International Journal of Strategic Property Management
ISSN: 1648-715X / eISSN: 1648-9179
2020 Volume 24 Issue 5: 300–312
https://doi.org/10.3846/ijspm.2020.12742

CAN MACHINE LEARNING ALGORITHMS ASSOCIATED WITH TEXT MINING FROM INTERNET DATA IMPROVE HOUSING PRICE PREDICTION PERFORMANCE?

Jian-qiang GUO[1], Shu-hen CHIANG [2,*], Min LIU[1], Chi-Chun YANG[3], Kai-yi GOU[4]

[1] *School of Business, Shandong University, Weihai, China*
[2] *Department of Finance, Chung-Yuan Christian University, Chungli, Taiwan*
[3] *Department of Accounting, National Taiwan University, Taipei, Taiwan*
[4] *School of Business, University of Sussex, Brighton and Hove, UK*

Received 02 July 2019; accepted 05 February 2020

Abstract. Housing frenzies in China have attracted widespread global attention over the past few years, but the key is how to more accurately forecast housing prices in order to establish an effective real estate policy. Based on the ubiquitousness and immediacy of Internet data, this research adopts a broader version of text mining to search for keywords in relation to housing prices and then evaluates the predictive abilities using machine learning algorithms. Our findings indicate that this new method, especially random forest, not only detects turning points, but also offers prediction ability that clearly outperforms traditional regression analysis. Overall, the prediction based on online search data through a machine learning mechanism helps us better understand the trends of house prices in China.

Keywords: housing frenzies, Internet search, text mining, machine learning.

Introduction

After the 1997 Asian financial crisis, the China government announced its privatization and commercialization of the domestic housing sector to pursue economic growth sustainability (Chen et al., 2011). However, many Chinese market characteristics have led to the present overheated housing phenomenon, including high growth and low consumption (high savings), fewer investment channels, the traditional view of land as wealth, the country's marriage and household registration systems, an urban-biased development strategy, economic growth as the top priority, and local fiscal deficits (Tsai & Chiang, 2019). The 2008 global financial crisis (GFC) forced the authorities to exercise an expansionary monetary policy that soon became associated with excessive liquidity and a mortgage credit boom, thus further skyrocketing housing prices. Wu et al. (2012) emphasized that China's housing mania far surpasses the U.S. housing boom during the 1995–2006 period. Furthermore, Glaeser et al. (2017) pointed out that the real estate industry plays an extraordinary role in urban China no matter from production or employment. Thus, our first economic motivation is to search for a feasible solution to the overheated housing question in China.

China as the second largest economy in the world is now enjoying the most successful Internet commerce and IoT (Internet of things) applications based on Internet externalities, which are magnified by a population of over 1.3 billion. Many large world-class firms such as Alibaba, Baidu, and Tencent are famous for proving the Internet's significance to China's overall economy. Compared with other countries, it is generally believed that the Internet will even play a more essential role in China's future economic development and commercial mode. We therefore suggest that applying Internet data via machine learning is a very natural way to explore high housing prices in China, which is our second economic motivation.

While in the past potential buyers would visit the offices of real estate agencies or read the real estate section of local newspaper, the primary format of a housing search nowadays is through online search engines (van Dijk & Francke, 2018). Thus, one could say that the vast majority of people looking to buy a house now by searching online fully shows the imperativeness of employing Internet

*Corresponding author. E-mail: *shchiang@cycu.edu.tw* search data for predicting housing prices (Rae & Sener, 2016). Moreover, Internet search data can encompass many different sources, including buyers, sellers, developers, and government agencies – that is, the demand and supply sides of the housing market can be integrated into comprehensive Internet information, unlike public data, that is mostly surveyed from a single, specific, and passive source. In other words, Internet data include a massive amount of valuable information that can help predict housing prices in a faster and more efficient way. As the most momentous decision for a household, it is quite clear that purchasing housing inevitably involves large-scale search activity, especially over the Internet (Maclennan & O'Sullivan, 2012). As such, housing search activity online is very useful for predicting housing prices and this is our third economic motivation.

To sum up, based on the above three economic motivations in China – global interest in housing frenzies, fast-growing Internet-related applications, and the strongest Internet search action for housing investment - the purpose of this paper is therefore to apply big data techniques like Internet search data and machine learning to obtain a better housing price prediction. We believe that a better prediction performance for housing prices is a critical step toward healthier real estate development in China.

