基于网络搜索数据的房地产价格预测

董 倩　孙娜娜　李 伟

内容提要:本文以北京、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,以来自我国最大搜索引擎的百度搜索指数为数据基础,使用 6 种计量模型分别对 16 个城市的二手房价格和新房价格进行了拟合和预测,得到预测二手房和新房价格变动情况的最优模型。结果显示:网络搜索数据不但能够较好地预测房价指数,而且能够分析经济主体行为的趋势与规律,有一定的时效性。预测的月度房地产价格能够比官方数据发布提前约两周时间。

关键词:网络搜索数据;房地产价格预测;交叉验证;支持向量机;随机森林

中图分类号:C812　　**文献标识码**:A　　**文章编号**:1002 - 4565(2014)10 - 0081 - 08

Real Estate Price Prediction based on Web Search Data

Dong Qian　Sun Nana　Li Wei

Abstract: This article provides an optimal model predicting the price trends in new and secondary housing market in 16 cities including Beijing, Shanghai, Tianjin, and some other relatively developed cities in China. Based on the Baidu Search Index (BSI), we fitted and forecasted the housing prices in both markets by using 6 analytical models. The results show that the web search data can not only predict the housing prices, but it can also derive some specific patterns and trends of economic agent behaviors. Besides, this prediction model is timely since it can predict the price trends of the real estate industry two weeks before official statistic agencies publish the data.

Key words: Web Search Data; Real Estate Price Prediction; Cross Validation; SVM; Random Forest

一、引言

随着经济社会的发展,房地产业成为国民经济发展的重要支柱,房屋在人民生活中的地位举足轻重,人们对房地产价格的关注度也越来越高。政府统计部门十分重视房价统计工作,为了适应不断变化的形势,对房价统计制度方法进行了一系列改革。改革后的房地产价格指数更加合理,但是由于该指数系列发布时间是每月中旬,其时效性仍无法满足大众的需求,而且在房价波动较大的时期,发布的房价指数与人们的直观感受也存在一定的差距。

在网络化、信息化高速发达的今天,搜索引擎为购房者在信息搜索方面起到了重要作用,提供了极大的便利,节省了他们收集信息的成本。从某种意义上说,搜索引擎实际上充当了潜在购房者、房地产商以及提供各类房产信息的网站之间的桥梁。因此,从对网络搜索数据的分析中可以发现房地产市场参与各方的心理预期、行为模式,而这些因素将直接影响到房地产价格。大数据时代,对于网络搜索数据的研究分析能够得出经济主体行为的趋势与规律,而且相比经济主体的实际行为具有一定的时效性。

事实上,利用网络搜索数据进行预测在商业界和学术界均有诸多探索和研究。最先利用 Google 搜索数据成功的预测美国流感疾病趋势是 Ginsberg et al. (2009),令健康行为指标等于互联网查询有关的流感的发病率,估计了美国每周的流感活动。这一预测方法被迅速的引入到失业率预测中来。Askitas 和 Zimmermann(2009)研究发现,基于某些关键词的预测比官方数据能够更早的显示失业趋势的变化。

目前,应用网络搜索数据预测价格指数的研究还较少,而预测房地产价格的研究更少。张崇等(2012)基于谷歌搜索数据的 CPI 预测具有一定的

转折点预测能力，平均提前周期为 4 个月。Rajendra Kulkarni 等人（2009）使用城市层面的谷歌搜索指数预测了 20 个城市经季节调整的 case-shiller 指数的变化情况。Lynn Wu 和 Erik Brynjolfsson（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。在美国，对于国家级的预测，与使用传统数据进行基准模型预测比较，利用搜索数据进行样本外预测的平均绝对误差更小。杨树新等（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。

可以看出，目前网络搜索数据应用最多的就是谷歌搜索数据。但是在我国，由于网络限制、习惯等的影响，应用最多的搜索引擎就是百度，因而应用百度搜索指数研究我国房价走势更符合实际情况。但是，目前应用百度搜索数据进行预测研究的文章非常少，尚处于起步阶段；以我国主要大中城市为研究对象，通过百度搜索数据预测房地产价格指数的研究还没有，而本文恰恰能够填补这方面研究的空白，同时也是对大数据应用于政府统计的有益探索，具有较强的理论和实际意义。

