2015 年第四届全国大学生
统计建模大赛

论 文 题 目 ：	大数据背景下基于网络搜索数据的商品房价格预测——以武汉市为例[1]
参 赛 学 校 ：	中南财经政法大学
参 赛 者 ：	张令令 孙金金 黄世祥
指 导 教 师 ：	李占风
完 成 时 间 ：	**2015** 年 **6** 月

中南财经政法大学

[1] 注:该论文获得由中国统计教育学会举办的“2015 年（第四届）全国大学生统计建模大赛”大数据统计建模类研究生组二等奖。

摘要

摘要：信息技术的发展及互联网的普及，使得消费者在做消费决策时更多的依赖网络搜索获取知识，网络搜索数据的趋势变化就直接反应市场需求的变化，并最终体现在商品市场价格的变化上，商品房房价的变化不外如是。本文基于这一思想，利用我国国内使用范围最广的百度搜索数据对武汉市商品房价格进行拟合和预测，选取与商品房价格搜索相关的 119 个关键词的百度指数，并用 Python 对关键词数据进行抓取。数据清洗时分别利用简单手动筛选、Pearson 相关系数筛选、线性回归筛选、逐步回归及 AIC 准则等步骤选出 8 个关键词作为最终自变量的关键词，分别是：公积金贷款额度，武汉公积金管理中心，公积金提取，按揭贷款利率，金地集团，武汉亿房网，租房子 58 同城，建材团购。实证研究部分对数据分别建立线性回归、回归树、bagging、随机森林和 SVM 等 5 个模型，并比较各模型的拟合和预测效果。结果显示：线性回归模型、随机森林模型和 SVM 模型拟合效果拟合效果较好且差异不大，其中以随机森林拟合效果最优，但在数据预测时，线性回归模型和 SVM 预测效果较好，平均误差率分别为 1.12%和 0.84%。本文所用方法可预测商品房房价领先官方发布数据 10-15 天。文章最后根据实证结果提出相关建议。

关键词：百度指数；搜索数据抓取；房价预测；随机森林；SVM

大数据背景下基于网络搜索数据的商品房价格预测
——以武汉市为例

一、问题描述

（一）研究的背景与意义

随着互联网的飞速发展，大数据时代已经悄然的进入了人们的生活当中，大数据开启了巨大的时代转型，就宏观经济分析而言，大数据时代带来的转变是重大且具有革命意义的[2]。大数据时代使得人们获取信息的渠道更加的宽广，从而影响着人们的行为，尤其是搜索引擎对人们的各项决策起到重要的作用。与传统的经济分析方法相比，运用海量数据的进行经济分析有明显的优势，从这些海量的数据中人们可以及时的了解经济的发展形势，正确的做出经济发展的预测，政府也可以合理的制定经济政策。

近年来，商品房价格的研究一直是各界学者关注的焦点，尤其是大中城市的商品房价格问题，更是我国政府、学术界、商界、广大购房者共同关注的重中之重。在传统的研究方法中，研究人员对数据的获取一般是从国家统计局、金融统计年鉴等,然而这些数据存在时滞性,如国家统计局关于2014年的数据会在2015年9月份公布。利用这些数据进行研究并提出政策建议明显不能满足政策制定部门的及时性需求。大数据时代，网络搜索行为直接预示经济发展趋势，网络搜索数据的时效性可以弥补传统方法的不足。目前，基于网络搜索数据的预测研究主要集中在股票、汽车等领域，关于房价预测的研究还不多见。少有的基于网络搜索数据的房价研究也是基于Google搜索指数，但国内主要运用的是百度搜索，所以Google搜索指数在国内并不具有代表性，所以本文利用百度指数对与武汉市商品房价格相关的关键词进行分析，并选取预测武汉市商品房价格的最优模型。

本文研究的意义是通过研究商品房价格及与其相关的网络搜索的关键词的相关关系，在众多的关键词中找到几个能解释商品房价格波动的关键词，从而使武汉市相关部门通过监测这几个关键词来了解商品房价格的变动趋势。利用线性回归、回归树、bagging、随机森林和SVM模型对商品房价格和关键词进行拟合，找到能精确预测房价的稳定模型，从而为相关部门对房价的预测提供最优的方案，由于武汉市房管局对武汉市商品房价格的公布一般是在次月的10—15号，

[2]刘涛熊,徐晓飞.大数据与宏观经济分析研究综述[J].国外理论动态,2015(01).

运用本文分析得到的预测模型能够使相关部门对商品房价格进行及时的预测，比官方数据提前 10—15 天得到武汉市商品房价格。

（二）国内外研究的现状

房地产价格预测成为近年来研究的热点，随着计量方法的成熟，国内外关于房地产价格预测的研究也逐渐增多。Browna[3](1997)等采用 Kalman 滤波的时序回归模型对英国房地产价格进行了预测，并证明其预测效果优于 VAR 和 ECM 模型；Gnirguis[4](2005)等综合采用指数平滑法、AR、VECM、GARCH、带自回归参数和随机参数的 Kalman 滤波等方法对美国房价的未来走势进行了预测，并通过比较验证证实带自回归参数的 Kalman 滤波和 GARCH 两种方法的预测效果最优；Selim[5](2008)采用人工神经网络和 Hedonic 回归模型对土耳其的房价进行预测，并认为人工神经网络的预测较优； Quigle[6]利用平行数据回归来分析与经济相关的指标来预测每个城市住宅价格走势；Hsu CC，Chen CY 和 Chia-Yon Chen 提出基于神经网络的灰色 GM(1，1)混合预测模型，但对具有季节性变化特征的数据无法使用这个模型来完成预测。国内关于房地产价格预测的研究主要是定量预测，武秀丽、张锋[7](2007)采用时间序列分析法，依据房地产价格时间序列的发展规律，进行房价的趋势预测，发现预测值与实际观测值的误差非常小，预测效果很好；王聪[8](2008)在考虑影响房价的各种因素的基础上，以我国 35 个大中城市相关数据为样本，采用多元回归分析和 LOGISTIC 分析两种方法构建大中城市房地产价格预测模型；王婧，田澎[9](2005)采用小波神经网络对房地产价格指数进行预测，并与指数平滑法和 RBF 神经网络预测做比较，结果表明采用小波神经网络能够更有效的预测房地产价格指数；程亚鹏，张虎，张庆宏[10](1999)用灰色预测方法 GM(1，1)模型建立了北京房地产价格指数的预测模型，并经过检验验证模型可靠；钱峰，吕效国，朱帆[11](2009)提出了一种结合非线性回归技术的灰色 GM(1，1)模型的改进模型，并通过预测我国的房地产价格指数验证了该方法

[3]Jane P,Browna, Haiyan Song & Alan Me Gillivray.Forecasting UK House Prices:A Time Varying Coefficient Approach[J].Econoinic Modelling,1997 (04).

[4]Harry S, Gnirgnis, Christos I, Giannikos & Randy I, Anderson.The US Housing Market: Asset Pricing Forecasts Using Time Varying Coefficients [J].The Journal of Real Estate Finance and Economics, 2005(01).

[5]Hasa Selim.Determinants of House Prices in Turkey: Hedonic Regression Versus Artificial Neural Network [J]. DogusUniversity Journal, 2008(01).

[6]Quinlan J R,Induction of decision trees.Machine Learning,1986,(01).

[7]武秀丽,张锋.时间序列分析法在房价预测中的应用[J].科学技术与工程.2007(11).

[8]王聪.基于多因素 LOGISTIC 的城市房地产价格预测模型研究[D].大连理工大学,2008(10).

[9]王婧,田澎.小波神经网络在房地产价格指数预测中的应用[J].计算机仿真 2005(07).

[10]程亚鹏,张虎,张庆宏.GM(1，1)模型在房地产价格指数预测中的应用[J].河北农业大学学报,1997(07).

[11]钱峰,吕效国.朱帆.灰色 GM(1，1)模型的改进模型在房地产价格指数预测中的应用[J].数学的实践与认识,2009(04).