During these times of soaring housing prices associated with high volatility, an effective housing policy must depend on how to correctly predict future housing prices by means of real-time information sources and new prediction methods. The former points to Internet data based on the following reasons. First, Internet data, which can be directly obtained by its users, are in real time. Second, Internet search data based on target orientation, rather than passive surveys, can greatly improve the prediction ability of housing prices. Third, Internet data are leading indicators of housing prices on the grounds that housing buyers often start their search for a house by browsing the Internet in advance (van Dijk & Francke, 2018). In contrast to Internet data, public data are classified into low-frequency, passive, and lagging information. As far as prediction methods are concerned, any method with the ability to include Internet data to further predict housing prices is a noteworthy choice.

Traditional econometrics cannot comprise Internet data with a great number of predictors on the grounds that the core of econometrics is to use limited variables – for example, estimated parameters of interest under the assumption of a specific and linear functional form (Choi & Varian, 2012; Wu & Brynjolfsson, 2013). Nevertheless, big data techniques provide us with greater opportunities to focus on a better way to predict housing prices via a machine learning mechanism with very flexible functions without any probability distribution covering the variety of information that exists through web mining. To sum up, given that machine learning can offer so many powerful statistical estimation advantages based on real-time and high-dimensional structure of the data as well as the most flexible function to consider interactions and nonlinear relationships among variables, we therefore introduce these new methods using Internet data into housing price predictions in the case of Shanghai, China.

Different from the past studies, we choose to apply the Baidu index as the leading web search engine rather than Google search (Wu & Brynjolfsson, 2013; Lee & Mori, 2016; Wu & Deng, 2015; Zheng et al., 2016) in order to take a closer look at housing markets through a Chinese interface. Moreover, we use a broader definition of text mining by considering all possible correlations between housing price and its keywords in order to more completely capture possible predictors that could affect Shanghai housing prices. This amounts to saying that using the Chinese version of an Internet search website (Baidu) and the introduction of text mining to expand more keywords as our predictors of housing price are our additional contributions to the real estate literature.

Based on monthly housing prices of Shanghai from 2011 to 2017, we first utilize text mining approaches to capture 29 keywords in relation to housing prices. Next, we apply three methods to predict Shanghai's housing prices and it is clear that the random forest as one type of machine learning algorithm offers the best predictive ability of housing prices according to different prediction criteria. On these grounds, we come to a conclusion that a solid forecast of housing prices based on Internet search data, text mining, and machine learning can help authorities to create an effective housing policy so as to develop a sound and stable housing market in the future.

The remainder of this paper is organized as follows. Section 1 reviews some important studies on housing prices, Internet search, and predictions. Section 2 outlines text mining and prediction methods based on machine learning. Section 3 presents and compares the descriptions of data; at the same time, text mining techniques are used to select useful keywords in relation to housing prices. Section 4 estimates and evaluates forecasting abilities among three models in order to present the importance of machine learning algorithms. Finally, a review of the conclusions is presented.

1. Literature review

In this section we first survey some studies regarding China's high housing prices in order to prove the importance of this topic. More importantly, we shall review many research studies that have touched upon Internet search data via two subsections. One focuses on predictions by a new explainable variable from Internet data under traditional regressive estimations, and the other completely applies related tools of big data, for example, machine learning to forecast economic changes.

1.1. High housing prices in China

It is surprising to find that China's real estate sector started to develop after the economic reform of 1998, but its overheated housing market and even housing frenzies have recently attracted global attention (Glaeser et al., 2017; Tsai & Chiang, 2019). Generally speaking, an overheated housing market can be investigated by the concepts of housing bubbles and housing diffusion effects. As far as housing bubbles are concerned, Hui and Yue (2006) and Tsai et al. (2015) both showed that housing bubbles have appeared in China's cities, while Ren et al. (2012) and Liu et al. (2016) suggested that there is no evidence of housing bubbles in China. In other words, whether housing bubble exists in China leaves room for a variety of doubts and interpretations. For intercity housing diffusions, Chiang (2014), Lee et al. (2016), and Weng and Gong (2017) all proved the existence of ripple effects among China's cities, except Gong et al. (2016), who found little evidence of spillovers among cities within the Pan-Pearl River Delta. According to evidence from housing bubbles and housing diffusions, it seems reasonable to conclude that housing frenzies are creating troubles in modern China, and so how to delicately forecast housing prices to set up useful and timely policy measures is the core of many economic questions.