本文拟利用网络搜索数据，找出影响我国房地产价格变动的因素，并通过模型，对房地产价格走势情况进行预测。本文的研究思路是：为了解决房地产价格的时效性问题，本文尝试利用百度搜索数据预测我国主要大中城市的新建住宅销售价格指数和二手住宅销售价格指数。由于网络搜索数据可以实时获取，可以把影响价格变化的即时因素带入预测模型，这样在每月月初就可以得到上月的新房和二手房价格指数，比官方数据发布提前两周左右，弥补了传统统计数据信息发布相对滞后的问题，同时该预测数据也可以作为传统房地产价格统计数据的有益补充和参考。

本文的主要创新之处在于：第一，采用网络搜索数据对房地产价格进行预测，在国内相关研究还很少见到。采用网络数据进行预测不但具有较好的预测效果，而且与传统预测方法和官方调查数据相比具有很强的时效性。第二，对每个城市采用交叉验证技术分别建立线性回归、回归树、Bagging、Boosting、随机森林和支持向量机模型进行预测，然后通过测试集 NMSE 和 MSE 的比较选取预测效果最优的模型。第三，由于百度搜索指数可获取的数据从 2011 年 1 月，截止到 2014 年 5 月仅有 29 个同比数据。为了弥补数据量较少可能带来的偏差，我们采用 3 折交叉验证技术，这样就保证了最终结果的精确性和可靠性。

二、变量描述和模型构建

（一）模型构建的前提假设

价格理论认为商品的价格由供求关系决定，房屋作为一种特殊的商品，具有消费品和投资品的双重属性，影响其供给和需求的因素也具有一定的特殊性。首先，房地产业作为国民经济的支柱产业，房地产价格与宏观经济形势息息相关，经济发展水平、居民收入、城镇化水平、物价、利率等宏观经济因素都会影响房地产价格的整体走势。其次，中国的房地产市场受政策因素影响比较大。主要有房地产制度、房地产价格政策、税收政策、城市发展战略、城市规划、土地利用规划等。最后，房屋作为消费品，与其自身和交易环节相关的区域因素、环境因素、个别因素等也会对各具体的房地产项目价格带来直接影响。在上述各种因素的共同作用下，房地产价格的变动整体上是有一定规律可循的。

作为供给与需求的市场微观主体，房地产开发企业及购房者在产生投资或消费需求后，都需要搜集与宏观经济、政策及与房屋本身相关的大量信息，而搜索引擎已成为最重要的信息入口。房地产企业和购房者的心理预期和行为在房地产市场和互联网上都有所反映，在房地产市场上体现为交易量和价格的变化；而在互联网上则体现为搜索内容、搜索量等指标的变化。因此本文认为，网络搜索数据与房地产价格变动存在一定的相关关系。基于大数据的相关性和实时性特点，利用可得的网络搜索数据，可以建立适当的模型对房地产价格进行短期预测。

（二）变量和数据描述

1. 研究对象。

从 2011 年 1 月起，国家统计局每月公布 70 个大中城市新建住宅销售价格指数，按面积划分的新建商品住宅销售价格指数和二手住宅销售价格指数，包括以 2010 年为基期的定基指数、同比指数和环比指数。由于本文要利用百度搜索数据预测房地产价格，考虑到在一些规模较小或者经济不太发达的城市，人们对房地产信息的收集可能更多地通过广告、朋友介绍或房产中介等渠道，通过网络搜索房产信息相对较少，因此我们选取规模较大、经济比较

发达、房地产交易相对活跃的 16 个城市作为研究对象,具体包括北京、上海、广州、深圳等 4 个一线城市,天津、重庆、南京、武汉、沈阳、西安、成都、杭州、青岛、厦门、长沙和海口等 12 个二线城市。

2. 数据描述。

百度搜索指数是以网民在百度的搜索量为数据基础,以关键词为统计对象,分析并计算出各个关键词在百度网页搜索中搜索频次的加权和。是当前互联网乃至整个数据时代最重要的统计分析平台之一,自发布之日起便成为众多企业营销决策的重要依据。根据搜索来源的不同,百度搜索指数可分为 PC 搜索指数(PC 指个人电脑)和移动搜索指数。该平台还可提供词汇媒体指数,用于评估媒体对某词汇的关注程度。

对于网络搜索数据的获取与处理,主要是基于百度指数这项服务,在百度指数当中输入关键词,就能够获得该关键词自 2011 年以来每日的搜索量。该搜索量为相对数据,即相对于当日百度总搜索量中该关键词的搜索率。这项功能反映了某一个关键词在某段时间里的关注程度。

