2014年全国统计建模大赛

（第四届）获奖论文选

全国统计建模大赛执行委员会
国家统计局统计教育培训中心 编

中国统计出版社
China Statistics Press

2014年全国统计建模大赛

（第四届）获奖论文选

全国统计建模大赛执行委员会
国家统计局统计教育培训中心 编

中国统计出版社
China Statistics Press

图书在版编目(CIP)数据

2014年(第四届)全国统计建模大赛获奖论文选 /
全国统计建模大赛执行委员会，国家统计局统计教育培训
中心编. —— 北京 ：中国统计出版社，2015.7
ISBN 978-7-5037-7470-6

Ⅰ. ①2… Ⅱ. ①全… ②国… Ⅲ. ①统计模型－文集
Ⅳ. ①C8-53

中国版本图书馆 CIP 数据核字(2015)第 154573 号

2014 年(第四届)全国统计建模大赛获奖论文选

作　　者/全国统计建模大赛执行委员会　国家统计局统计教育培训中心编
责任编辑/张　赏
特约编辑/孙　慧　李　锐
封面设计/李雪燕
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲 6 号　邮政编码/100073
电　　话/邮购(010)63376909　书店(010)68783171
网　　址/http://www.zgtjcbs.com
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/710×1000mm　1/16
字　　数/465 千字
印　　张/26.5
版　　别/2015 年 7 月第 1 版
版　　次/2015 年 7 月第 1 次印刷
定　　价/52.00 元

编者说明

自2008年开始，国家统计局在全国政府统计系统每两年举行一次全国统计建模大赛，要在统计青年中营造创新进取、钻研业务的氛围，逐步提高全系统利用统计模型分析处理数据、运用数据的能力。

本次建模大赛是继2008年、2010年、2012年之后的第四届全国统计建模大赛。历时数月的紧张角逐，2014年9月由全国统计系统55支代表队、160多位青年统计人参加的第四届全国统计建模大赛在北京落下帷幕。经过论文提交、培训以及答辩等环节，大赛最终评选出特等奖1名，一等奖8名，二等奖10名。

2014年全国统计建模大赛着力于探讨大数据在政府统计中的应用，应对大数据时代对统计工作提出的机遇和挑战，跟进统计学的新发展，积极打造统计基础数据搜集“第二轨”。大赛于2014年5月发出通知，6月底之前提交参赛作品。2014年9月，大赛执委会选拔出20支参赛队进京参加决赛。经过两天紧张激烈的决赛答辩，2014年全国统计建模大赛结果于9月26日上午揭晓，并在国家统计局举行了隆重的颁奖典礼。9月29日在国家统计局首次召开了“大数据与统计建模”联网直播视频报告会，从获奖代表队中选出5支队伍就各自利用大数据进行统计建模的论文成果进行了主题汇报。各省统计局、国家统计局各调查总队均同期在线收看了这次主题汇报。

国家统计局副局长、全国统计建模大赛组委会主任委员张为民出席了“大数据与统计建模”视频报告会。他指出：国家统计局各有关专业司、各省统计局、国家统计局各调查总队要充分重视大数据时代为统计系统带来的机遇和挑战，深入研究大数据在统计工作当中的应用，奋力推进统计事业改革发展。

中国统计教育学会网站http：//www. sescn. org. cn开设了历届统计建模获奖论文展示专栏，内容包括历届统计建模获奖论文。本书编选了此次大赛的部分获奖作品，并且请有关参赛队员根据专家提出的意见逐一进行了修改和完善。尽管如此，作品中也难免有错漏不当之处，还望各位专家和读者批评指正。

全国统计建模大赛执行委员会

国家统计局统计教育培训中心

2015. 6

目录

一、全国统计建模大赛特等奖论文

房租指数编制中算法模型选择和应用

北京市统计局、国家统计局北京调查总队

冯艳、吴寒、肖京涛

摘 要

编制价格指数最核心、最基本的要求是同质可比，从而保证指数反映的是纯价格变化，而不包括质量改变带来的价格变化。为保证同质可比，对价格进行质量调整是价格指数构建中最重要的实践问题之一，它是决定价格指数精确性的关键问题，同时也是目前编制过程中最难以解决的问题。