1.2. Internet search and traditional economic prediction

Information is always the most important factor for any economic issue. Along with the ever-growing advancement in new communication technology, the Internet generates a huge amount of data encompassing words, graphs, messages, etc. Thus, how to collect and analyze big data has now become essential to economic research, and housing price prediction is no exception.

Internet search data have been applied in many fields, including epidemiology by Ginsberg et al. (2009). For economic topics, Choi and Varian (2012) used Google search data to predict five kinds of economic questions. Baker and Fradkin (2017) employed the Google Job Search index (GJSI) from Google search data, but found no effect of the unemployment insurance policy on job search. Ettredge et al. (2005) and Askitas and Zimmermann (2009) both used Google search data to discuss the U.S. unemployment rate. Guzman (2011) quoted Google search data to predict inflation.

As far as real estate is concerned, Internet search data are also widely used to predict housing prices on the grounds that the housing transaction decision must depend on a housing search process in advance, especially in any period with fast-growing housing prices (Piazzesi et al., 2020; Rae, 2015; Maclennan & O'Sullivan, 2012). In other words, since the highest search intensity exists in an overheated housing market, using Internet search data based on big data techniques is a very natural experiment for housing price prediction in the face of China's current housing frenzy. Beracha and Wintoki (2013) applied Google search data as the search intensity index to display better forecasting ability of housing prices. Wu and Deng (2015) adopted Google search data to create an information flow index (IFI) at both national and urban levels to estimate spillover effects among urban housing markets. Lee and Mori (2016), who followed the idea of Da et al. (2011), selected the search volume index (SVI) from Google search data to calculate conspicuous effects on higher housing premiums (prices). Zheng et al. (2016) introduced Google search data to set up the confidence index to explore the possibilities of rising housing prices in China. Chauvet et al. (2016) evaluated mortgage default risk by use of a new index, which they referred to as mortgage default risk index (MDRI), through Internet search data. Rae and Sener (2016) selected Rightmove, covering more than 90% of real estate transactions in the UK, to explore the spatial distribution of housing searches. Similarly, Piazzesi et al. (2020) introduced Trulia as a leading online housing market portal to understand search behavior in San Francisco. Van Dijk and Francke (2018) applied Internet search data from Funda, the largest housing website in the Netherlands, to measure a market tightness indicator through Internet search data, while van Veldhuizen et al. (2016) again used Google search data to find that Internet search data can provide useful information for housing transactions in the Netherlands.

To sum up, it is clear that Internet search data have been used to predict many economic questions, including housing prices; at the same time, Google search data are the most often-used resources. However, we also see that buyers and analysts refer to their local Internet search engine to see which location is strongly preferred in a specific housing market. The final and most important point is that all the above studies still chose to apply a traditional econometric approach to evaluate an additional benefit of estimation results by adding a new independent variable through the calculation of Internet search data; at the same time, they nearly all conclude that econometric estimations with an additional variable from Internet search data consistently generate better empirical results.

1.3. Text mining and machine learning

As mentioned above, we see two possible limitations in these articles from the viewpoint of big data. First, although many studies have started to utilize Internet search data to evaluate and predict economic changes, including housing prices, they only have developed a new index as another explainable variable through the Google search data engine on the grounds that this is the simplest way to maintain traditional econometric estimations[1]. A question now arises as to whether they should omit new Internet tools, like machine learning, to predict economic variables. Second, searching for keywords in relation to housing prices can be found in many cases. For example, Wu and Brynjolfsson (2009) selected "real estate" and "real estate agency" to forecast housing prices, while Beracha

[1] Except for time-series econometrics, Tan et al. (2017) and Lee et al. (2019) applied grey algorithm and hierarchical model to predict property prices.

and Wintoki (2013) introduced "real estate" and "rent" to predict future housing prices. Wu and Deng (2015) employed the keyword "house price" to discuss intercity housing diffusions, and Zheng et al. (2016) used "housing price" associated with "rising" or "increasing" to predict housing prices in the future. In other words, they all arbitrarily selected some keywords as the predictors to predict housing prices. What seems to be lacking, however, is to quote a text mining approach to objectively expand the field of keywords.