的有效性和准确性；基于近几年数据挖掘技术的发展，不少学者也将这一技术利用到房价预测的研究中，倪大鹏[12](2013)结合本地房地产价格数据对房地产价格与影响房地产价格的重点因素之间的关系进行研究，实现合理的价格在多种因素影响下的预测；罗婧和朱建峰[13](2013)利用数据挖掘技术提取与房地产价格相关的因素，再构建空间面板模型进行弹性分析，实证结果得出与房价密切相关的因素；杨刚[14](2014)从现有的房地产交易数据中挖掘得到关联规则，从影响房地产价格的众多因素中提出关联度比较高的因素作为参数，并对模型有效性进行了验证。上述研究所用到的预测方法都是从对房地产经济影响因素出发，对历史样本数据的依赖性较强，随着互联网的广泛普及，基于互联网搜索数据的预测已经成为一个学术热点，例如 Ginsberg[15]等(2009)通过分析流感的看诊量和部门相关关键词的搜索量之间的相关关系，建立了以相关关键词的 Google 搜索量为基础的预测模型并取得了较好的预测效果；Lynn Wu 和 Erik Brynjolfsson[16](2014)发现谷歌的房屋搜索指数能够很好地预测未来住房市场的销售量和销售价格；杨树新，董纪昌[17]等(2013)以全国房屋销售价格指数为对象，研究了谷歌搜索关键词与房屋销售价格指数的相关性并通过实证检验得出房屋销售价格指数与提前 5 个月搜索指数的相关性最大；董倩[18]等（2014）以北京、上海、天津、重庆等 16 个大中城市的二手房价格和新房价格为研究对象，得到预测二手房和新房价格变动情况的最优模型。

二、指标的选取和数据描述

（一）指标选取的理论基础

互联网的普及，搜索使得互联网拥有丰富而又全面的超海量信息。用户在网上进行消费和搜索信息时，会将自己的搜索痕迹留在搜索引擎中，所有搜索记录的信息汇集在一起形成了庞大的数据库。这些数据中包含了用户的行为趋势，反应了用户在某一时段的经济或社会行为，由此我们可以对用户的经济社会行为的相关性进行研究。

[12]倪大鹏. 基于数据挖掘的房地产定价方法研究[D].大连理工大学,2013.

[13]罗婧,朱建锋. 基于数据挖掘与空间计量的房地产价格经验分析[J].开发研究,2013(05).

[14]杨刚.基于数据挖掘的房地产价格分析预测研究[D].南昌大学,2014(05).

[15]Jeremy Ginsberg, Matthew H, Mohebbi, Rajan S,Patel,Lynnette Bramme,Mark S,Smolinski,Larry Brilliant.Detecting Influenza Epidemics Using Search Engine Query Data[J]. Nature,2009,457(2).

[16]Wu L,Brynjofsson E,The Future of Prediction:How Google searches Foreshadow Housing Prices and Sales [C] .Working Paper,2014.

[17]杨树新,董纪昌,李秀婷.基于网络关键词搜索的房地产价格影响因素研究[J].2013(03).

[18]董倩,孙娜娜,李伟.基于网络搜索数据的房地产价格预测[J].2014(10).

对商品房价格的预测，要考虑商品房既是投资品也是消费品的特殊性，不论是作为投资还是消费，对于房地产开发商和消费者来说，房地产的投资建设和房屋购买都是一项重大的经济决策。在对商品房进行投资和消费时，房地产开发商和消费者都会考虑众多因素。对于消费者，在购买房屋之前会考虑与房屋相关的各种因素，由于互联网的便利性，人们更倾向于利用网络搜索了解相关信息。从消费品的角度来看，消费者购房是为了以后的居住，所以消费者关注的主要因素是：房屋的价格、户型、房贷利率、楼盘的地理位置、物业和房地产公司等。作为投资品，买房作为投资主要从两个方面获利，一方面是从房地产市场的涨价来获利；另一方面是将房屋进行出租获得租金，因此投资者主要关注的因素是，房屋的价格、市场利率、楼盘的地理位置等。因此人们在对商品房进行购买或者投资之前会利用搜索引擎来了解上述因素的信息，从而做出合理的决策。即人们的搜索行为与房价走势之间存在一定的相关关系，如图 1。

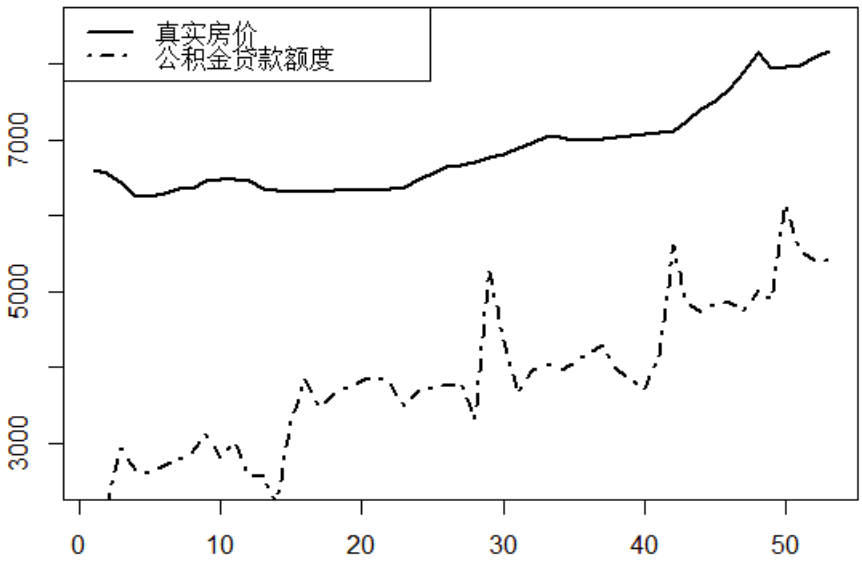

图 1 关键词“公积金贷款额度”和真实房价的趋势图

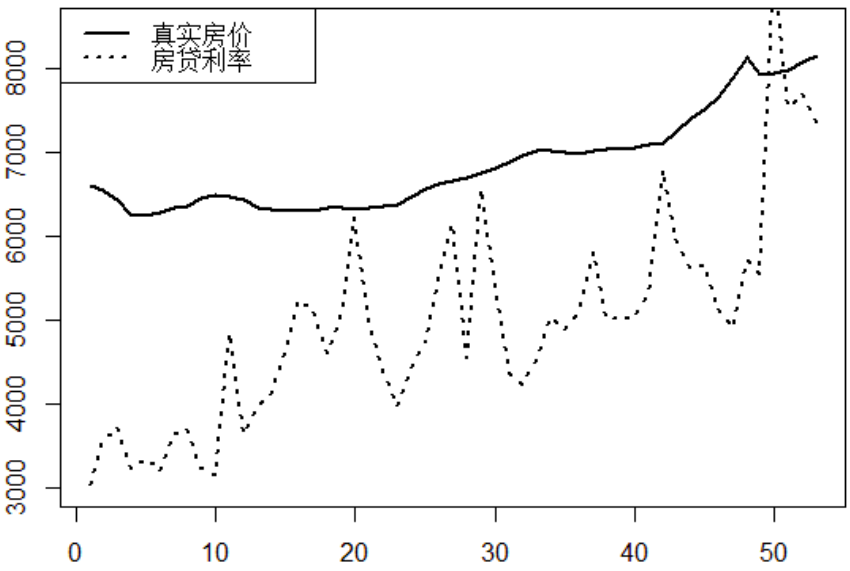

图 2 关键词“房贷利率”和真实房价的趋势图

由上图 1 和图 2 可以看到，忽略数量上差异，关键词“公积金贷款额度”、“房贷利率”和真实房价之间具有相似的发展趋势，因此可以推断上述两个关键词与房价有着较强的相关关系，所以我们认为可以根据关键词和真实房价的相关关系，选取与房价相关关系较大的关键词进行房价预测。

（一）指标的选取

1. 百度指数的介绍

百度指数是以百度用户的搜索行为为基础的，将用户的搜索行为记录下来，用以反映不同关键词在过去一段时间里的“用户关注度”和“媒体关注度”。用户可以从百度指数中提取某个关键词在百度的搜索规模的数据。百度指数的主要功能模块有：基于单个词的趋势研究（包含整体趋势、PC 趋势还有移动趋势）、需求图谱、舆情管家、人群画像；基于行业的整体趋势、地域分布、人群属性、搜索时间特征。百度指数主要向用户提供以下时间段的数据搜索量，“最近 7 天”、“最近 30 天”、“最近 90 天”、“最近半年”、和全部的数据搜索量。目前百度指数所能提供的时间段是 2011 年 1 月 1 日至今，本文选取的百度指数的时间跨度均为 2011 年 1 月 1 日——2015 年 5 月 31 日。本文仅介绍百度指数的趋势研究和需求图谱两个模块，以下以“武汉房价”这个关键词为例介绍百度指数这两个模块的搜索及数据的提取。