本研究试图选择一种算法模型，解决房租指数编制中的同质可比问题。数据选取2013年1月—2014年5月北京主要大中型房屋中介公司的全部房屋租赁成交记录，以特征价格理论为基础构建模型。在综合考量数据建模和算法建模的基础上，选用算法建模方法建立模型。选择线性模型、决策树、Adaboost、Bagging、随机森林、神经网络、支持向量机等7种模型，通过多个数据集的比较验证，最终选择随机森林算法模型对房租进行评估预测，实现同质量的房屋在基期和报告期都有租赁价格，保证计算价格指数的样本质量不发生变化，由此计算出的房租指数反映纯价格变化。

该模型不仅可以解决房租指数编制中的同质可比问题，而且可进一步分析各种房屋特征对房租影响的重要性，同时，在数据基础完备的条件下，本研究的方法和思路可应用于价格指数编制中其他类别的质量调整。

关键词：算法模型　特征价格理论　价格指数　同质可比

一、问题的提出及研究意义

居民消费价格指数(Consumer Price Index，简称CPI)，是度量一定时期内居民消费的商品和服务价格水平变动的相对数，反映的是供求因素变化等引起的纯价格变动，不应反映商品和服务的质量变化。因此，“同质可比”是对样本的基本要求。然而在实践中，由于一些商品的特殊性或产品更新换代较快等原因，选取的样本在不同时期发生了质量变化。使用质量发生变化的样本计算价格指数，将影响指数的精确性。SNA(1968)就提出，在编制价格指数时应考虑产品质量的差

异。SNA(1993)进一步对产品质量差异和变化问题进行了讨论。

然而,剔除质量引起的价格变动——这一被称作是"质量调整"的工作非常困难,至今仍未得到很好解决。各国统计机构在编制价格指数时,通常使用可比替换、虚拟、样本更新、专家判断等方法处理样本质量变化问题。其核心思路或是忽略微小的质量变化,或是选取质量类似的样本进行替代,在很大程度上都存在人为判断,主观成分较多。很多发达国家采用特征回归法(即 Hedonic 法)进行质量调整,其核心是确定哪些特征会对产品的价格产生影响,通过构建一个函数进行回归估计,将产品特征与价格之间的关系加以量化,从而对价格进行质量调整。尽管该方法是目前质量调整中最有效的一种方法,但其也有一定缺陷。其中最主要的就是在构建函数时有一定的主观假设,假设商品的各个特征与价格之间是线性关系(或者经过变换成为线性)。而且,采用回归的方法估计函数时也需要满足一定的假设,如变量之间的独立性、服从正态分布等。这些统计学意义上的假设在真实的数据集中很难满足。

我国编制 CPI 的商品和服务项目是根据全国城乡近 13 万户居民家庭消费支出构成资料和有关规定确定的,包括食品、衣着、居住等 8 大类,262 个基本分类。CPI 就是由这些基本分类的价格指数通过权数加权平均计算而得。私房房租和虚拟住房估算租金是构成 CPI 的两个基本分类,而在这两个基本分类价格指数的编制过程中就存在不同质可比问题。由于房屋具有异质性和低流动性等特点,加之现行的价格统计是以抽样调查为核心方法,当月租赁成交的住房,在下一个月很难抽到完全相同的成交样本,目前主要采用专家评估、采价员判断等主观方法,对质量变化大的样本进行剔除、选择质量类似的房屋样本替换或者评估未成交的原质量房屋租赁价格。2010 年以来,全国 CPI 编制方法制度改革,私房房租和住房估算租金权数上调,在北京 CPI 的构成中,私房房租和住房估算租金的权数在 16%左右,成为仅次于食品的第二大影响因素。因此,解决房租指数编制中的同质可比问题,提高其精确性任务很迫切。目前,美国、瑞士、新西兰等国家使用 Hedonic 法对房租指数进行质量调整,我国尚未采用数理方法或统计模型等工具进行质量调整。

本研究改变编制价格指数中的抽样思维,尝试使用大数据,通过建立模型,解决房租指数的质量调整问题。通过挖掘[①]北京市房屋租赁成交数据,用总体而非抽样数据构建模型,评估预测房屋租金,实现同质量的房屋在基期和报告期都有租赁价格,保证计算价格指数的总体数据质量不发生变化,由此计算出的房租指数达到同质可比要求,从而能够更科学的反映北京房屋租赁市场租金的变化情况。此方法实现了对同质可比的判断从主观向科学的过渡,对提高居民消费价格