3. 变量描述。

(1)被解释变量。被解释变量分别是 16 个城市的二手住宅销售价格指数(以下简称二手房价格)和新建商品住宅销售价格指数(以下简称新房价格)。采用 2012 年 1 月至 2014 年 5 月共 29 个月的月度同比数据,来源于国家统计局网站。

(2)解释变量。解释变量是与二手房和新房价格相关的某些关键词的网络搜索指数。网络搜索中不同关键词代表的含义不同,因此需要采用科学的方法对关键词进行筛选。

首先,根据人们在房屋购买决策中考虑的主要方面选定初始关键词。具体包括宏观经济形势和房地产市场整体走势、与房地产市场密切相关的政策、与房屋本身和交易细节直接相关的各类信息等,共选取 15 个初始关键词。

其次,利用百度搜索引擎的关键词自动推荐技术,得到与二手房价格相关的 101 个关键词,与新房价格相关的 59 个关键词。剔除重复和数据量较少的关键词,组成关键词库。

最后,对数据进行移动平均处理,转化成月度数据,分别计算每个关键词与二手房价格和新房价格的相关系数,检验每个关键词与房地产价格指数的相关性,并据此对关键词进行筛选。

经过多次比较和筛选,对于 16 个城市的二手房价格预测,我们最终选取了 12 个关键词,分别是:房价走势、房源、装修、房产网、公积金、房贷利率、房产税、房屋出租、房产中介、二手房、二手房交易流程、二手房交易税费。

对于新房价格预测,最终选取了 8 个关键词,分别是:房价走势、房源、装修、房产网、公积金、房贷利率、新楼盘、保障房。

4. 数据预处理。

为了与因变量(二手住宅销售价格指数和新建商品住宅销售价格指数)保持一致,我们对所有关键词的搜索指数做如下处理:首先将根据日搜索指数计算月度平均搜索指数,然后将月度平均搜索指数转换为同比数据,最终得到 16 个城市所有关键词从 2012 年 1 月到 2014 年 5 月的月度同比数据。

(三)房地产价格预测模型构建

1. 回归模型。

由于每个城市中二手房和新房价格的影响因素不同,因此分别构建模型。

(1)对于二手住宅销售价格指数:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \varepsilon_i,$$
$$\varepsilon_i \sim N(0,1), i = 1,\cdots,n; j = 1,\cdots,m \quad (7)$$

其中,$Y_1, Y_2, \cdots, Y_{16}$ 分别表示为 16 个城市:北京、上海、广州、深圳、天津、重庆、南京、武汉、沈阳、西安、成都、杭州、青岛、厦门、长沙和海口。$x_j (j = 1,\cdots,12)$ 为适用于二手房的 12 个关键词搜索指数,分别是:房价走势、房源、装修、房产网、公积金、房贷利率、二手房交易流程、二手房交易税费、房产中介、二手房、房产税和房屋出租,即 $m = 12$。

因此,最终 16 个城市中,每个城市都建立一个回归模型,并且由于每个城市二手房价格的影响因素不同,最终得到的自变量以及自变量个数均有可能不同。

(2)对于新建商品住宅销售价格指数:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \varepsilon_i,$$
$$\varepsilon_i \sim N(0,1), i = 1,\cdots,n; j = 1,\cdots,m \quad (8)$$

其中,$Y_1, Y_2, \cdots, Y_{16}$ 分别表示为16个城市,$x_{.j} (j = 1,\cdots,8)$ 为适用于新房的 8 个关键词搜索指数,分别是:房价走势、房源、装修、房产网、公积金、房贷利率、新楼盘、保障房),即 $m = 8$。

因此,最终 16 个城市中,每个城市都建立一个

回归模型，并且由于每个城市新房价格的影响因素不同，最终得到的自变量以及自变量个数均有可能不同。

2. 交叉验证技术下的多种模型。

针对每个城市，我们首先对全部数据做一个简单的线性回归，但是发现 R^2 的结果不太理想，而且大部分自变量的 p-value 大于 0.005。然后我们用函数 cor(correlation)计算了各个变量之间的相关系数和两两散点图。从相关系数和两两散点图中看不出各个变量之间的任何模式，因此本文运用 3 折交叉验证技术，采用线性回归、回归树、Bagging、Boosting、随机森林、支持向量机 6 种模型进行预测，并对各种模型的预测结果进行比较。