All these things make it clear that, even after applying Internet search data, most studies still lack actual big data applications such as text mining and machine learning. In fact, Nardo et al. (2016) mentioned that using text mining to create monitoring variables can be considered like a story that depicts which independent variables are closely related to a dependent variable; moreover, machine learning can supply the "best" story to describe the final result of a story by the best prediction performance. Thus, we want to apply text mining and machine learning to develop our story between keywords and housing prices. Varian (2014) similarly pointed out that big data possess three benefits: more powerful data manipulation, more potential predictors, and more flexible relationships. These three benefits fully correspond to Internet search data, text mining, and machine learning, respectively. However, past studies only touched the surface of Internet data – namely, the first benefit from massive Internet data collection. On the other hand, this paper places Internet search data, text mining, and machine learning together in order to fill the gap in the past research, while at the same time providing the best story of housing price prediction. We believe that applying these big data techniques to improve the forecasting ability among complicated interrelationships is necessary when exploring and resolving housing troubles in China.

Jirong et al. (2011) applied various models of machine learning over the last few years to forecast housing prices in China to prove that machine learning outperforms the other models based on its housing price prediction. Park and Bae (2015) collected daily housing price data (5,359) of Fairfax County, Virginia in the U.S. from the multiple listing service of metropolitan regional information systems (MRIS) during 2004 to 2017 and then selected 28 variables based on a hedonic-based method to compare prediction performances among four classifiers from machine learning algorithms. Plakandaras et al. (2015) found that the predictive ability of U.S. housing prices based on machine learning is clearly better than traditional vector regression (VAR) and Bayesian VAR models. Mullainathan and Spiess (2017) and Chen et al. (2017) both proposed the hedonic price theory to derive many independent variables and then applied machine learning techniques to obtain better housing price predictions in the U.S. and Taiwan, respectively.

Based on housing tenures, machine learning methods have been further extended from housing prices to housing rents in the case of China. For example, Hu et al. (2019) adopted housing rent data from on-line housing rental websites (OHRWs) in order to obtain a better understanding of fine-scale and real-time housing rent information in Shenzhen (as the most popular immigrant city), where young immigrants in pursuit of new job opportunities generally need much more rental spaces against fast-growing high housing prices. They first applied rental data from 8117 communities at the most disaggregated scale as well as a set of independent variables based on the hedonic theory and then implemented 6 machine learning algorithms to evaluate which method is the best fit for housing rental data according to prediction performance. The results revealed that two algorithms, including random forest, can be used to trace housing rental dynamics in the future. Chen et al. (2016) also quoted on-line rental information as a reliable source of real-time and fine-scale housing rental data at the most basic level in Guangzhou, determining the independent variables by nighttime lights and several types of points of interest (POIs). Based on the above information, they also chose 6 machine learning methods to predict housing rents; at the same time, to fill out no observation data in some special locations. Based on the above studies using machine learning algorithms, it is clear that they mostly focused on cross-sectional prediction performance with a large number of cross-section units across a relatively short time interval, whereas our study pays more attention on a time-series prediction for a future housing price trend. The most important point herein is to directly appeal to Internet search data by a text mining methodology to further explore the real intentions of housing participants, rather than utilize a prior theoretical foundation like the hedonic theory.

Judging from the above, the logarithmic increase in Internet information has not only changed people's everyday lives, but has also provided many more possibilities for predicting economic variables. Although a great deal of effort has been made at introducing Internet search data to analyze economic topics, surprisingly few studies have so far been comprehensively applied to forecast economic variables, including housing prices via an integration of text mining and machine learning. To our knowledge, this is the first paper that exercises a broader version of text mining in order to capture more keywords in relation to housing prices and then uses some tools of machine learning - for example, the elastic net model and random forest – to predict housing prices in Shanghai. We expect that this research will spur more interests in big data applications in the area of predictions, which would help efficiently set up workable and useful policy measures.

2. Text mining and machine learning algorithms

In this section we shall first describe the Internet penetration rate and the main Internet search engines in China. In turn, we introduce a broader view of text mining here in order to capture more keywords as a set of the predictors for housing prices. Finally, we outline some prediction models of machine learning – for example, the elastic net model and random forest – in order to compare with the traditional linear regression model.

2.1. Internet and search engines in China

The fast-growing development of the Internet has deeply penetrated and impacted all dimensions of people's lives, and it has pushed big data techniques to become a lot more popular now. In fact, China not only has the largest population in the world at 1.3 billion, but it also has the biggest Internet commerce economy. At the end of 2017, registered users of Internet search engines hit 751 million, associating with a popularity rate of 54.3%; moreover, this upward trend has continued mostly unstoppable as shown in Figure 1.[2] Based on this, applying machine learning methods associated with Internet search data can help us to predict housing prices in China.