① 本文所有操作基于 Excel 和 R 软件实现。

指数编制方法的科学性具有重要意义。

二、相关理论综述

(一)特征价格理论

1967 年,Ridker 首次在住宅市场的研究中引入特征价格理论。近年来,随着房地产市场的迅猛发展,对房地产价格的研究越来越多,特征价格理论逐渐成为该领域的重要基础理论之一。该理论主要是将房地产看作一种产品,其效用则来源于房地产本身的一些特征变量,每个特征所带来的效用总和即形成了房地产价格的决定因子。该理论得出的"房地产价格差异是由其自身特征的不同引起的"结论也得到了学界广泛的认同[1]。

考虑到与买卖类似,房屋本身特征的不同也决定了房屋租赁价格水平的差异。本研究将特征价格理论应用到对房屋租赁价格的研究中,基于该理论构建模型,对房租进行评估与预测。

根据特征价格理论,可将影响房租的因素分为两部分——市场因素和房屋自身特征。市场因素包括供需结构、宏观政策等;房屋自身特征包括地理位置、配套设施、居住舒适度等。将各因素对房租的构成用函数表示:

$$P = f(X,Y) = g(x_1,x_2,\cdots) + h(y_1,y_2,\cdots)$$

其中:P 为房屋租金,X 代表市场因素,Y 代表房屋自身特征;

$g(x_1,x_2,\cdots,x_m)$ 表示市场因素对租金的影响函数;

$h(y_1,y_2,\cdots,y_n)$ 表示房屋自身特征对租金的影响函数。

在房屋租赁市场中,受信息不对称、存在时间交易成本、议价谈判能力不同等因素影响,实际的租金成交价 Z 与理论值 P 存在随机扰动,用 ω 表示,实际的租金成交价可表示为:

$$Z = P + \omega = g(x_1,x_2,\cdots) + h(y_1,y_2,\cdots) + \omega$$

房屋租赁价格指数反映的是纯价格变动,即市场因素对价格的影响,剔除由于房屋自身特征变动和随机扰动因素的干扰,理论上应在 Y 不变条件下,计算不同时期价格 P 的变动,即:

$$\begin{aligned} PriceIndex &= P_t/P_0 = P(X_t,Y)/P(X_0,Y) \\ &= (g(x_{1t},x_{2t},\cdots) + h(y_1,y_2,\cdots))/(g(x_{10},x_{20},\cdots) \\ &\quad + h(y_1,y_2,\cdots)) \end{aligned}$$

其中,P_t 和 P_0 分别表示基期和报告期的房租。

设通过模型对房租的拟合价格为 $\hat{P}$,如使用市场全部成交记录建立模型 M,$\hat{P}$ 完全等于 P;对于用样本数据建立的模型 M',$\hat{P}$ 是 P 的估计,若模型标准均方误差(NMSE)是可接受的,$\hat{P}$ 可近似看作 P。

利用基期和报告期实际房屋租赁成交记录分别构建模型 M_0 和 M_t，通过该模型拟合房屋在基期和报告期的租金，分别记为 $\hat{P}_0$ 和 $\hat{P}_t$。那么：

$$\Pr iceIndex = P_t / P_0 \approx \hat{P}_t / \hat{P}_0$$

(二)可供选择的算法模型

Breiman 概括了统计建模的两种类别：数据建模和算法建模。数据建模是传统统计中的建模方式，如 logistics 回归，线性回归和 Cox 回归等。算法建模则是整合数据挖掘和机器学习中的各类算法机理的建模方式。

在大数据时代，算法建模相对于数据建模有很大不同。以数据建模中常用的多元线性回归为例，该方法是在假设总体服从特定概率分布的基础上，建立线性函数关系，对参数进行估计。尽管该方法具有操作简便，容易解释，易于外延和分析等优势，但也存在很多先天性不足，如需要假定函数形式，对非线性关系的样本拟合性较差。