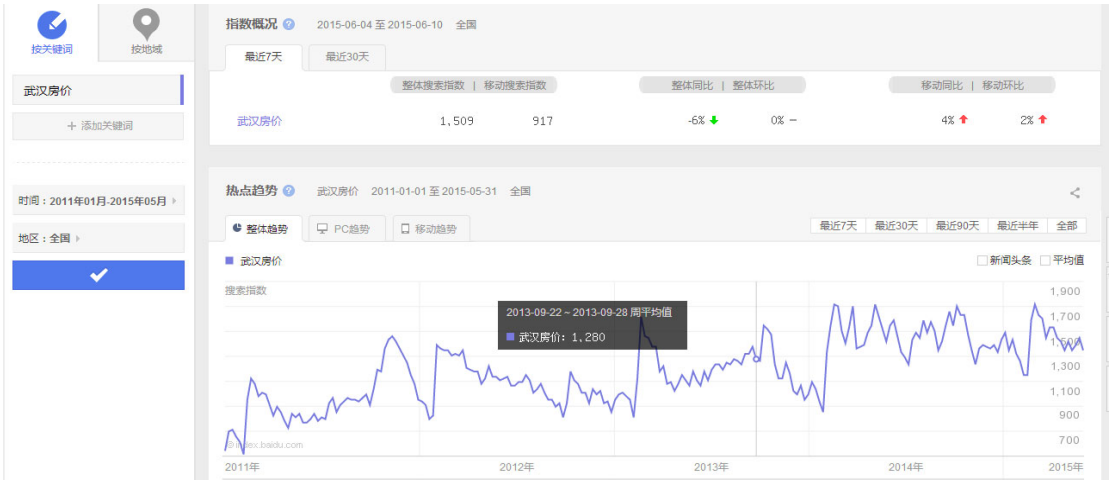

图 3 2011 年 1 月至 2015 年 5 月关键词“武汉房价”的周平均搜索趋势图

从图 3 中可以看出，在 2011 年 1 月份和 2015 年 5 月份这个时间段内，每年搜索量分别在 2011 年 11 月份、2012 年 2 月份、2013 年 2 月份、2014 年 5 月份和 2015 年的 3 月份出现高峰，数值分别为 1461、1401、162、1730、1729，由搜索量的数值可以看出，近几年关于武汉房价的整体搜索趋势是上升的。

图 4 与“武汉房价”相关的 15 个关键词及上升最快的搜索词

当点击需求图谱事就会出现上图，该图给出了与关键词“武汉房价”相关的 15 个关键词，以及根据搜索量排序得到 8 个上升最快的关键词，可以从这里提取少量与关键词“武汉房价”相关的关键词。

2. 指标的选取

根据指标选取的理论基础的分析可以知道用户的搜索行为反应了用户的近期的行为，但由于搜索用户的异质性，不同的用户对某一事物的关注度也有所不同，因此使得搜索的关键词也具有多样性，而且关键词应该包含与房价相关的各个方面，根据实际的生活经验选取人们买房主要考虑的因素进行关键词的选取，本文从 11 个方面进行关键词的提取，这 11 个方面涵盖了武汉房价、武汉房地产、房贷利率等，并将这 11 个方面作为初始的关键词，并运用百度关键词挖掘进行关键词的搜索，最终的到 119 个关键词，见附录 1。

（二）数据的描述

数据从百度指数的网页源代码中提取，点击鼠标右键审查元素可以看到制作网页的源代码，进一步找到代表图中数据的源代码，通过分析源代码进行数据的抓取。对于源代码数据的抓取是运用 Python2.7.10 进行操作的，代码见附录 2。剔除没有数据对应的关键词及数据量很少的关键词，得到的关键词的数据见数据包。武汉市商品房价格来自 wind 数据库，选取的时间区间是 2011 年 1 月至 2015 年 5 月。

对于得到关键词的数据进行处理，由于 2011 年 1 月份至 2015 年 5 月份共有 231 个周平均数据，由于本文研究是数据是月度数据，要将关键词的周度数据转

化为月度数据，即按照日历中的各月的周数进行加总，将每月最后几天的数据按所在周的天数作为权重，将该周的数据按权重分配到相邻的中，得到月度的数据。

四、模型的提出

本文的实证部分用到的机器学习算法有回归树、bagging、随机森林和 SVM 方法，将各种方法的理论部分介绍如下：

（一）回归树

当决策树的输出变量（因变量）是分类变量时，叫分类树，而当决策树的输出变量为连续变量时称为回归树。回归树不用假定经典回归中的诸如独立性、正态性、线性或者光滑性等等，无论自变量是数量变量和定性变量都同样适用[19]。当线性回归中关于自变量和因变量的线性关系假设不成立，且自变量为连续变量时，则通常可建立回归树模型，回归树模型的建立通过持续的（或递推的）分层将样本不断细分（亦即分枝），而分枝点是能够使得两分枝的反应变量的变异最大的预测变量的某个值，这样各节点内样本的同质性不断增强，最终达到节点内样本同质或由于样本数量过少无法继续分层，这里将终节点称为叶，而分枝开始的节点被称为根，一般既要保证回归树包含了足够的信息，又要把并不重要的枝节去掉。回归树模型的拟合效果一般较线性回归效果优，而且回归树模型的假设不如多元线性回归模型严格，适用范围更广[20]。

（二）bagging

bagging 算法的基础是重复取样，其通过产生样本的 Bootstrap 样本作为训练集，每一次随机地从大小为 n 的训练集中抽取 n 个样本作为此次的训练样本集。这种训练集被称作原始训练集合的 Bootstrap 复制，这种技术也叫 Bootstrap 综合，即 bagging。原始训练集中的某些样本可能在新的训练集中出现多次，而另外一些样本则可能一次也不出现。用该种训练集来训练分类器，产生 t 个分类器，然后采用投票表决方法来进行组合，产生一个最终的分类器。在这 t 个分类器中，单个的分类器识别率不一定非常的高，但是它们集成后的结果却有着很高的识别率[21]。相应的 bagging 算法的过程图如下：

[19]吴喜之.统计学:从概念到数据分析［M］.北京:高等教育出版社 ,2008.

[20]莫春梅,倪宗瓒,高凤琼.回归树的建模与应用[J].中国预防医学杂志,2002(09).

[21]张翔,周明全,耿国华,侯凡.Bagging 算法在中文文本分类中的应用[J].计算机工程与应用,2009(05).

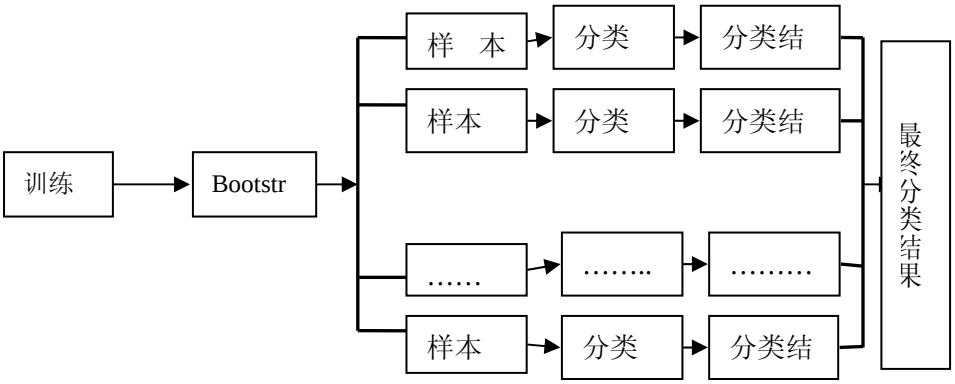

图 5 bagging 算法流程图

（三）随机森林

随机森林可以看成是 bagging 和随机子空间的结合，是由一系列的分类器组合在一起进行决策，期望得到一个最“公平”的集成学习方法。如图所示，构造每一个分类器需要从原数据集中随机抽取出一部分样本作为样本子空间，然后再从样本子空间中随机的选取一个新的特征子空间，在这个新空间中建立决策树作为分类器，最后通过投票的方法得到最终决策。这种方法的优点在于对于数据，它可以产生高准确度的分类器和处理大量的输入变量；在决定类别时，评估出变量的重要性，而且在建造森林时，它可以在内部对于一般化后的误差产生不偏差的估计[22]。