我们对这 6 种算法都用 3 折交叉验证的方法来判断其结果的可靠性。通过构造交叉验证随机建立的 3 个训练集和测试集，每个数据集建立 6 个模型，对每个训练集和测试集分别得到 6 个标准化均方误差(NMSE)和均方误差(MSE)，再算出 3 次 NMSE 和 MSE 的平均值。令 $\bar{y}$ 为因变量均值，$\hat{y}$ 为从训练集得到的模型对一个数据集(可能是训练集本身也可能是测试集)的预测值，这里 NMSE 的定义为：

$$\text{NMSE} = \frac{\overline{(y-\hat{y})^2}}{(y-\bar{y})^2} = \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \qquad (9)$$

MSE 定义为：

$$\text{MSE} = \overline{(y-\hat{y})^2} = \sum(y-\hat{y})^2 \qquad (10)$$

NMSE 的值越小，模型的拟合度越好；MSE 的值越小，模型的稳定性越好。

三、房地产价格预测实证分析

(一)二手住宅销售价格指数预测

本文的研究对象是 16 个大中城市的二手房价格和新房价格。对于二手房价格，16 个城市的价格走势预测步骤基本相同，这里以北京二手住宅销售价格指数的数据分析过程和预测结果为例进行说明。本文采用 R 技术进行模型建立和分析，具体过程如下。

第一步，对因变量 Y(二手住宅同比销售价格指数)进行分析，得到柱状图和 Q-Q 图，从图中可以看出因变量不符合正态分布。

第二步，根据模型(7)进行线性回归，结果显示所有自变量的 P-value 均大于 0.05，影响不显著，因此需要寻找更适合的模型进行分析。

第三步，计算各个变量之间的相关系数，绘制两两散点图。相关系数和散点图能够直观地反映变量之间的关系，但看不出各个变量间的任何模式。

第四步，应用 AIC 准则在逐步回归函数中选取自变量，最终选取 7 个对北京二手房价格指数影响最大的自变量，分别是房价走势、房源、装修、公积金、房贷利率、二手房交易流程和房屋出租搜索指数。

从这些变量可以看出，北京市居民在进行二手房交易时，除了对房源、装修等房屋本身的特性关注外，对房价走势、公积金、房贷利率等经济形势和政策比较关心。此外，还有二手房交易特有的交易流程、房屋出租信息也是进行二手房交易时重点关注的内容。

由于我国地区经济发展不均衡，各地居民收入水平、文化都存在一定差异，因此每个城市居民在进行房屋交易时关注的重点不尽相同，对二手房价格指数影响较大的自变量也不尽相同。其他城市选取的自变量见表 1。

表 1 影响 16 个大中城市二手房价格的主要关键词搜索指数

城市	搜索关键词
北京	房价走势、房源、装修、公积金、房贷利率、二手房交易流程、房屋出租
上海	房价走势、房源、装修、房贷利率、二手房交易流程、二手房交易税费、房产中介、房屋出租
广州	装修、房产网、公积金、二手房交易流程、房屋出租
深圳	装修、公积金、二手房交易流程、二手房交易税费、房产中介、房产税、房屋出租
天津	房产网、公积金、二手房交易流程、二手房交易税费、房屋出租
重庆	房价走势、装修、公积金、房贷利率、二手房交易流程、二手房、房产税、房屋出租
南京	房价走势、装修、房产网、公积金、二手房交易流程、二手房交易税费、二手房、房产税、房屋出租
武汉	房价走势、装修、房产网、公积金、二手房交易税费、二手房、房屋出租
沈阳	房价走势、装修、房产网、公积金、房贷利率、二手房交易税费、二手房、房产税
西安	房价走势、装修、房产网、公积金、二手房交易税费、房产税、房屋出租
成都	房价走势、房源、装修、公积金、房贷利率、二手房交易流程、二手房交易税费、房产中介、二手房、房产税
杭州	公积金、二手房交易流程、房产中介、二手房
青岛	房价走势、房源、公积金、二手房交易流程、二手房交易税费、二手房
厦门	房源、装修、二手房交易流程、房产中介、二手房、房屋出租
长沙	装修、房产网、公积金、二手房交易税费、二手房、房屋出租
海口	房价走势、房源、房贷利率、二手房交易流程、房产中介、房产税