与传统数据建模不同，使用算法建模进行研究能够充分体现其数据挖掘的优势，不需要对函数形式进行事先假定，避免了假设误差，也不需要对数据进行独立、同分布等一些统计学意义上的假设。算法建模首先关注的是预测，其次才是模型解释。

1. 主要算法模型概述

根据本文的研究目的，适宜使用数据挖掘中的分类技术。用于分类的主要算法模型有决策树以及决策树的组合算法，如：Adaboost、Bagging、随机森林等，神经网络、支持向量机等。

(1)决策树(tree)

决策树算法是通过构造决策树来发现数据中蕴涵的分类规则。对离散变量做决策树被称为分类树，对连续变量做决策树被称为回归树。决策树的生产过程本质是对训练样本集的不断分组过程。首先由训练样本集生成决策树，其次对决策树进行剪技，即对上一阶段生成的决策树进行检验、校正的过程，主要是用新的样本数据集(称为测试数据集)中的数据校验决策树生成过程中产生的初步规则，将那些影响预测准确性的分枝剪除。

(2)Adaboost(Adaptive Boosting)

这是一种迭代式的组合算法，所用的基础分类器(如：决策树)一开始可能较弱，即出错率较高。然后，随着迭代，不断地通过自助法(bootstrap)加权再抽样，根据产生新样本改进分类器。每一次迭代都针对前一个分类器对某些观测值的误分缺陷加以修正，通常是在有放回抽取样本时，对那些误分的观测值增加权重，相当于对正确分类减少权重，这样在新的样本中就可能有更多的前一次分错的观测值，再形成一个新的分类器进入下一轮迭代。而且在每轮迭代时都对这一轮产生的分类器给出错误率，最终结果由各个阶段的分类器按照错误率加权(权重目

的是惩罚错误率大的分类器)投票产生。

(3)Bagging(Bootstrap Aggregating)

它是利用了自助法(bootstrap)的放回抽样,对训练样本做 k 次放回抽样,每次可抽取和样本量同样的观测值,由于是放回抽样,于是就有了 k 个不同的样本。进而对每个样本生成一个决策树,每个树都对一个新的观测值产生一个预测。如果目的是分类,那么由这些树的分类结果的多数产生 bagging 的分类;如果目的是回归,则由这些树的结果的平均得到因变量的预测值。

(4)随机森林(random forests)

随机森林是一个包含多个决策树的分类器,并且其输出的类别是由个别树输出的类别的众数而定。随机森林也是进行许多次自助法放回抽样,所得到的样本数目及由此建立的决策树数量要大大多于 bagging 的样本数目。与 bagging 的关键区别在于,在生成每棵树的时候,每个节点的变量都仅仅在随机选出的少数变量中产生。因此,不但样本是随机的,就连每个节点的产生都有相当大的随机性。随机森林让每个树尽量增长,而且不进行修剪。随机森林不惧怕很大的维数,即使是数千变量,它也不必删除变量。

(5)神经网络

神经网络是一种旨在模仿人脑结构及其功能的信息处理系统,是由大量的人工神经元按照一定的拓扑结构广泛互连形成的,并按照一定的学习规则,通过对大量样本数据的学习和训练,把网络掌握的"知识"以神经元之间的连接权值和阈值的形式储存下来,利用这些"知识"可以实现某种人脑功能的推理机。

(6)支持向量机(Support Vector Machine)

它是针对线性可分情况进行分析,对于线性不可分的情况,通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分,从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能。该算法基于结构风险最小化理论,在特征空间中建构最优分割超平面,使得学习器得到全局最优化,并且在整个样本空间的期望风险以某个概率满足一定上界。

2. 模型选择依据

交叉验证是机器学习中检验模型效果常用的方法。常用的交叉验证检验有:2 折交叉验证,5 折交叉验证,10 折交叉验证。以 10 折交叉验证为例,主要作法是将数据平均分为 10 份,每次取出 1 份作为测试集,剩下的 9 份作为训练集,最后将 10 次的平均结果作为预测误差的估计。在评价算法模型的回归性能时,常用的判断误差的标准为标准化的均方误差(NMSE),NMSE 最小的模型被选为最优模型。

$$NMSE = \frac{(y-\hat{y})^2}{/\ (y-\bar{y})^2} = \frac{\sum (y-\hat{y})^2}{\sum (y-\bar{y})^2}$$