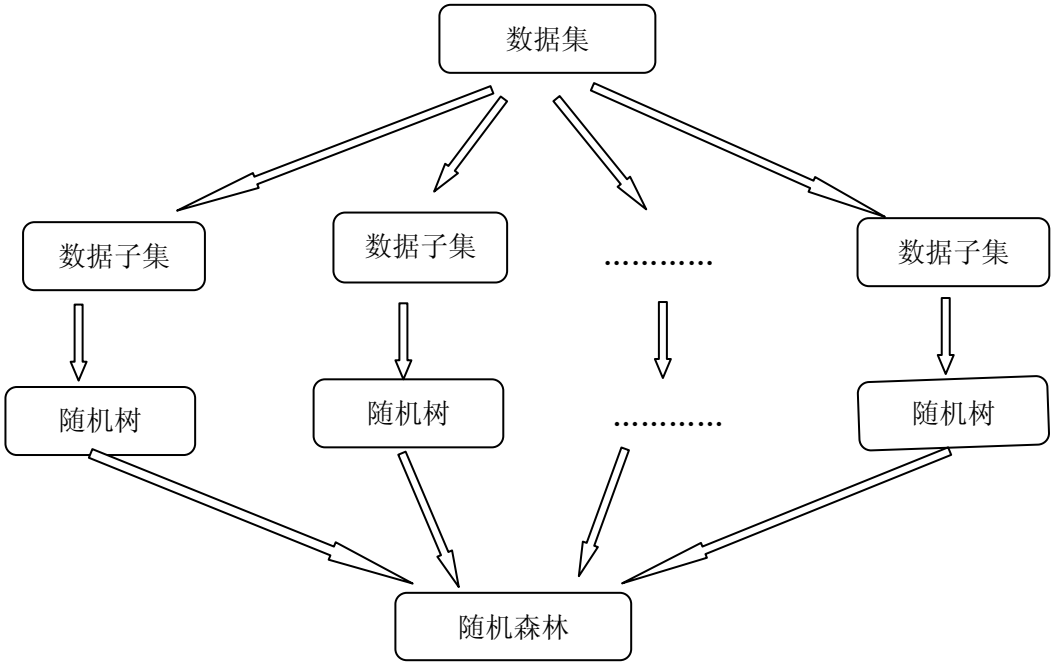

图 6 随机森林构造结构图

[22]Breiman L,Random forests[J].Machine Learing,2001(45).

（四）SVM

支持向量机(SVM)是在统计学习理论和结构风险最小原理基础上发展起来的一种新的机器学习方法，是解决非线性分类、函数估算、密度估算等问题的有效手段，主要思想是建立一个最优决策超平面，使得该平面两侧距平面最近的两类样本之间的距离最大化，从而对分类问题提供良好的泛化能力根据有限的样本信息在模型中特定训练样本的学习精度和无错误地识别任意样本的能力之间寻求最佳最精确的结果，保证了模型具有全局最优、最大泛化能力、推广能力强等优点，在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合中，能够很好地解决许多实际预测问题[23]。

SVM 的线性回归形式如下，设样本为 n 维向量，某区域的 k 个样本及其值表示为：$(x_1,y_1),.....(x_k,y_k)\in R^n$，线性函数设为：

$$f(x)=w*x+b \qquad (1)$$

并假设所有训练数据都可以在精度ε下无误差地用线性函数拟合，即：

$$\begin{cases} y_1-w*x_1-b\le\varepsilon \\ w*x_1+b-y_1\le\varepsilon \end{cases} \quad i=1,2,\cdots\cdots,k \qquad (2)$$

考虑到允许拟合误差的情况，引入松弛因子$\xi_i\ge 0$和$\xi^*{}_i\ge 0$则式（2）变成

$$\begin{cases} y_1-w*x_1-b\le\varepsilon+\xi_i \\ w*x_1+b-y_1\le\varepsilon+\xi^*{}_i \\ \xi_i\ge 0 \\ \xi_i^*\ge 0 \end{cases} \quad i=1,2,\cdots\cdots,k \qquad (3)$$

则回归估计问题就转化为在约束条件（3）下最小化误差：

$$\min R(w,\xi,\xi^*)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{k}(\xi_i+\xi_i^*) \qquad (4)$$

五、模型的求解和检验

在本文的实证部分，首先确定最终的影响武汉商品房价格最显著的几个关键词变量，然后分别用线性回归、回归树、bagging、随机森林和 SVM 等 5 个模型对数据进行拟合和预测，并比较各模型拟合及预测的效果，得到最佳的拟合模型和预测模型。数据分析的全部工作通过 R3.1.1 软件实现，代码见附录 4。

[23]肖轩.灰色神经网络与支持向量机预测模型研究[D].武汉理工大学,2009(05).

关于模型的拟合和预测有几点说明：拟合样本区间是 2011 年 1 月到 2015 年 2 月，共 50 个自然月的数据，用于预测检验的样本区间是 2015 年的 3 月到 5 月三个月的数据。在模型拟合效果的比较上，本文引入 MSE 和 NMSE 分别代表模型的稳定性和拟合度，两者定义如下：

$$MSE = \frac{1}{n}\sum (Y - \hat{Y})^2$$

$$NMSE = \frac{MSE}{\mathrm{var}(Y)} = \frac{n-1}{n}\frac{\sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2}$$

（一）搜索关键词的确定

本文共搜集的 119 个百度指数关键词中，若全部引入模型进行拟合和预测显然工作量巨大，并且会出现严重的多重共线性导致模型效果不佳，因此须筛选出少量具有代表性的关键词，再进行模型的拟合，最终的关键词的确定经过以下几个步骤：

1.简单手动筛选。观察所有关键词的数据，将数据趋势变化不明显或者几乎没有变化趋势的关键词剔除；

2.Pearson 相关系数筛选。分别计算各关键词与因变量的 Pearson 相关系数，将与因变量相关系数小于 0.6 的剔除，本文选择 0.6 作为临界值，是综合考虑筛选后的变量个数和各变量间关系模式不明显等因素的结果。经过该步骤筛选得到的关键词及其与因变量的相关系数如表 1。

3.线性回归筛选。前两步筛选出的全部变量作为自变量对因变量进行简单线性回归，并对模型进行多重共线性和自相关进行诊断，最后利用逐步回归及 AIC 准则选出最终的关键词，。

表 1 相关系数筛选后的关键词及其与房价的相关系数表

武汉房价	武汉房价趋势	武汉房地产公司	金地集团	公积金查询	租房子 58 同城	住房公积金	公积金提取
0.6906	0.654	-0.7362	0.8009	0.7337	0.8708	0.6499	0.8243
楼盘名称	武汉公积金	武汉公积金管理中心	房贷利率	建材团购	按揭贷款利率	房贷基准利率	搜房网武汉
-0.702	0.6489	0.8589	0.6895	0.6885	0.8719	0.6892	0.8651
二手房出售	二手房转让	房贷利率是多少	楼盘网	楼盘开盘	公积金贷款额度	万科金色城市	小户型装修
0.6999	0.723	0.8199	0.7956	-0.7095	0.8096	0.7013	0.8028
户型图大全	户型	个人公积金余额查询	租房合同	链家租房	九正建材网	武汉亿房网	建材加盟
0.7736	0.8494	0.7568	0.6633	0.9667	0.6351	-0.8311	0.6412

经过简单手动筛选，共剔除 21 个关键词，相关系数的筛选，共剔除 50 个关键词，而最后一步的线性回归筛选，则剔除了 39 个关键词，因此最后在线性回归中存在 8 个作为最终自变量的关键词，分别是：公积金贷款额度，武汉公积金管理中心，公积金提取，按揭贷款利率，金地集团，武汉亿房网，租房子 58 同城，建材团购。

从得到的最终的这些关键词可以看出，武汉市民决定是否买房这一重要经济行为时，密切关注的并非是房源和房价走势等基础性问题，而是更加关注公积金和房贷利率等宏观政策性问题，买房时会综合考虑工作单位的公积金福利和当地各银行提供的房贷利率的优惠。同时也会密切关注公积金的管理制度的实施情况，从网上搜集关于买房的相关信息（如金地集团和武汉亿房网），关注租房和建材市场的动向。