Wu and Deng (2015) and Zheng et al. (2016) both chose the Google search engine to collect Internet data in order to explore China's housing market. However, it is clear that the leading search engine in China actually is run by Baidu, rather than Google. According to StatCounter global statistics (February, 2019) as in Figure 2, the market shares of the top five search engine webs in China are Baidu (74.63%), Shenma (13.52%), Sogou (4.78%), Haosou (3.16%), and then Google (2.03%), respectively. Thus, although Google is the best known search engine with a global market share of more than 92%, Baidu is undoubtedly the leading search engine in China on the grounds that Baidu's website fits in with China's Internet users' utilization of simplified Chinese characters (for words) rather than an English word interface. In addition, it is generally established that housing transactions mainly come from local Chinese habitants. We therefore decide to take the Baidu search engine as the source of our Internet search data to investigate China's housing market.

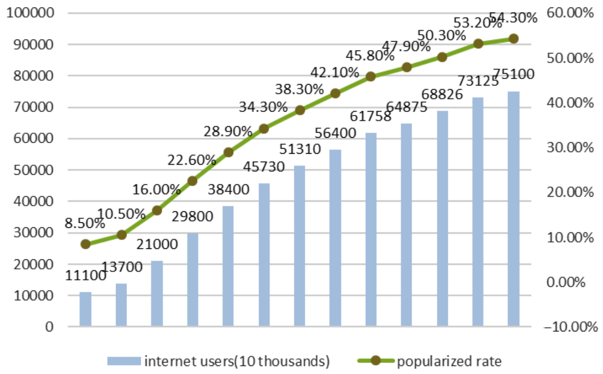

Figure 1. Internet popularity trend in China (2005–2017)

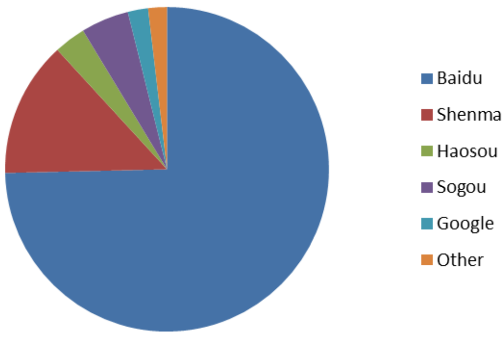

Figure 2. Market shares of search engines in China (Feb. 2019)
(source: Statcounter global stats)

2.2. Text mining

Even though we know that Internet data can represent potential emotions of Internet users, how to determine the critical keywords as adequate predictors in order to trace their true motivations for housing transactions remains an unsettled question. Text mining here provides a workable answer by the extraction of high-quality information from unstructured data – for example, words – in order to identify reasonable keywords in relation to housing prices.

To search for all possible representations of housing prices, we first apply the keyword tool of the Baidu search engine to collect the first type of initial keywords. Compared to the first type of keywords from Baidu, the second type of keywords mainly stems from an academic resource – namely, Chinese National Knowledge Infrastructure (CNKI), which is the largest Chinese full-text database, including more than 9.000 journals across the fields of economics, education, business, and others. Next, we apply Citespace software for analyzing the cluster of possible keywords and find 6.316 papers in relation to housing prices. Specifically, this software generates a connectedness map of Chinese keywords, where bigger (smaller) words imply more (less) correlations with housing prices, and they are classified into the second type of keywords.

Due to overly professional and academic writings, we take the second keywords by web crawler to gather more words into our training bank. However, this expanding-keyword process may lead to long-tail keywords, and so we further employ Jieba to solve this question via application of a useful Chinese word segmentation (participle) module. We set up a corpora bank that is used to manage a natural language from the texts. Furthermore, we introduce word2vec as a neural network to train a large corpus of text into a vector space and eventually derive many keywords relative to our focus on housing prices. Finally, we delete some repeated and meaningless terms by applying SQL (structured query language) to detect two kinds of keywords from the Baidu website data and CNKI database. The remaining forms our keywords for housing prices.[3]

[2] Popularity rate is the ratio of Internet users to population.

[3] As Varian (2014) mentioned, the SQL system can be used to manipulate big data.