从表 1 中可以发现,不同城市居民在进行二手房交易时,既有共同点,也有各自的特点。在大多数城市,人们考虑购买二手房时,首先要了解“二手房交易流程”和“二手房交易税费”,在影响 16 个城市二手房价格的重要关键词搜索指数中,这两个关键词分别出现了 12 次和 10 次;其次重点关注的就是房价走势和公积金等房产政策。所以,通过搜索引擎搜索“房价走势”和“公积金”的也比较多,分别为 10 次和 13 次。在购买二手房后,主要考虑的就是“装修”和“房屋出租”,这两个关键词分别出现了 12 次和 11 次。

第五步,构造交叉验证随机建立的 3 个训练集和测试集,每个数据集建立 6 个模型,分别计算 6 个模型的 NMSE 和均方误差 MSE,结果见表 2。其中主要关注测试集的 NMSE 和 MSE,NMSE 代表模型的拟合度,NMSE 的值越小,表明模型的拟合度越好;MSE 代表模型的稳定性,MSE 的值越小,表明模型的稳定性越好。

就北京二手住宅的数据而言,使用 SVM 方法时,模型的拟合度和稳定性最好。通过对 SVM 方法的反复试验,最终得出了最优的 SVM 模型。在 SVM 的 R-code 运行过程中最优模型的参数包括:$\varepsilon = 0.1$,$Cost = 1$,$\delta = 0.12218$。模型的预测效果见图 1。

表 2　北京市二手房各种模型所得的 NMSE 和 MSE

方法	训练集		测试集	
	NMSE	MSE	NMSE	MSE
线性回归	0.0755	10.2347	0.0755	5.4720
回归树	0.9222	67.1800	0.7306	50.4010
m-boosting	0.8455	63.2983	0.6992	47.8127
bagging	0.9073	66.3040	0.7304	50.3750
随机森林	0.2680	18.0992	0.0684	4.8674
SVM	0.2478	16.6498	0.0371	2.8414

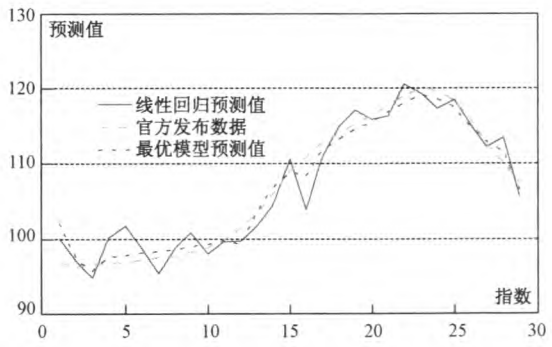

图 1　北京市二手房价格指数实际值与预测值对比图

图 1 中显示的是由线性回归所得的北京二手住宅同比销售价格指数预测值,国家统计局发布的北京二手住宅同比销售价格指数,以及从 6 种模型中选出的最佳模型支持向量机(SVM)所得出的二手住宅同比销售价格指数预测值三条曲线。从图 1 可以看出,三条曲线的变化趋势基本一致,与线性回归模型预测结果相比,由最优模型 SVM 所得出的预测值与国家统计局发布的北京二手住宅同比销售价格指数更接近。

由于影响各城市二手房走势的因素不尽相同,各地居民在购房时关注的重点也不完全一致,因变量和自变量数据特性的差异导致各个城市预测效果最好的模型也不尽相同。16 个大中城市二手房价格指数最优的预测模型如表 3 所示。

同样对除北京外的其他 15 个城市的二手房价格指数进行预测,可以看出,对所有城市而言,最优模型和线性回归模型的预测结果与实际值的走势都基本一致,与线性回归模型相比,选取的最优模型的预测结果与官方公布的实际值更为接近。尤其是对重庆和深圳而言,线性回归结果偏离实际值较大,这说明线性回归模型不适合对这两个城市的二手房价格指数进行预测。

表 3　16 个大中城市二手房价格最优预测模型

序号	城市	性能最优模型	稳定性最优模型
1	北京	SVM	SVM
2	上海	SVM	SVM
3	广州	SVM	SVM
4	深圳	SVM	SVM
5	天津	SVM	SVM
6	重庆	随机森林	随机森林
7	南京	SVM	SVM
8	武汉	SVM	SVM
9	沈阳	随机森林	随机森林
10	西安	随机森林	随机森林
11	成都	随机森林	随机森林
12	杭州	SVM	SVM
13	青岛	随机森林	随机森林
14	厦门	随机森林	随机森林
15	长沙	随机森林	随机森林
16	海口	随机森林	随机森林