三、模型的构建

(一)数据来源

尽管目前,网络是人们获取海量信息,取得大数据的主要来源,但对于计算房租指数而言,不能使用主流租房网站的数据计算价格指数最主要的原因是:租房与其他商品网上交易行为不同,价格并非网上所见即所得,交易需要在线下完成,网络数据为挂牌价而非实际成交价。因此,研究计算价格指数的数据应为实际交易完成后,取得的实际成交价。

北京市城镇居民租赁住房情况调查显示,目前中介公司在我市租赁市场中起到了重要作用,租住私房的家庭中,半数以上是通过中介租房,而且这一比例有继续上涨的趋势。尽管在租赁市场中还有一些是通过私人关系、小广告等其他非正规渠道租房,但中介公司作为正规的租赁企业,对市场的价格变化起到风向标作用,非正规渠道的房租价格会跟随着市场正规企业的变化而变化。因此,中介公司全部的房屋租赁交易记录能够反映房屋租赁市场的价格变化趋势。

本研究使用的数据集来自北京市住房和城乡建设委员会获得26家中介公司从2013年1月—2014年5月共24万余条租赁交易行政备案记录。这26家公司交易量较大,市场占有率达到90%以上。北京市住房和城乡建设委员会对房屋中介机构有监督管理的行政职责,因此该数据可信度高、是实际成交价,符合研究编制房租价格指数的基本要求。而且该数据集数据量大、结构多样、有实效性、有研究价值,符合大数据包含的四个V的特征(Volume数量、Variety多样性、Velocity速度、Value价值)。因此,可把本研究的数据集近似看作房租交易的大数据,使用机器学习、算法模型等大数据研究方法进行研究。

(二)数据预处理

该数据集中每条交易记录包含14个属性,可归纳为4方面:交易特征、房屋特征、配套设施和区位特征(见表1)。

表1 房屋租赁交易记录属性

属性分类	编号	属性名称
交易特征	1	记录编号
	2	成交时间
	3	成交价格
房屋特征	4	房屋类别
	5	楼层总高度
	6	所处楼层
	7	建筑面积
	8	房屋结构

续表

属性分类	编号	属性名称
配套设施	9	装修程度
	10	家具情况
	11	电器情况
区位特征	12	行政区划
	13	环路
	14	所处区域

目前尽管各中介公司执行房屋租赁备案制度，但有些记录登记不严格、不规范，数据库存在缺陷，原始记录较为混杂。在构建模型之前，对数据进行初步分析，并做如下预处理。

1. 属性合并归类——保证信息规范

房屋租赁交易备案记录中很多属性为人工填写，各个房屋租赁中介公司填报标准不统一，导致内涵相同的属性表述不一致，如：在“房屋类别”中，有“普通住宅”、“一般商品房”，“商品房”等相似内容。为使属性内容更为规范、简单、清晰，将其归并定义为“普通住宅”。在备案记录中，“房屋类别”、“房屋结构”、“装修程度”三个属性均有类似情况，分别进行合并归类（具体整理结果见附录 1）。

2. 属性重定义——保证信息准确

“所处楼层”、“总楼层”这两个属性有一定关联性，“所属楼层”的含义根据“总楼层”高度不同而不同。如：某记录中“所处楼层”为 6 层，对于总楼层为 6 层和总楼层为 18 层来说，意义不同。因此重新定义“楼层比例”替代“所处楼层”，准确反映房屋在整栋建筑中位置不同对租金的影响（见表 2）。

表 2　楼层比例设置规则

类　别	楼层比例所在区间
地下室	$(-\infty,0]$
低　层	$(0,0.33)$
中　层	$[0.33,0.66)$
高　层	$[0.66,1]$

注：楼层比例＝所处楼层/总楼层

3. 剔除异常值——保证信息合理

剔除数据集中明显不符合逻辑的异常值，但大数据的处理基于数据的多样性与混杂性，因此剔除条件设置的比较宽松。

条件 1：楼层总高度＞80 层[①]；

① 目前，北京最高楼国贸三期为 80 层。

条件2:建筑面积<10平方米;