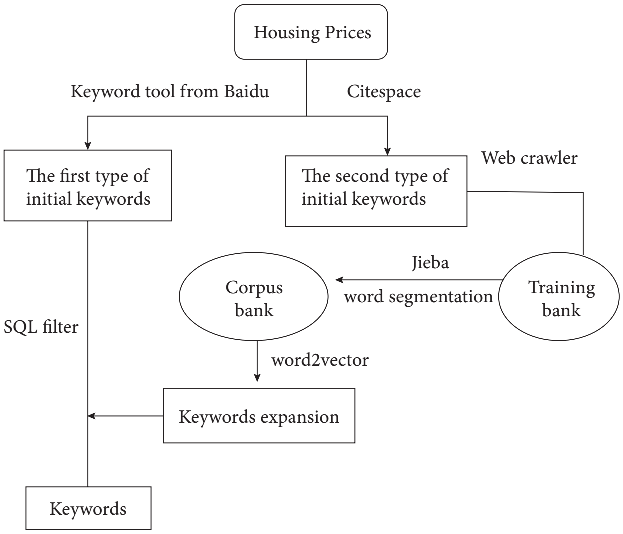

Figure 3. The structure of text mining to search for keywords

As stated above, we obtain all possible keywords that are directly and indirectly related to housing prices through two channels: one is extracted from Baidu to display the preferences of ordinary Internet users – for example, buyers and sellers of real estate; while another is obtained by the CNKI database covering academic journals in order to show the concept of housing prices from the viewpoint of scholars and experts. After combining two kinds of keywords from different angles, we are very confident that all potential keywords regarding housing prices have been included. Figure 3 presents all these steps to obtain these keywords. Lastly, we again want to emphasize the importance of a Chinese interface to demonstrate the local appetite for housing assets. Baidu's website and Jieba software are two typical examples using Chinese words.

We finally must begin to quantify these keywords by inserting them into the Baidu search indices, which are similar to Google Analytics. They are available since 2011 on daily and monthly bases to calculate search volumes from the Baidu search engine to collect all structured data based on the Baidu indices of our keywords.

2.3. Prediction models using machine learning

We now try to apply the three models to predict housing prices in Shanghai, including the traditional linear regression model and two models based on machine learning algorithms. The elastic net model is regarded as a parametric prediction that is an extended regression model to non-linear forms, while the random forest model is regarded as a non-parametric prediction that is expanded from a single tree – for example, the decision tree (Mullainathan & Spiess, 2017).

Machine learning is an especially noteworthy approach by use of a flexible function covering a large number of related keywords and millions or even billions of observations. However, it is very important to note that machine learning, which can comprise very large dimensions of explainable variables, only focuses on prediction, rather than parameter estimation, on the grounds that the initial idea behind machine learning is to propose a better prediction performance by means of the most complicated and flexible interactions among all variables (Mullainathan & Spiess, 2017). At the same time, to prevent excessive complexity or overfitting, machine learning often introduces a validation mechanism as a form of regularization to choose the model's optimal depth. The most common way is ten-fold cross-validation, which divides the data into ten subsets (folds) in order to train and test the data for how well your chosen model performs in this section.

2.3.1. Generalized linear regression model

This model is a linear multiple regression over many independent variables (predictors, x) based on the assumption of a normal distribution of the residuals (ε_i) as (1) with P predictors:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i \, . \quad (1)$$

To search for the best linear unbiased estimator (BLUE) regarding the impacts of P predictors on the dependent variable y, we must minimize the loss function – namely, the sum of squares residuals (SSR) below:

$$\min\left(SSR\right) = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ji} \right)^2 . \quad (2)$$

The biggest shortcoming of this approach is multicollinearity, which happens in extremely high correlations among variables. Another restriction of a linear regression is to limit many possible interactions among variables in order to maintain a linear function. These two questions will be resolved by the following methods from machine learning algorithms.

2.3.2. Elastic net model

When our data are relatively fat (namely, lots of predictors), we must select adequate features - namely, the variable selection (Variant, 2014) – so as to simultaneously simplify the model, to avoid the overfit problem, and to reduce training time. To achieve these goals, we set up a penalized regression by using different regularizations. LASSO (least absolute shrinkage and selection operator) and ridge regression are two notable examples.