(二)新建商品住宅销售价格指数拟合和预测

限于篇幅,这里同样仅列出对北京新建商品住宅销售价格指数的数据分析过程和预测结果。其他城市的预测过程基本相同,不再赘述。

第一步,对因变量 Y(新建商品住宅同比销售价格指数)进行分析,形成柱状图和 Q-Q 图,得到因变

量不符合正态分布。

第二步，根据式(8)进行线性回归，结果显示所有自变量的 P-value 均大于 0.05，影响不显著，因此需要寻找更适合的模型进行分析。

第三步，计算各个变量之间的相关系数，绘制散点图。相关系数和散点图能够直观地反映变量之间的关系，但看不出各个变量间的任何模式。

第四步，应用 AIC 准则在逐步回归函数(AICstep)中选取自变量，最终选取 3 个对北京新房价格指数影响最大的自变量，分别是房价走势、房源、装修。从这些变量可以看出，北京市居民在购买新房时，更多地关注房源、装修等房屋本身的特性，以及未来的房价走势，对于公积金、房贷利率等相关政策关心较少。值得一提的是，北京市居民在购买新房时，对保障房的关注并不多，这可能说明目前北京市保障房建设的力度还不足以影响居民购买新房的决策，例如很多人在购买新房时，由于保障房的位置较偏远、申请购买的周期较长，不得不放弃这一选择，而只能购买普通商品房。

对于其他城市，影响居民购买新房的因素则不尽相同。其他城市选取的自变量见表 4。与二手房类似，人们在进行新房交易时，通过搜索引擎搜索最多的关键词是“房价走势”、“装修”，在影响 16 个城市新房价格指数的主要关键词搜索指数中，这两个关键词都出现了 12 次，其次，“公积金”和“房产网”出现了 10 次，但是“保障房”只出现了 7 次，这表明，不只是北京，全国主要大中城市的保障房建设都有待加强和改善，使得保障房真正成为人们购买新房时的一个重要选择。

第五步，构造交叉验证随机建立的 3 个训练集和测试集，每个数据集建立 6 个模型，分别计算 6 个模型的 NMSE 和均方误差 MSE，结果见表 5。其中主要关注测试集的 NMSE 和 MSE，NMSE 代表模型的拟合度，NMSE 的值越小，表明模型的拟合度越好；MSE 代表模型的稳定性，MSE 的值越小，表明模型的稳定性越好。

显然，就北京新建商品住宅的数据而言，使用 SVM 方法时，模型的性能和稳定性最好。通过对 SVM 方法的反复试验，最终得出了最优的 SVM 模型。在 SVM 的 R-code 运行过程中最优模型的参数包括：$\varepsilon = 0.1$，$Cost\ c = 1$，$\delta = 0.53856$。模型的预测效果见图 2。

表 4　影响 16 个大中城市新房价格指数的主要关键词搜索指数

城市	搜索关键词
北京	房价走势、房源、装修
上海	房源、装修、保障房
广州	房价走势、装修、房产网、公积金、房贷利率、新楼盘、保障房
深圳	房价走势、装修、房产网、公积金、新楼盘、保障房
天津	房价走势、房产网、公积金
重庆	房价走势、装修、公积金、房贷利率
南京	房价走势、房产网、公积金、房贷利率、新楼盘
武汉	装修、房产网、公积金、房贷利率、保障房
沈阳	房价走势、装修、公积金
西安	装修、房产网、公积金、房贷利率
成都	房价走势、装修、房贷利率
杭州	房价走势、装修、房产网、公积金、新楼盘、保障房
青岛	房价走势、房产网、新楼盘
厦门	房产网、新楼盘、保障房
长沙	房价走势、装修、公积金
海口	房价走势、房源、装修、房产网、房贷利率、新楼盘、保障房

表 5　北京新房各种模型所得的 NMSE 和 MSE

方法	训练集		测试集	
	NMSE	MSE	NMSE	MSE
线性回归	0.2095	14.2550	0.1177	8.5054
回归树	1.4388	93.4264	0.7085	48.5757
m-boosting	1.1842	80.5367	0.6852	46.6549
bagging	1.1852	80.3941	0.7118	48.8358
随机森林	0.3446	22.0088	0.0715	5.0169
SVM	0.2638	17.6334	0.0542	3.8142

