方面的差异,国外学者常用的某些特征变量(如地下室、花园等)并不完全适合国内住房的实际状况。因此,结合国内外学者的研究成果和中国房地产市场的实际情况,笔者将影响中国住宅房产价格的特征因素分成了"区位特征"、"建筑特征"和"邻里环境"三大类共计21项,如表1。

1.2　随机森林模型的建立

随机森林模型是Breiman(2001)在建立分类和预测模型时首次提出的[4]。在计算中,随机森林由众多单棵分类回归树(CART)进行组合得到,然后通过投票法的运用可得到最后的分类结果。该算法具有需要调整的参数较少、不必担心过度拟合、分类速度快、能高效处理大样本数据、能估计特征因素的重要性以及有较强的抗噪音能力等特点。与传统的线性回归方法不同的是,使用随机森林进行

基金项目:国家自然科学基金委创新研究群体科学基金资助项目(70921001);国家自然科学基金委重大国际合作项目(71210003);教育部哲学社会科学研究重大课题攻关项目(08JZD0016;10JZD0020)

作者简介:徐　戈(1978-),女,湖南长沙人,博士研究生,讲师,研究方向:房地产理论,绿色经济。
张　科(1979-),男,湖南长沙人,博士研究生,研究方向:房地产理论。

表1　　影响二手房价格的特征变量及其含义

特征分类	变量名	变量含义
区位特征	到CBD时间	从住宅小区开车到CBD的最短时间
	公交线路	住宅小区周围公共交通的便利程度
	地铁站	住宅小区周围地铁站的个数
建筑特征	建筑面积	二手房的建筑面积
	卧室数	二手房的卧室个数
	所在层	二手房所在楼层
	总层	二手房总楼层
	朝向	二手房朝向
	容积率	住宅小区的容积率
	装修	二手房的装修程度
	建筑年龄	二手房的房龄
	小区户数	住宅小区总户数
邻里环境	绿化	住宅小区的绿化质量
	卫生	住宅小区的卫生质量
	空气	住宅小区的空气质量
	小区周边自然环境	周边自然环境质量
	物业管理费率	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
	生活配套	住宅小区附近生活配套设施
	临近学校	住宅小区是否临近学校
	停车位	住宅小区的停车位情况

研究能够充分体现其数据挖掘的优势，且用该方法不用对函数形式事先进行假定，避免了假设误差。

运用随机森林方法进行房产价值进行评估研究，实际上是基于其样本内高精度拟合学习规律，样本外高置信度推广知识的能力。随机森林具有分类和回归两种技术，本文研究的不动产价值评估属于回归预测问题，随机森林回归（RFR）的基本思想是：首先利用自助法抽样，从原始数据集抽取B个样本，且每个样本的样本容量都与原始数据集相同；然后对B个样本分别建立B棵树，得到B个结果；最后，对这B个结果取平均值来得到最终的预测结果。基于随机森林的房产估价模型计算过程如下：

房产估值的随机森林是B棵树 $\{T_1(X), \ldots, T_B(X)\}$ 的集合，其中，$X=\{x_1, \ldots, x_p\}$ 是住宅房产的P维特征向量。集合将会产生B个结果 $\hat{Y}_1 = T_1(X), \ldots \hat{Y}_B = T_B(X)\}$，其中 $\hat{Y}_b, b=1, \cdots, B$ 为第b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。给定一列特征变量的数据进行训练，$D=\{(X_1, Y_1), \ldots,(X_n, Y_n)\}$，$X_i\ i=1, \cdots, n$ 指住宅房产的特征向量，而 Y_i 指房产的挂牌价格，随机森林回归算法实现流程为：

（1）原始数据样本含量为n，应用bootstrap有放回地随机抽取B个自助样本集，并由此构建B棵回归树，每次bootstrap抽样未被抽到的样本组成了B个袋外数据（out-of-bag，OOB），作为随机森林的测试样本；

（2）设原始数据的变量个数为P，则在每一棵回归树的每个节点处随机抽取 m_{try} 个变量作为备选分枝变量，然后在其中根据分枝优度准则选取最优分枝。在随机森林回归中，参数 $m_{try}=P/3$，在这个方法中，m_{try} 是唯一的调整参数。树可以最大化的生长，无须剪枝；