（二）模型估计

正如上文在筛选关键词时提到的，在筛选过程中，我们构建了一个线性回归方程，估计的结果表 2。

表 2 线性回归的拟合结果

变量	系数	标准误	p 值
截距项	5699.761	700.128	0.000(***)
公积金贷款额度	-0.046	0.026	0.091(*)
武汉公积金管理中心	0.073	0.026	0.007(***)
公积金提取	-0.038	0.012	0.004(***)
按揭贷款利率	0.166	0.049	0.002(***)
金地集团	1.514	0.191	0.000(***)
武汉亿房网	-0.863	0.153	0.000(***)
租房子 58 同城	2.142	0.368	0.000(***)
建材团购	0.188	0.039	0.000(***)
$\overline{R}^2$ = 0.9828 *F*统计量 = 352(***)			

注：“***”表示在 0.01 的置信水平下显著，“**”表示在 0.05 的置信水平下显著，“*”表示在 0.1 的置信水平下显著。

从回归结果看，总体拟合效果很好，变量系数均通过了显著性检验，因此该线性模型应该具有较好的预测效果。但进一步考察模型假设，我们可通过图形简单刻画因变量的分布情况，并画出它的直方图与其对应的正态分布图进行比较，结果见图 7。

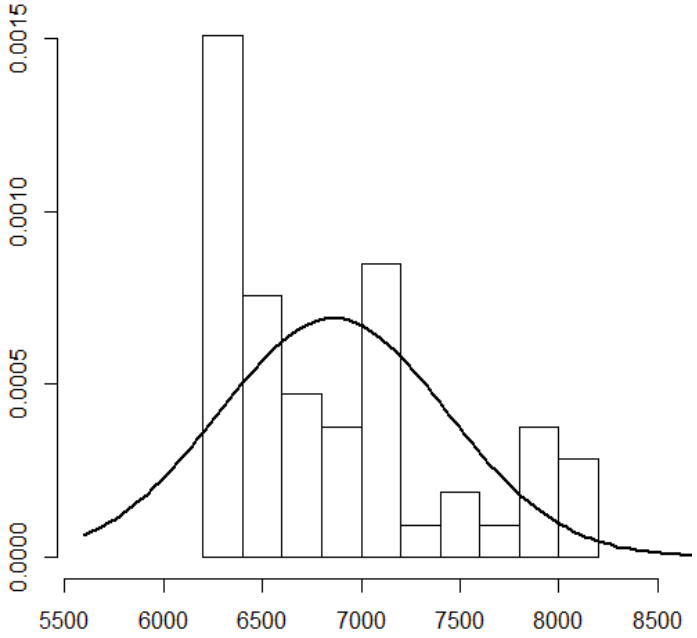

图 7 房价直方图及与相应正态分布的比较图

对武汉商品房价格的数据进行正态性检验，JB 统计量的值为 9.4838，对应的 p 值为 0.0087，即拒绝原假设，认为该组数据不服从正太分布。从图 7 可看出因变量房价的分布明显不服从正态分布，因此在用线性回归进行模型模拟时并不能满足其对因变量分布的基本假定，故在下文中本文相继用回归树、bagging、随机森林及 SVM 等 4 个模型对数据进行了模拟，下面将逐一介绍拟合过程。

回归树模型的拟合通过如下步骤完成：（1）用 R 中的 rpart 包中的 rpart 函数因变量与各自变量的树回归；（2）对得到的树回归模型进行交叉验证，看是否存在过度拟合。（3）若存在过度拟合现象，则使用剪枝函数(prune)修正模型；（4）再次用交叉验证技术对修正模型进行过拟合验证，若仍存在过拟合现象，则重复进行步骤（3）和（4），直到过拟合现象消除，并画出最终的回归树结果。最终回归树结果如图 8（X19 和 X6 表示关键词金地集团和公积金查询）：

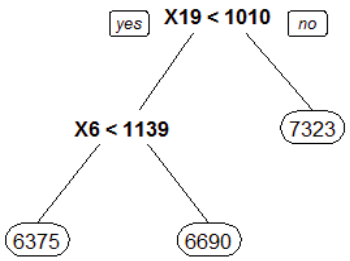

图 8 回归树结果图

上文建立的回归树明显过于简化，在数据拟合和模型的预测上也存在过于粗略的问题，因此进一步使用 bagging 提高模型的拟合精度，bagging 模型的基本步骤：（1）不断放回的对训练样本进行再抽样，抽取 n 个自助样本；（2）对每个样本建立一个回归树，并平均每个样本对应的预测值作为总体的预测值；（3）进行交叉验证，看是否存在过拟合现象，若存在，则通过控制对应参数进行修正。bagging 模型的拟合和预测的效果将在下文各模型的比较中呈现。

bagging 更进一步就是随机森林模型，它同时对训练样本和变量进行抽样。随机森林模型的拟合还拥有回归树和 bagging 不具有的优势就是它不用进行交叉验证，因为随机森林不存在过拟合的现象，同时还可以输出自变量的重要性度量，运行结果如表 2。

表 3 随机森林模型下各变量的重要性

变量	均方误差递减下的重要性	精确度递减下的重要性
公积金贷款额度	22279.16	898044.50
武汉公积金管理中心	60205.74	2021314.30
公积金提取	35084.10	1502289.00
按揭贷款利率	32629.56	1501574.60
金地集团	42184.00	1982673.30
武汉亿房网	25492.16	1312322.00
租房子 58 同城	45893.46	1781718.60
建材团购	33614.62	1235673.00

通过表 3 中的结果可以看出，在均方误差递减意义下的重要性排名中，排在前三位的关键词分别是武汉公积金管理中心、租房子 58 同城和金地集团，在精确度递减意义下的重要性排名中，排在前三位的关键词分别是武汉公积金管理中心、金地集团和租房子 58 同城。

SVM 模型的估计也直接通过 e1071 包中的 svm 函数实现，经过不断调试，得到的最优模型的估计参数分别为 $\cos t = 1$，$tolerance = 0.001$，$epsilon = 0.1$，$rho = -0.445$，其中，cost 表示代价值，tolerance 表示迭代的终止允许项，epsilon 表示迭代中二次规划的终止条件，rho 表示判断函数的常数项。模型的拟合和预测的效果将在下文各模型的比较中呈现。

（三）预测结果分析

在完成对 5 种模型的估计后，分别计算出各模型的 MSE 和 NMSE，对各模型的拟合度和稳定性进行评估，所得结果见表 4。

表 4 武汉商品房价格各种模型的 MSE、NMSE 和排名

模型 指标	线性回归	回归树	Bagging	随机森林	SVM
MSE	3620.652	63683.680	21096.150	3205.383	4279.567
NMSE	0.01406	0.24736	0.08194	0.01245	0.01662
排名	2	5	4	1	3

从模型的拟合度和稳定性来看，回归树和 Bagging 模型都存在明显的不足，在 MSE 和 NMSE 值上都远远大于另外三个模型，尤其回归树模型在所有模型中的效果最差。线性回归、随机森林和 SVM 这三种模型的效果较好，其中随机森林模型的拟合度和稳定性最佳，但与线性回归和 SVM 差别不大，可以预见的是这三个模型都具有较为精确的预测能力。为了更直观的比较，我们可通过图形展示三个模型的拟合效果，具体情况可见图 9。