LASSO is a penalty regression with a quadratic loss function that introduces a penalty term associated with SSR as in (3). From (3), it is clear that $\lambda \sum_{j=1}^{p} \left| \beta_j \right|$ is a kind of shrinkage penalty, and λ is the "tuning" parameter. On the other hand, the ridge regression is another penalty regression with a quadratic regularizer that inserts another penalty term into the original SSR term as in (4), where $\sum_{j=1}^{p} \beta_j^2$ is another kind of shrinkage penalty, and λ is still a tuning parameter or complexity parameter.

$$\min_{\beta}\left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ji} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right) = \min\left(SSR + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right); \quad (3)$$

$$\min_{\beta}\left(\sum\nolimits_{i=1}^{n}\left(y_i-\beta_0-\sum\nolimits_{j=1}^{p}\quad\right)^2+\lambda\sum\nolimits_{j=1}^{p}\beta_j^2\right)=\min\left(SSR+\lambda\sum\nolimits_{j=1}^{p}\beta_j^2\right). \tag{4}$$

Thus, LASSO combined with ridge regression is called the elastic net model, and this model possesses a penalty factor as in (5).

$$\lambda\sum\nolimits_{j=1}^{P}\left[\left(1-\alpha\right)\left|\beta_j\right|+\alpha\beta_j^2\right]. \tag{5}$$

The estimation method in (5) contains two methods as special cases. If α = 1, then there is only the quadratic constraint, which is a ridge regression. If α = 0, then this is called LASSO. To sum up, applying different penalty strategies can help select useful predictors to improve the overfitting question.

2.3.3. Random forest

Traditional linear regression does not involve non-linear and complicated interactions among variables, but regression trees can solve these questions, especially for high-dimensional datasets. However, a single tree may generate poor performance, and so adding randomness into the decision tree via bootstrap, bagging, and boosting can improve the prediction ability greatly, and this means random forest, which uses many trees. Howard and Bowles (2012) stated that random forest is the most successful learning algorithm for prediction.

There are many steps in order to show random forest as follows (Variant, 2014):

1. Select a bootstrap sample of the observations to grow a tree.
2. At each point of the tree, choose a random sample of the prediction to make the next decision.
3. Repeat step 2 many times to grow a forest of trees.
4. Average the results of all trees to calculate the prediction performance.

2.3.4. Cross-validation

Machine learning divides the data into three parts: training, testing, and validation. Training data can obtain a model, while validation and testing the data can help choose a better model. As mentioned above, to avoid excessive complexity we must select a good "tuning" parameter – for example, the optimal variable selection in the elastic net model and the optimal depth of the tree in the random forest. We summarize k-fold cross-validation (k = 10 is the most common choice) as follows.

1. Divide the data into roughly k equal subsets (folds) and label them by s = 1,.., k. Start with subset s = 1.
2. Pick a value for the tuning parameter.
3. Fit your model using k–1 subsets other than subset s.
4. Predict subset s and measure the associated loss.
5. Stop if s = k; otherwise, increment s by a and go to step 2.

Cross-validation is applied so as to increase the efficiency of the prediction procedure. Here, we randomly partition the sample into equally-sized subsamples (folds). Finally, we pick the parameter with the best estimated average performance.

3. Data description and predictors by text mining

In this section, we shall first outline housing prices in Shanghai. Next, we exercise text mining as section 2.2 to show our keywords as predictors here to predict housing price in the next section.

3.1. Housing prices of Shanghai

Shanghai as the economic center in China is a famous international city, has a population of over 20 million, is the largest city in the country, and is the second largest metropolitan area throughout the world. Based on its powerful economic competitiveness, such as owning the highest ratio of educated employees, having a financial market, and being a transportation and trade hub, its gross regional product per capita has been over US$20,000 since 2017; at the same time, the role of Shanghai in China's economy is very important, making up 3.6% of total gross domestic product. More importantly, the real estate sector in Shanghai is one of the six largest industries. Based on the above, we think that selecting Shanghai's housing prices is adequate for our study.

To understand the merits of housing price predictions based on the three models, we must collect actual housing prices as the starting points of our study. Since 2006, the National Bureau of Statistics (NBS) of China has officially announced the monthly housing sale prices of seventy large- and medium-size cities, which have been further classified by three levels; Shanghai belongs among the first-tier cities.[4] At the same time, the data come from new housing transactions and not second-hand transactions. These data are widely used to investigate the trends in China's housing market (Liu et al., 2016), and so we quote Shanghai's housing prices from this database. However, the Baidu search data only trace back to 2011, and thus we select housing prices of Shanghai from 2011 to 2017 with 84 observations in order to meet data consistency between housing prices and Baidu search data.

3.2. Keywords in relation to housing prices

Through the process of text mining as section 2.2, we obtain all 29 keywords in order to set up the predictors of housing prices; moreover, we classify them into 4 groups as in Table 1 based on economic viewpoints – for example, macro-level policies, local attributes, housing market characteristics, and housing costs, respectively. This table shows that even when we introduce text mining to capture the intentions of online housing buying behaviors, these keywords are still related to traditional economic theories. However, compared to economic theory, text mining can help us be closer to the Internet world of housing searchers.