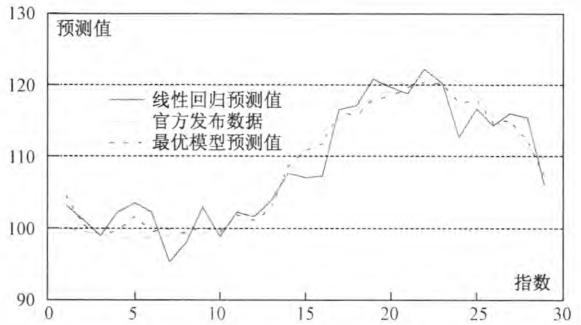

图 2　北京市新房价格指数实际值与预测值对比图

这个预测曲线图的三条曲线是由线性回归所得的北京新建商品住宅销售价格指数同比预测值。国家统计局发布的北京新建商品住宅销售价格指数同比同比值，以及从 6 种模型中选出的最佳模型支持向量集(SVM)所得出的新建商品住宅销售价格指数同比预测值组成。我们可以看出三条曲线的变化趋势基本一致，由最优模型 SVM 所得出的预测值比线形回归所得的预测值更接近国家统计局发布的北京新建商品住宅销售价格指数同比值。

由于影响各城市新房价格的因素不尽相同,各地居民在购房时关注的重点也不完全一致,因变量和自变量数据特性的差异导致各个城市预测效果最好的模型也不尽相同。对于其他城市最优的预测模型如表 6 所示。

表 6　16 个大中城市新房价格最优预测模型

序号	城市	性能最优模型	稳定性最优模型
1	北京	SVM	SVM
2	上海	SVM	SVM
3	广州	SVM	SVM
4	深圳	随机森林	随机森林
5	天津	SVM	SVM
6	重庆	SVM	SVM
7	南京	SVM	SVM
8	武汉	SVM	SVM
9	沈阳	SVM	SVM
10	西安	随机森林	随机森林
11	成都	随机森林	随机森林
12	杭州	SVM	随机森林
13	青岛	随机森林	随机森林
14	厦门	SVM	SVM
15	长沙	SVM	SVM
16	海口	SVM	SVM

从表 3 和表 6 可以看出,对于本文的数据而言,在使用的 6 种方法中,SVM 和随机森林表现最佳,随后是 bagging、m-boosting、线性回归、回归树模型。首先,对于 16 个城市中二手房和新房选择的模型中,支持向量机(Support Vector Machine, SVM)(Cortes&Vapnik,1995)表现最好,因为它在解决小样本、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。支持向量机方法是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的,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。其次,随机森林之所以能够得出较为准确的预测值,是因为分类的过程中,对于每个随机产生的决策树中的分类输入特征向量,森林中每棵树对样本进行分类,根据每个树的权重得到最后的分类结果。所有的树训练集都是使用同样的参数,但是训练集是不同的。随机森林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。

同样对除北京外的其他 15 个城市的新房价格指数进行预测,可以看出,对所有城市而言,最优模型和线性回归模型的预测结果与实际值的走势都基本一致,与线性回归模型相比,选取的最优模型的预测结果与官方公布的实际值更为接近。尤其是对重庆和海口而言,线性回归结果偏离实际值较大,这说明线性回归模型不适合对这两个城市的新房价格指数进行预测。

四、结论与展望

(一)研究结论

本文基于百度搜索指数,采用线性回归、回归树、随机森林、Bagging、Boosting 和支持向量机 6 种模型和交叉验证技术分别对北京、上海、广州、深圳、天津、重庆等 16 个城市的二手住宅销售价格指数和新建商品住宅销售价格指数进行了拟合和预测,得出如下主要结论:

第一,通过运用交叉验证技术在 6 种模型中选择的最优模型成功地预测了 16 个城市的二手住宅和新建住宅销售价格指数。总体来看,线性回归模型和最优模型的预测结果与实际值的走势基本一致,但是最优模型的预测值与实际值更接近。线性回归模型对大部分城市的预测效果较好,但是对重庆、深圳的二手房价格和重庆、海口的新房价格拟合度不够理想。

第二,根据各模型测试集的 NMSE 和 MSE 结果,在使用的 6 种方法中,SVM 和随机森林表现最佳,其次是 bagging、m-boosting、回归树、线性回归模型。首先,在 16 个城市二手房和新房价格预测模型中,支持向量机表现最好,因为它在解决小样本、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。其次,随机森林也能够得出较为准确的预测值。其优点在于对于数据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。