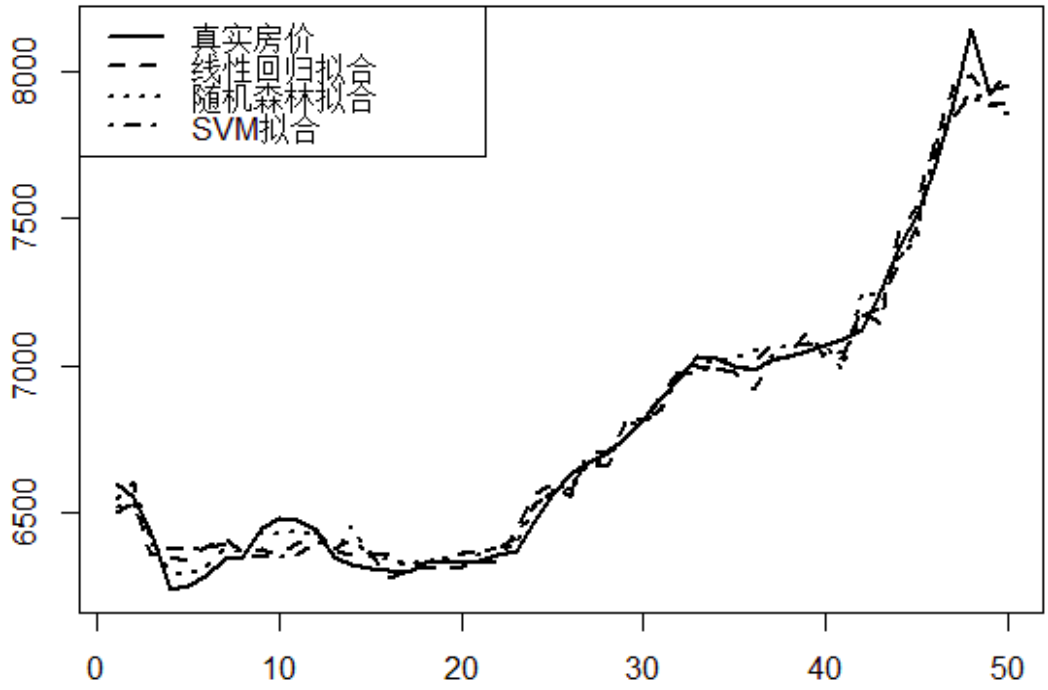

图 9 武汉商品房价格实际值与各模型拟合值的比较

从图 9 中可以看出，线性回归、随机森林和 SVM 这三个模型在样本期内的预测值（即拟合值）的趋势与武汉商品房房价的实际值的变化趋势高度一致，同时从各模型拟合值曲线与真实值曲线的贴近程度看，随机森林模型拟合值曲线比线性回归模型拟合值曲线更贴近真实值，而线性回归拟合值曲线则比 SVM 模型拟合值曲线更贴近真实值。由此也可以看出，在拟合效果较好的三个模型中，随机森林的拟合值总体上更贴近真实值。

最后分别用各个模型对样本期外的三个月的武汉商品房房价进行预测，并与真实值进行比较，计算平均误差率，所得结果见表 5。

表 5 武汉商品房房价各模型预测值与真实值比较

时间	实际房价	线性回归预测	回归树预测	bagging 预测	随机森林预测	SVM 预测
201503	7978.8	7981.61	7322.60	7574.10	7788.57	8066.89
201505	8089.84	8164.44	7322.60	7565.68	7769.25	8109.40
201505	8159.08	7962.91	7322.60	7574.32	7845.12	8062.48
平均误差率	——	1.12%	9.32%	6.24%	3.40%	0.84%

如表 5 所示，三行取值分别为武汉商品房房价 2015 年 3 月到 2015 年 5 月的真实值和各模型的房价预测值。模型拟合效果较差的回归树和 bagging 模型的预测效果仍然不理想，与真实房价相比存在较大偏差，令人意外的是拟合效果最佳的随机森林模型预测效果并不理想，通过观察图 9 最后一段图形可以发现，在实际房价突然上升的一个月中，随机森林模型并没有准确捕捉到这一突然变化，因此出现后几期拟合值与实际值相差较大的现象，并导致最终预测值的效果不理想。而线性回归模型和 SVM 模型很好的捕捉到了样本期最后的这一突然变化，因此预测值具有较好的效果，平均误差率仅有 1.12%和 0.84%。总体上来看，SVM 模型拥有最好的预测效果。

六、结论和建议

本文在分析商品房价格和百度搜索指数关系的基础上，对武汉市商品房价格和与其相关的关键词进行分析，对选取与武汉市商品房价格 119 了个关键词利用进行筛选，根据逐步回归及 AIC 准则选出最终的 8 个关键词，分别是公积金贷款额度，武汉公积金管理中心，公积金提取，按揭贷款利率，金地集团，武汉亿房网，租房子 58 同城，建材团购。

对筛选出来的 8 个关键词与武汉市商品房价格运用线性回归、回归树、bagging、随机森林和 SVM 模型进行建模，得到对武汉市商品房价格拟合效果最好而且模型比较稳定的模型是随机森林模型。将各个模型对样本期外的三个月的武汉商品房房价进行预测，并与真实值进行比较，计算平均误差率。结果是线性回归模型和 SVM 模型平均误差率为 1.12%和 0.84%，比其他的模型预测效果更佳，而且可以提前 10—15 天对武汉市商品房价格进行预测。

由以上结论提出的相关建议如下：

1.武汉市政府对商品房价格趋势的了解可以关注公积金贷款额度，武汉公积金管理中心，公积金提取，按揭贷款利率，金地集团，武汉亿房网等 8 个关键词的搜索量。

2.如果相关部门对武汉市商品房价格的波动趋势仅仅是通过关键词的搜索

量进行商品房价格的拟合，可以选用拟合度较高而且比较稳定的随机森林模型，如果要对武汉市商品房价格进行预测，应该采用预测精度较高的 SVM 模型。

3.相关部门可以在月底将得到的关键词的百度指数运用 SVM 模型对商品房价格进行及时预测。

七、模型的评价与改进

模型的评价：论文中选用更适用于我国经济问题的百度指数进行研究，使得研究更具有代表性。基于基础的关键词运用关键词挖掘进行关键词的选取，使得对关键词的选取更加全面。选用五中模型分别对武汉市商品房价格进行模拟和预测，得到的结果更具有说服力。模型建立和求解中用到的机器学习算法能够克服传统方法的局限性，得到的结果精度较高，时效性较强。目前利用百度指数的相关研究较少，关键词是根据实际经验选取的，带有一定的主观性，而且百度关键词挖掘推荐的关键词的有效性不能很好的甄别。在实际生活中一大部分人会利用百度搜索信息来了解关于房地产的信息，但还有一部分人没有利用百度搜索相关信息，而是通过实体广告、朋友亲戚介绍来购房的，所以百度搜索数据不能完全代表购房者和房地产投资者的整体行为信息，如果要利用本文的结论对消费者整体的消费行为进行研究，还需要进一步的验证。

模型的改进：将百度搜索的信息的有效性进行甄别，并将关键词的选取建立一套行之有效的理论体系。由于数据搜集的局限性，本文的变量全部来自与房价相关的关键词，如果加入其他与房价相关的经济变量模型的预测精度会更高一些。

参考文献

[1] Breiman L. Bagging Predictors[J]. Machine Learing, 1996(24).

[2] Breiman L, Random forests[J]. Machine Learing, 2001(45).

[3] Harry S, Gnirgnis, Christos I, Giannikos & Randy I, Anderson. he US Housing Market: Asset Pricing Forecasts Using Time Varying Coefficients [J]. The Journal of Real Estate Finance and Economics, 2005(01).

[4] Hasa Selim. Determinants of House Prices in Turkey: Hedonic Regression Versus Artificial Neural Network [J]. Dogus University Journal, 2008(01).

[5] Jane P, Browna, Haiyan Song & Alan Me Gillivray. Forecasting UK House Prices: A Time Varying Coefficient Approach[J]. Econoinic Modelling, 1997 (04).

[6] Jeremy Ginsberg, Matthew H, Mohebbi, Rajan S, Patel, Lynnette Bramme, Mark S, Smolinski, Larry Brilliant. Detecting Influenza Epidemics Using Search Engine Query Data[J]. Nature, 2009,457(2).

[7] Quinlan J R.Induction of decision trees.Machine Learning,1986,(01).

[8] Wu L, Brynjofsson E,The Future of Prediction: How Google searches Foreshadow Housing Prices and Sales [C] .Working Paper, 2014.

[9] 程亚鹏，张虎，张庆宏. GM(1，1)模型在房地产价格指数预测中的应用[J]. 河北农业大学学报，1997(07).

[10] 董倩，孙娜娜，李伟.基于网络搜索数据的房地产价格预测[J].2014(10).

[11] 罗婧，朱建锋．基于数据挖掘与空间计量的房地产价格经验分析[J]．开发研究，2013(05).

[12] 刘涛熊，徐晓飞．大数据与宏观经济分析研究综述[J]．国外理论动态，2015（01）.

[13] 莫春梅，倪宗瓒，高凤琼．回归树的建模与应用[J]．中国预防医学杂志，2002(09).

[14] 倪大鹏．基于数据挖掘的房地产定价方法研究[D]．大连理工大学，2013.

[15] 钱峰，吕效国，朱帆．灰色 GM(1，1)模型的改进模型在房地产价格指数预测中的应用[J]．数学的实践与认识，2009(04).

[16] 王聪．基于多因素 LOGISTIC 的城市房地产价格预测模型研究[D]．大连理工大学，2008(10).

[17] 王婧，田澎．小波神经网络在房地产价格指数预测中的应用[J]．计算机仿真 2005(07).

[18] 武秀丽，张锋．时间序列分析法在房价预测中的应用[J]．科学技术与工程．2007(11).