（3）重复上面的步骤，直到B棵树全部建好。

完成以上步骤，随机森林的训练集就建好了。最后，把测试集的自变量输入到建立好的基于随机森林房产估值模型中，得到房地产的估价结果。模型的基本思路如图1所示：

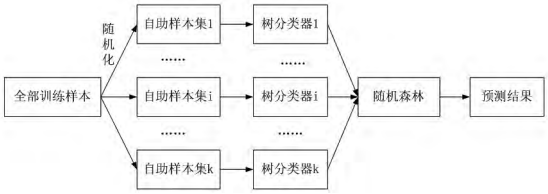

图1　二手房产估值的随机森林模型

1.3　特征变量重要性评价

对于随机森林回归中的变量重要性评分，使用基于permutation随机置换的残差均方减小量进行衡量，其具体过程为：

（1）对每一个bootstrap抽取的自助样本建立一个回归树模型，同时使用该模型对相应的袋外数据OOB进行预测，得到B个袋外数据的残差均方，记为 $MSE_1, MSE_2, \cdots, MSE_B$。

（2）变量 X_i 在B个OOB样本中的随机置换，形成新的OOB测试样本，然后用已建立的随机森林对新的OOB进行预测，与第一步的计算方法相同，得到随机置换后的OOB残差均方，得到如下矩阵：

$$\begin{bmatrix} MSE_{11} & MSE_{12} & \cdots & MSE_{1B} \\ MSE_{21} & MSE_{22} & \cdots & MSE_{2B} \\ MSE_{31} & MSE_{32} & \cdots & MSE_{3B} \\ \vdots & \vdots & \vdots & \vdots \\ MSE_{p1} & MSE_{p2} & \cdots & MSE_{pB} \end{bmatrix} \quad (1)$$

（3）用 $MSE_1, MSE_2, \cdots, MSE_B$ 与如上矩阵对应的第i列向量相减，平均后再除以标准误则得到变量 X_i 的重要性评分，即

$$\mathrm{sc}\,ore_i = (\sum_{j=1}^{B}(MSE_j - MSE_{ij})/b)/S_E, (1 \le i \le p) \quad (2)$$

随机森林通过给各特征随机地加入噪声干扰，观察模型准确率的变化，并利用准确率降低的幅度来衡量特征的重要性。若给某特征减少噪声后，模型准确率上升，则表明该特征重要程度较高。

2　基于广州市天河区的实证

2.1　样本区域选取

作为全国一线城市的典型代表，广州市房地产市场较为成熟，2012年广州二手住宅登记交易套数和面积分别达到5.9万套及494.3万平方米，广州市二手住宅全年交易宗数自2007年起已经连续5年超过了一手住宅。因此，本文选择二手市场成熟度高的广州市为样本选择对象。进一步的，由于天河区是广州市二手房房源最多、交易量最大和交易程度最活跃的行政区，同时考虑到样本覆盖面积和样本数量适中的原则，笔者将研究区域选定为天河区内东起广州环城高速，西至华南快速，南起珠江，北至广园快

速的梯形区域，该区域的总面积约为17平方公里，共涵盖有49个住宅小区。在该区域内，采用随机抽样的方式选择了二手房产的数据样本320个。

2.2 数据来源与量化处理

本文在实证过程中需要利用的数据包括二手房的挂牌价格数据和该二手房的价格特征数据（如表1的指标）。数据的来源途径主要包括：

（1）从全球排名第一的房地产网络平台——搜房网中的二手房数据库中选择二手房产的基本数据。这些基本数据包括：挂牌交易价格、小区名称、建筑面积、总价、建筑年龄、卧室数、住宅楼层、总层数、朝向、装修如何及小区总户数等。

（2）为获得住宅小区调查数据，笔者对研究区域内的住宅小区进行实地考察并走访小区居民。获取数据信息的主要方式是问卷调查，调查的主要内容包括小区周边自然环境、小区内部环境、小区内休闲运动配套情况、周边的生活教育配套气氛、小区的交通状况等。

（3）笔者还采用了百度地图软件测量了二手住宅所在小区距天河CBD（天河城广场）最短驾车时间，小区周边500米内公交线路的条数，和小区周边1000米内地铁站个数，用以反映住宅小区的区位特征。

经过以上数据收集和整理，笔者从320个样本中得到了298个无缺项的完整数据。然后对数据进行量化处理，量化处理的方式和过程为：1）对于建筑面积、物业管理费、卧室数、所在楼层、总层数、容积率、建筑年龄、小区户数、到天河CBD的驱车时间、公交线路条数、地铁站数量、装修豪华度12个变量，直接采用原始数据或进行简单变换；2）对于小区周边自然环境、小区内绿化、卫生、空气、停车位5个变量，采用Likert5级量化表予以量化；3）对于运动设施和生活配套2个变量，使用综合性指标进行衡量并根据所包含的内容进行评分；4）对于房屋朝向和临近学校2个变量，采用虚拟变量的方法予以量化。