条件3:成交价格<100元/月/套。

共剔除250条记录,剩余244418条。经过数据预处理后,数据集较为清晰,结构有所优化(见表3)。

表3　预处理后数据集情况

属性分类	属性名称	属性内容
交易特征	记录编号	I1～I244418
	成交时间	2013年1月－2014年5月
	成交价格	房屋租金
房屋特征	房屋类别	按房屋性质分为普通住宅、别墅、商铺等11类
	楼层总高度	所在楼宇的总高度
	楼层比例	地下室、低层、中层、高层
	建筑面积	房屋实际建筑面积
	房屋结构	按居室个数划分为1室、2室等11类
配套设施	装修程度	毛坯房、低档装修、中档装修、高档装修
	家具情况	有或无
	电器情况	有或无
区位特征	行政区划	16个①
	环路	2环内、2－3环、3－4环、4－5环、5－6环、6环外
	所处区域	100个②

4. 数据缺失值处理

在住房租赁行政备案记录中,“房屋类别”、“楼层总高度”、“所在楼层”、“房屋结构”、“装修程度”5个属性不是必填项,数据缺失情况较为严重(见表4、表5)。

表4　各属性缺失记录情况

	房屋类别	楼层总高度	楼层比例	房屋结构	装修程度
缺失数量(条)	119092	105244	105244	27125	199681
占比(%)	48.7	43.1	43.1	11.1	81.7

表5　数据缺失属性情况

缺失属性数(个)	记录数(条)	占比(%)
0	11332	4.6
1	53915	22.1
2	78183	32.0
3	66701	27.3
4	30091	12.3
5	4196	1.7

① 含北京经济技术开发区,不含房山。

② 依据北京市住房和城乡建设委员会划分标准。

若删除全部有缺失属性的记录，仅保留属性完备的记录，则数据集中只剩11132条记录，占原数据集的4.6%，数据损失过大。但由于缺失数据较多，全部进行插补不合理，因此，考虑仅对缺失1个属性的数据记录分别使用回归法、众数填补法、最临近替代法进行插补。

表6　插补、删除缺失值与全部数据对比表

指　标	类别	全部数据	回归插补	众数插补	临近插补	删除缺失
记录数	—	244418	65247	65247	65247	11332
租金（元/月/套）	最小值	150	200	200	200	200
	最大值	500000	351313	351313	351313	31000
	均值	4245	3919	3919	3919	3827.5
	中位数	3800	3600	3600	3600	3600
面积（m^2）	最小值	10	10	10	10	10
	最大值	3900	3900	3900	3900	356
	均值	71.66	68.7	68.7	68.7	69.2
	中位数	65	64	64.4	64.4	65
环路（%）	2环内	7.2	8.9	8.9	8.9	9.4
	2～3环	21.3	24.6	24.6	24.6	25.5
	3～4环	27.8	27.1	27.1	27.1	26.1
	4～5环	17.2	16.0	16.0	16.0	15.4
	5～6环	24.6	22.8	22.8	22.8	23.2
	6环外	2.0	0.6	0.6	0.6	0.5
区域（个）		100	93	93	93	82
房屋类型（%）	普通住宅	95.5	98.8	99.5	99.4	99.4
	平房	0.18	0.33	0.33	0.41	0.6
总楼层（层）	最高	70	70	70	70	41
	均值	12.77	13.07	13.07	13.07	13
	中位数	11	11	11	11	11
楼层位置（%）	地下室	0.4	0.4	0.4	0.4	0.3
	底部	30.6	30.7	30.7	30.7	31.8
	中部	33.2	33.2	33.2	33.2	32.7
	高部	35.8	35.8	35.8	35.8	35.3
房屋结构（%）	一居	35.9	35.2	34.7	35.6	36.1
	二居	46.8	48.1	48.9	47.4	48.7
	三居	15.5	15	14.9	15.3	13.8
装修程度（%）	一般装修	32.2	30	21.7	30.1	33.9
	精装修	65.6	68.6	76.9	68.4	64.7
家具配套（%）	齐全	45.1	58.4	58.4	58.4	18.5
	不齐全	54.9	41.6	41.6	58.4	81.5
电器配套（%）	齐全	45.1	58.4	58.4	58.4	18.5
	不齐全	54.9	41.6	41.6	41.6	81.5