[19] 吴喜之．统计学：从概念到数据分析［M］．北京：高等教育出版社，2008．杨刚．基于数据挖掘的房地产价格分析预测研究[D]．南昌大学，2014(05).

[20] 肖轩．灰色神经网络与支持向量机预测模型研究[D]．武汉理工大学，2009(05).

[21] 杨刚．基于数据挖掘的房地产价格分析预测研究[D]．南昌大学，2014(05).

[22] 雍凯．随机森林的特征选择和模型优化算法研究[D]．哈尔滨工业大学，2008(12).

[23] 杨树新，董纪昌，李秀婷.基于网络关键词搜索的房地产价格影响因素研究[J].2013(03).

[24] 张崇，吕本富，彭赓，刘颖. 网络搜索数据与 CPI 的相关性研究[J]. 管理科学学报． 2012．15(7): 50-59.

[25] 张翔，周明全，耿国华，侯凡. Bagging 算法在中文文本分类中的应用[J]. 计算机工程与应用，2009(05).

附录

附录 1：原始关键词

武汉房价	武汉房价、武汉房价网、武汉房价走势、房价网、武汉房价走势 2015、 武 汉房价地图、黄冈房价网、2012 武汉房价
武汉房地产	武汉房地产、武汉房地产信息网、武汉房地产市场信息网、武汉市房地产管理局、武汉房地产公司排名、武汉房地产网、武汉房地产交易中心、（武汉地区）金地集团武汉房地产开发有限公司、武汉房地产公司、武汉万科魅力之城、黄石房地产
武汉搜房网	武汉搜房网、武汉搜房网二手房、搜房网武汉、武汉二手房、武汉搜房网新房、搜房网武汉装修、武汉装修网、武汉装修公司、武汉搜房网租房、武汉租房、武汉租房网、武汉得意生活网、武汉亿房网
武汉楼盘	武汉楼盘、武汉新楼盘、楼盘网、楼盘开盘、房产信息网新楼盘、楼盘风水、超低价楼盘、武汉楼盘地图、楼盘广告、武汉楼盘信息、武汉最新楼盘、一手楼盘、楼盘名称、湖北蕲春新楼盘、楼盘出售、万科金色城市、武汉楼盘信息网、武汉别墅楼盘、武汉楼盘开盘、武汉楼市、武汉万科楼盘、武汉物业
户型	装修、小户型装修、户型图、小户型装修实景图、小户型装修实例、小户型大空间、小户型、跃层户型、交换空间小户型设计、小户型装修设计、户型图大全、小户型装修效果图、50 平米小户型装修效果图、40 平米小户型装修效果图、户型、30 平米小户型装修效果图、小户型装修案例、别墅户型图、60 平米小户型装修效果图、小户型室内装修设计、一居室小户型装修图、小户型设计、三室一厅户型图、小户型空间创意设计、户型风水、小户型家装、户型设计、小户型装修图片、loft 户型、户型分析、联排别墅户型图、户型图怎么看、80 平米小户型装修效果图、户型平面图、户型介绍
租房	赶集网租房、58 同城租房网、租房、租房网、租房子 58 同城、58 同城租房、租房合同、租房合同范本、出租房、安居客、58 同城出租房屋个人、好租、丁丁租房、租房协议、链家租房
建材	中国建材网、建材市场、建材、建材网、中国建材在线、新型建材、建材在线、九正建材网、北新建材、建材团购、建筑建材、中国建材、建材加盟、建材加盟、建材家居网、建材团购、防水建材
二手房	二手房、58 同城二手房、二手房出售、二手房出售、赶集网二手房、二手房转让、二手房交易流程、买二手房注意事项、二手房网、二手房过户费怎么算、二手房信息网、二手房买卖、武汉二手房、二手房买卖合同、个人二手房、求购二手房、
房贷利率	房贷利率、最新房贷利率、首套房贷利率、房贷利率是多少、建行房贷利率、房贷利率表、房贷利率计算、房贷利率打折、房贷利率怎么算、房贷利率下调、房贷利率调整、房贷利率优惠、银行贷款利率、按揭贷款利率、 房贷基准利率、房贷利率一般是多少、房贷最新利率、最新房贷利率

公积金	公积金查询、住房公积金查询、个人公积金余额查询、公积金、住房公积金、住房公积金提取条件、公积金查询个人账户、公积金提取、公积金贷款、公积金贷款利率、公积金贷款额度、住房公积金提取、公积金管理中心、武汉公积金、武汉公积金管理中心、公积金贷款流程

附录 2：Python 数据抓取源程序

```
file_in = '/Users/John/Desktop/in.txt'
file_out= '/Users/John /Desktop/out.txt'
up = 130.0
height = 207.666666
low = up + height
def get_value(maxy, miny, site):
    return maxy - (site - up) / height * (maxy - miny)
if __name__ == '__main__':
    reader = open(file_in)
    if not reader:
        print 'no file!'
    line = reader.readline()
    miny = (float)(line.strip())
    line = reader.readline()
    maxy = (float)(line.strip())
    data = reader.readline()
    attrs = data.split(',')
    values = []
    for attr in attrs:
        if attr[0] == 'M':
            continue
            #f = (float)(attr[1:])
            #values.append(f)
        elif 'L'in attr:
            tw = attr.split('L')
            f = (float)(tw[0])
            values.append(get_value(maxy, miny, f))
        else:
            f = (float)(attr)
            values.append(get_value(maxy, miny, f))
    writer = open(file_out, 'w')
    length = len(values)
    for i in range(0, length):
        v_s = '%f' %values[i]
        if i != 0:
            writer.write('    ')
        writer.write(v_s)
    writer.write('\r\n')
```

```
writer.close()
print 'process down, data keep in file: ', file_out
```

附录 3：关键词和商品房价格的相关系数

中国建材网	建材市场	建材	建材网	中国建材在线	新型建材	建材在线	九正建材网	建材团购	建筑建材
-0.5678	0.3709	-0.189	-0.3421	-0.3762	0.0244	-0.0809	0.6351	0.6885	0.06
中国建材	建材加盟	二手房	58 同城二手房	二手房出售	二手房转让	二手房交易	二手房网	二手房信息网	武汉二手房
-0.3434	0.6412	0.4923	0.5697	0.6999	0.723	-0.0085	-0.0957	0.5882	-0.0187
二手房买卖合同	个人二手房	求购二手房	买二手房要交哪些税	武汉物业	房贷利率	最新房贷利率	首套房贷利率	房贷利率是多少	建行房贷利率
-0.3183	0.6003	0.4703	0.4703	-0.5762	0.6895	-0.1304	0.3337	0.8199	0.5974
房贷利率表	房贷利率计算	房贷利率调整	房贷利率优惠	按揭贷款利率	房贷基准利率	最新房贷利率	公积金查询	住房公积金查询	个人公积金余额查询
0.5003	-0.0509	0.3589	0.193	0.8719	0.6892	-0.1304	0.7337	0.1193	0.7568
公积金	住房公积金	住房公积金提取条件	公积金提取	公积金贷款	公积金贷款利率	公积金贷款额度	住房公积金提取	公积金管理中心	武汉公积金
0.5913	0.6499	0.4176	0.8243	0.3267	0.1989	0.8096	0.1856	-0.1745	0.6489
武汉公积金管理中心	公积金贷款流程	武汉房价	武汉房价网	武汉房价趋势	武汉房地产	武汉房地产信息网	武汉房地产市场信息网	武汉房地产公司	武汉房地产网
0.8589	-0.2614	0.6906	-0.1482	0.654	-0.606	0.0494	0.1453	-0.7362	-0.1396
金地集团	武汉万科魅力	黄石房地产	武汉搜房网	武汉搜房网二	搜房网武汉	武汉二手房	武汉装修网	武汉装修公司	武汉租房

	之城			手房					
0.8009	-0.6037	-0.3239	0.0496	-0.4664	0.8651	-0.0187	0.5635	0.1681	-0.6074
武汉租房网	武汉得意生活网	武汉亿房网	武汉楼盘	武汉新楼盘	楼盘网	楼盘开盘	楼盘风水	楼盘广告	楼盘名称
-0.5384	0.5427	-0.8311	0.5538	-0.0857	0.7956	-0.7095	-0.5117	-0.5957	-0.7023
楼盘出售	万科金色城市	武汉楼市	武汉万科楼盘	武汉装修	小户型装修	户型图大全	户型	武汉赶集网租房	58 同城租房网
-0.1801	0.7013	0.1198	-0.3476	-0.2093	0.8028	0.7736	0.8494	0.1361	0.5524
租房	租房网	租房子 58 同城	租房合同	租房合同范本	安居客	好租	链家租房		
0.4767	0.4046	0.8708	0.6633	0.4831	0.5959	0.0538	0.9667		