2.3 实证分析

2.3.1 训练集和测试集的确定

从经过量化处理之后的298个数据样本中随机抽出269个数据作为训练集，用于建立随机森林回归模型；剩下的29个数据作为测试集，用于检测模型的评估效果。比较训练集合测试集数据的描述性统计结果表明：训练集样本与测试集样本在分布范围上比较一致，测试集样本基本能够反映实际市场的情况，能够将模型的特征因素对于二手房价格的影响程度较好的体现出来，可以用于特征价格理论的参数回归和随机森林方法的住宅价格预测效果的比较。

2.3.2 相关参数的确定

利用R中附带的Random Forest软件包程序对所构建的随机森林方法模型进行拟合回归预测。建立模型需要对参数 m_{try} 进行设定。根据A. Liaw(2002)理论[5]，建立随机森林回归的参数 $m_{try}=p/3$ 时表现较好，其中，p为特征变量个数。本文中，p=20，因此将 $m_{try}=6,7$ 分别测试，在对 m_{try}，$ntree$ 的最佳组合进行确定时，试算取最优是普遍所采用的方法。通过比较发现 $m_{try}=7$，$ntree=1000$ 时模型评估效果更好。

2.3.3 实证结果及分析

使用上述最优参数 $m_{try}=7$，$ntree=1000$，对269个样本进行训练，就能得到随机森林模型。在该样本训练集内，根据随机森林模型回归得到总价与挂牌价之间特征描述；同时运用以上所建立的随机森林模型对测试集中的29个样本进行二手房价格评估，训练集和测试集进行拟合评估得到的结果如表2所示：

表2 训练样本和测试样本结果描述

	训练样本	测试样本
拟合优度	0.9180	0.750
平均平方根误差	0.0034	0.0235
平均绝对误差	0.0435	0.0876
平均相对误差	2.78%	5.30%

通过表2可知，从模型整体上看，训练集样本的拟合优度为0.9180，测试集样本的拟合优度为0.750，说明随机森林具备较好的学习和挖掘样本数据规律的能力，同时也具备了良好的拓展效果。就平均平方根误差、平均绝对误差和平均相对误差而言，测试样本比训练样的误差偏大一些，但都在可接受的范围内。从单样本的预测匹配度和绝对误差来看，绝对误差和相对误差值都相对较小，对于该误差产生的原因，不排除在挂牌过程中由于某些房主急需资金或是个人预期不同导致挂牌价格并非完全市场化的情形。如果此类情况确有存在，可以通过引入反映这类异常情况的变量到其他特征中来进行相应的修正。此外，若更多的正常挂牌案例包含在所采集的样本中，市场实际和消费者的需求将得到更有效地反映，误差也会减少；若是有细分市场的存在，也可以在各细分市场的基础上将模型重新建立，可以减少误差。研究结果表明：在随机森林的预测模型的基础上，通过随机森林回归可以合理的估算房地产的价格；而且随机森林方法在模型的推广泛化上具有一定优势。

2.3.4 特征变量重要性排序

随机地给各特征变量加入噪声干扰，可以通过观察准确率降低的程度来衡量特征变量的重要性程度。在变量处加入噪声干扰时，把模型准确率所增加的均方差记为%IncMSE，该数值越大说明所对应的变量越重要。在本文的研究中，利用随机森林模型得到各个特征变量对二手房价的重要性具体如表3所示：

表3反映了影响样本区域二手房价格因素的重要性顺序，结果表明：在广州市天河区的这段样本区域内，“到天河CBD的驱车时间”、“物业管理费”、“空气”、“公交线路”、“建筑年龄”这几个特征变量的重要性名列前茅，是影响该地区二手房价格的最重要因素。研究结果一方面可以为二手房产交易商和二手房持有者在定价方面提供依据；另一方面，开发商在建设商品房时可以综合考虑以上

表3　　样本区域内二手房价格特征变量的重要性排序

重要性次序	特征变量	%IncMSE(所增加的均方差)	重要性次序	特征变量	%IncMSE(所增加的均方差)
1	到天河CBD的驱车时间	55.0416732	11	运动设施	17.3412411
2	物业管理费	34.230079	12	卫生	17.0118396
3	空气	24.992668	13	装修	16.3588693
4	公交线路	23.7539536	14	容积率	15.7929289
5	建筑年龄	22.3811542	15	绿化	15.703027
6	小区户数	21.7071577	16	停车位	15.6777592
7	总层数	21.4880079	17	地铁站	13.579617
8	卧室数	21.1764148	18	生活配套	9.793283
9	所在楼层	17.8225594	19	临近学校	5.1379321
10	小区周边自然环境	17.4893966	20	朝向	0.2750178