第三,从影响 16 个大中城市二手住宅销售价格指数的主要关键词搜索指数来看,在大多数城市,人们考虑购买二手房时,首先要了解“二手房交易流程”和“二手房交易税费”,在影响 16 个城市二手房价格的重要关键词搜索指数中,这两个关键词分别出现了 12 次和 10 次;其次重点关注的是房价走势和公积金等房产政策,分别为 10 次和 13 次;在购买

二手房后，主要考虑的就是“装修”和“房屋出租”，这两个关键词分别出现了 12 次和 11 次。

第四，与二手房类似，人们在进行新房交易时，通过搜索引擎搜索最多的关键词是“房价走势”、“装修”，在影响 16 个城市新房价格指数的主要关键词搜索指数中，这两个关键词都出现了 12 次；其次，“公积金”和“房产网”均出现了 10 次，但是“保障房”只出现了 7 次，这表明，当前在主要大中城市保障房建设的力度还不足以影响居民购买新房的决策，很多人在购买新房时，由于保障房的位置较偏远、申请购买的周期较长，不得不放弃这一选择，而只能购买普通商品房。可见，保障房建设还有待加强和改善，使得保障房真正成为人们购买新房时的一个重要选择。

（二）研究展望

本文基于百度搜索指数，采用交叉验证技术和 6 种模型成功地拟合和预测了 16 个大中城市的新房和二手房价格指数，预测标准均方误差和均方误差最低达到 0.0152，不但具有良好的预测效果，而且与传统预测方法和官方调查数据相比具有很强的时效性。由于百度搜索指数每日实时更新，因此基于预测模型在每月 1 日即可得到上月的新房和二手房价格指数的预测数，比官方统计数据提前了两周。

事实上，本文的研究思路和方法还可以进一步拓展到其他官方统计的月度公布数据，如 CPI、PPI、居民收入、居民消费支出等，都可以通过选取适当的相关度较高的关键词，采用本文的研究方法进行预测，结果不但具有很强的时效性，对官方调查数据也是有益的补充。此外，如果能够获取月度劳动力就业情况的调查数据，还可以对月度失业率进行预测，弥补城镇登记失业率的不足。而且，可以预见，随着搜索数据量的积累，预测的精度将会越来越高。

参考文献

[1] Askitas N., Zimmermann K. F, Google Econometrics and Unemployment Forecasting[C]. Working Paper, 2009.

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

[3] Breiman, L., J. H. Friedman, R. A. Olshen, C. J. Stone. Classification and Regression Trees[M]. Chapman and Hall, New York, 1984.

[4] Cho H i, Varian H. Predicting the Present with Google Trends[C]. Technical Report, 2009, Google Inc.

[5] Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine Query data [J]. Nature, 2009, 457: 1012 - 1014.

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

[7] Jurgen A. Doornik. Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data [C]. Working Paper, 2009.

[8] Kulkarni R., Haynes K., Stough R., et al. Forecasting Housing Prices with Google Econometrics: A Demand Oriented Approach [C]. Working Paper, 2009.

[9] Schmidt T, Vosen S, Forecasting Private Consumption: Survey-based Indicators vs. Google Trends [C]. Ruhr Economic Papers, 2009.

[10] Wu L., Brynjolfsson E., The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales [C], Working Paper, 2014.

[11] 马建堂. 大数据在政府统计中的探索与应用[M]. 北京：中国统计出版社，2013.

[12] 徐继华，冯启娜，陈贞汝. 智慧政府：大数据治国时代的来临[M]. 北京：中信出版社，2014.

[13] 杨树新等. 基于网络关键词搜索的房地产价格影响因素研究[J]. 新疆财经大学学报，2013(3)：5 - 12.

[14] 杨欣，等. 基于网络搜索数据的突发事件对股票市场影响分析[J]. 数学的实践与认识，2013(12)：17 - 28.

[15] 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，2013.

[16] 张崇，等. 网络搜索数据与 CPI 的相关性研究[J]. 管理科学学报，2012(7)：50 - 59.

作者简介

董倩，女，1983 年生，2012 年毕业于美国伊利诺伊大学，获统计哲学博士学位，现为国家统计局统计科学研究所助理研究员。研究方向为竞争风险分析、贝叶斯、机器学习、数据挖掘。

孙娜娜，女，1987 年生，经济学硕士，统计师，现就职于国家统计局统计科学研究所。研究方向为数理统计。

李伟，女，1977 年生，管理学博士，现任国家统计局统计科学研究所副研究员。研究方向为经济金融统计分析。

（责任编辑：曹 麦）