附录 4：武汉市商品房价格预测相关代码

```
##############数据的读入与处理###############

ssdata<- read.csv("I:/搜索数据.csv",header=F)
houp<- read.csv("I:/商品房交易数据.csv",header=F)
ssdata<- as.matrix(ssdata)
mdata = matrix(0,nr=58,nc=98)
for(i in 0:57){
   j0 <- i+1
   j1 <- 4*i+1
   j2 <- 4*i+2
   j3 <- 4*i+3
   j4 <- 4*i+4
   mdata[j0,]<- ssdata[j1,] + ssdata[j2,] + ssdata[j3,] +ssdata[j4,]
}
mdata<- mdata[1:53,]
mdata<- data.frame(mdata)
setwd("D:\\")
write.table(mdata,"sample.csv",sep=",")

zdata<- cbind(houp,mdata) #包括因变量在内的所有数据
class(zdata)
names(zdata)
```

```
##############数据清洗###############

COO<- cor(zdata)
xx<- COO[1,]
xn<- which(xx> 0.55)
####通过 excel 将无关变量删除，再读入数据####
modata<- read.csv("I:/建模的数据.csv",header=F)
modata<- as.matrix(modata)
nrow(modata)
ncol(modata)
tmodata<- matrix(0,nr=58, nc=47)
for(i in 0:57){
   j0 <- i+1
   j1 <- 4*i+1
   j2 <- 4*i+2
   j3 <- 4*i+3
   j4 <- 4*i+4
   tmodata[j0,]<- modata[j1,] + modata[j2,] + modata[j3,] +modata[j4,]
}

tmodata<- tmodata[1:53,]
setwd("D:\\")
write.csv(tmodata,"sample.csv",sep=",")
##构造建模数据和预测数据
sydata<- cbind(houp,tmodata)
scol<- c(4,14,16,19,21,24,26,35,38,43,44,45)
jmdata<- sydata[1:50,-scol]

names(sydata)<- c("Y","X1","X2","X3","X4","X5","X6","X7","X8","X9","X10","X11",

"X12","X13","X14","X15","X16","X17","X18","X19","X20","X21","X22",

"X23","X24","X25","X26","X27","X28","X29","X30","X31","X32","X33",

"X34","X35","X36","X37","X38","X39","X40","X41","X42","X43","X44",
                              "X45","X46","X47")

predata<- sydata[51:53,-scol]
names(jmdata)<-    c("Y","X1","X2","X4","X5","X6","X7","X8","X9","X10","X11",
                          "X12","X14","X16","X17","X19","X21","X22",
                          "X24","X26","X27","X28","X29","X30","X31","X32","X33",
                          "X35","X36","X38","X39","X40","X41","X45","X46","X47")
names(predata)<-    c("Y","X1","X2","X4","X5","X6","X7","X8","X9","X10","X11",
```

```
                    "X12","X14","X16","X17","X19","X21","X22",
                    "X24","X26","X27","X28","X29","X30","X31","X32","X33",
                    "X35","X36","X38","X39","X40","X41","X45","X46","X47")
te1.fit<-
lm(Y~X1+X5+X6+X9+X12+X14+X16+X19+X21+X24+X26+X29+X30+X31+X32+X33
          +X35+X38+X40+X41+X46+X47,data=jmdata)
ste.fit<- step(te1.fit,data=jmdata)
summary(ste.fit)
te2.fit<- lm(Y ~ X5 + X6 + X12 + X16 + X19 + X24 + X35 + X46,data=jmdata)
summary(te2.fit)
Var.Y<- var(jmdata$Y)
mpre.lm<- te2.fit$fitted
tnmse1<- mean(resid(te2.fit)^2)/Var.Y
mse1<- mean(resid(te2.fit)^2)
col<- c(5,6,12,14,16,19,28,35)
tpredata<- predata[,col]

pre.lm<- predict(te2.fit,newdata=tpredata)     #预测集

----------------------------------------------------------------

library(DMwR)
library(rpart)
library(ipred)
library(randomForest)
####回归树算法####
tmodel.tree=rpart(Y ~ X5 + X6 + X12 + X16 + X19 + X24 + X35 + X46, data=jmdata)
rpart.plot(tmodel.tree)
mpre.tree <- predict(tmodel.tree, data=jmdata)
tnmse2 <- mean((mpre.tree- jmdata$Y)^2)/Var.Y
mse2<- mean((mpre.tree- jmdata$Y)^2)

pre.tree<- predict(tmodel.tree,newdata=tpredata)     #预测集
##bagging(袋袋算法)##
tmodel.bagging <- bagging(Y ~ X5 + X6 + X12 + X16 + X19 + X24 + X35 + X46,
                          data=jmdata, nbagg=1000)
mpre.bagging=predict(tmodel.bagging,jmdata)
tnmse3 <- mean((mpre.bagging- jmdata$Y)^2)/ Var.Y
mse3<- mean((mpre.bagging- jmdata$Y)^2)

pre.bagging<- predict(tmodel.bagging,newdata=tpredata)     #预测集
##随机森林算法##
tmodel.forest <-randomForest(Y ~ X5 + X6 + X12 + X16 + X19 + X24 + X35 + X46,
                             importance=T,data=jmdata)
```

```
mpre.forest=predict(tmodel.forest, jmdata)
tnmse4 <- mean((mpre.forest- jmdata$Y)^2)/ Var.Y
mse4<- mean((mpre.forest- jmdata$Y)^2)

pre.forest<- predict(tmodel.forest,newdata=tpredata)     #预测集
##SVM##
library(nnet)
library(e1071)

norm.data <- scale(jmdata)
tmodel.svm <- svm(Y ~ X5 + X6 + X12 + X16 + X19 + X24 + X35 + X46, norm.data)
mpre.svm<- predict(tmodel.svm, norm.data)
mpre.svm<- mpre.svm*sd(jmdata$Y) + mean(jmdata$Y)
Var.y<- var(scale(jmdata$Y))
tnmse5<- mean(resid(tmodel.svm)^2)/Var.y
mse5<- mean(resid(tmodel.svm)^2)*Var.Y

pre.svm<- predict(tmodel.svm,newdata=scale(tpredata))       #预测集

mix.pre<- cbind(jmdata$Y,mpre.lm,mpre.tree,mpre.bagging,mpre.forest,mpre.svm)
mix.pre<- as.ts(mix.pre)
setwd("D:\\")
write.csv(mmix.pre,"预测值.csv",sep=",")

mmix.pre<- cbind(pre.lm, pre.tree, pre.bagging, pre.forest, pre.svm)
setwd("D:\\")
write.csv(mmix.pre,".csv",sep=",")

print(c(tnmse1,tnmse2,tnmse3,tnmse4,tnmse5))



--------------------------------------------------------------------------------
###############文中图形的绘制代码###############

mix.pre<- read.csv("I:/模型实际及拟合值.csv",header=T)
mix.pre<- as.ts(mix.pre)
mix.pre<- data.frame(mix.pre)

houp<- as.ts(houp)
plot(houp, xlab="", ylab="",ylim=c(3000,8500), lwd=2, lty=1)
lines(tmodata[,13],lwd=2,lty=3)
legend("topleft",c("真实房价","房贷利率"), lwd=2, lty=c(1,3))

attach(mix.pre)
```

```
plot(as.ts(真实房价),xlab="",ylab="",lwd=2,lty=1)
lines(线性回归拟合,lwd=2,lty=2)
lines(随机森林拟合值,lwd=2,lty=3)
lines(SVM 拟合值,lwd=2,lty=4)
legend("topleft",c("真实房价","线性回归拟合","随机森林拟合","SVM 拟合"),
       lty=1:4,lwd=2)
plot(houp, xlab="", ylab="",ylim=c(2500,8500), lwd=2, lty=1)
lines(tmodata[,5],lwd=2,lty=4)
legend("topleft",c("真实房价","公积金贷款额度"), lwd=2, lty=c(1,4))
houp<- houp[,1]
dens <- density(houp)
xlim<- range(dens$x) ;ylim<- range(dens$y)
m<- mean(houp);s<- sd(houp)
hist(houp,breaks=10,xlim=xlim,xlab="",ylab="",main="",prob=T)
curve( dnorm(x, m, s), lwd=2, add=T)
```