研究结果，在楼盘选址、建筑特性等方面进行有针对性的取舍，因为这是购房者在选购二手房产时最关注的因素。

值得说明的是，以上关于特征变量的重要性排序是基于所选择的样本区域的结果，而在其他区域或其他城市，如“地铁站”、“靠近学校”等因素可能会有更加重要的排序。也就是说，研究结果依赖于研究样本的实际状况。

3　与传统方法的比较分析

为了检验随机森林模型在房产价格评估中的优劣，本文选择传统的价格评估技术——多元线性回归模型来进行对比研究。房产价格评估的多元线性回归模型为：

$$P=a_0+\sum a_i Z_i+\varepsilon \quad (3)$$

其中 Z_i 代表房产的特征向量。

同样利用天河区数据样本，选用逐步筛选法确定最后进入回归方程的特征变量，结果如表4所示：

表4　　逐步筛选法确定特征变量

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
9	.936	.876	.872	.11893793

i. Predictors: (Constant), 物业管理费, 到天河CBD的驱车时间, 空气, 建筑年龄, 绿化, 装修, 生活配套, 小区户数, 朝向

由表4可以看出，最后进入多元线性回归模型的特征变量有9个，即物业管理费、到天河CBD的驱车时间、空气、建筑年龄、绿化、装修、生活配套、小区户数和朝向。对方程进行多元线性回归，得到如下回归结果：

二手住宅单价=1.486+0.192*物业管理费−0.026*到天河CBD的驱车时间+0.122*空气−0.022*建筑年龄+0.042*绿化+0.047*装修+0.037*生活配套+9.972*小区户数+0.035*朝向

对回归方程进行显著性检验，F检验表明：所有自变量的回归系数不同时为零，也就是说因变量和自变量全体之间确实存在线性关系，因此多元线性回归模型是合理的。运用该模型对训练集和测试集样本进行价格预测，结果如表5所示：

表5　　多元线性回归结果及比较

	全体样本−传统模型	训练样本−随机森林	测试样本−随机森林
拟合优度	0.645	0.9180	0.750
平均平方根误差	0.0274	0.0034	0.0235
平均绝对误差	0.1128	0.0435	0.0876
平均相对误差	7.40%	2.78%	5.30%

由表5可以得出，无论从整体的拟合优度来看，还是从平均绝对误差、相对误差、以及平均平方根误差的比较，应用随机森林模型进行评估的效果都要优于传统的多元线性回归方法，也从一个方面说明了应用随机森林进行房产价格评估的优越性，就整体效果看随机森林模型具有重要的实际应用价值。

4　结语

本文构建了房产评估的随机森林模型，并以二手住宅的价格评估为例展开了实证研究。在特征价格理论框架下，本文选择了区位特征、建筑特征和邻里环境三大类，包含“到CBD的驱车时间”等21个变量。通过对广州市天河区特定研究区域内的49个小区298个二手房挂牌案例进行实地调查、搜房网查询、百度地图计算等方式，获取了研究区域内二手房价格的特征价格数据，并展开实证研究。研究结果表明：训练集样本的拟合优度达到了91.8%，测试集样本的拟合优度达到了75.0%，随机森林模型适用于房产价格的评估。进一步的，本文将研究结果与传统多元线性回归模型进行了比较分析，研究结果表明：无论是训练样本还是测试样本，随机森林模型的预测精度都显著优于传统方法。本文的研究为二手房产的批量评估、房产税税基的批量计算等现实问题提供了一种科学有效的解决方案。

参考文献：

[1]EA Antipov, EB Pokryshevskaya. 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,(4).

[2]Lee Junsoo, Kwak Seung-Jun List, John A. Average Derivative Estimation of Hedonic Price Models [J]. Environmental and Resource Economics,2000,3(16).

[3]李贵孚,柳青.特征价格模型在信息商品定价中的应用[J].情报科学,2006,(11).

[4]Breiman L. Random Forests[J]. Machine Learning,2001,45(1).

[5]Andy Liaw. Classification and Regression by Random Forest [J]. R News,2002,2(3).

（责任编辑/亦　民）