[4] The other three first-tier cities are Beijing, Guangzhou, and Shenzhen.

Table 1. Keywords of housing prices in relation to economic aspects

Economic aspects	Keywords
Macro policies (7)	Urbanization, rail transportation, real estate policy, pension fund, macro control, monetary policy, inflation
Local attributes (6)	Shanghai's second-hand house, house web, house, house price, rental house and school district house
Housing market characteristics (9)	Second-hand house, second-hand web, housing, housing price, housing frenzies, rising prices, price/income ratio, house, rent house web
Housing costs (7)	Housing fee, housing tax, mortgage calculator, mortgage interest, down payment, property tax, decoration

Moreover, Table 2 shows the descriptive statistics of the Baidu indices from the 29 main keywords during the period 2011–2017 on a monthly basis. We obtain the data of keywords by using the Baidu search engine to explore the relative advantages of the three models (generalized regression, elastic net, and random forest models) based on their prediction ability of housing price in Shanghai.

4. Estimation results and prediction performance

In this section, we first implement the generalized regression model, elastic net model, and random forest, respectively to present their individual estimation results. Besides, we apply total-sample and out-of-sample prediction methods to evaluate their predictive abilities of housing prices in Shanghai.

Table 2. Descriptive statistics of Baidu indices of the main keywords for 2011–2017

Keywords	Min	Max	Mean	SD	Kurtosis	Skew
Urbanization (城镇化)	218	3151	858.58	659.95	1.61	2.42
Second-hand house (二手房)	4385	20116	9622.51	4045.52	0.72	–0.59
Housing fee (二手房税费)	131	2344	578.63	492.37	1.98	3.22
Second-hand web (二手房网站)	131	524	228.41	91.67	1.49	1.22
Housing (房产)	1730	6744	2627.94	852.88	3.05	11.69
Housing tax (房产税)	1333	16372	4280.14	2552.16	1.90	5.62
Mortgage calculator (房贷计算器)	3060	45456	15694.46	9396.91	0.78	0.26
Mortgage interest (房贷利率)	838	3491	1736.18	585.96	0.76	0.24
Real estate policy (房地产政策)	32	711	252.24	116.14	1.70	3.49
Housing price (房价)	2329	12280	4208.37	1896.32	2.18	5.84
Housing frenzies (房价暴跌)	71	2055	491.42	418.16	1.23	1.23
Rising prices (房价上涨)	141	1569	286.24	210.79	3.90	18.95
Price/income ratio (房价收入比)	133	1010	294.46	145.68	2.39	8.12
Pension fund (公积金)	3565	11906	7686.77	1820.67	0.14	–0.16
Rail transportation (轨道交通)	503	1484	833.00	209.60	0.98	0.61
Macro-control (宏观调控)	429	1443	1000.85	280.11	–0.35	–0.99
Monetary policy (货币政策)	386	1448	884.83	250.85	0.11	–0.68
House (楼盘)	365	1026	669.27	159.92	0.44	–0.38
Down payment (买房首付)	200	504	298.77	69.47	1.09	0.54
Second-hand house, Shanghai (上海二手房)	310	28625	2687.54	4449.72	4.30	20.78
Shanghai's house web (上海房产网)	340	1527	921.28	336.71	–0.67	–0.91
House price, Shanghai (上海房价)	398	11702	2219.92	1399.85	4.20	26.26
Shanghai house (上海楼盘)	165	523	253.54	71.13	1.63	2.79
Shanghai rental house (上海租房)	1582	4152	2540.76	565.72	0.57	–0.33
Inflation (通货膨胀)	1479	3354	2111.22	378.99	1.24	2.06
Property tax (物业税)	117	446	195.35	61.85	1.57	3.08
School district house (学区房)	186	3818	929.15	612.93	1.99	6.61
Decoration (装修)	3700	31572	8685.14	5038.58	2.31	6.17
Rent house net (租房网站)	179	923	433.21	132.04	1.22	2.39

Note: Keywords are English words associated with the Chinese meanings in parentheses.

4.1. Estimation process

We first apply (1) to estimate a generalized regression model using search data of keywords regarding housing prices. An initial result is shown in Table 3, including keywords, coefficients, and their significant levels. Here, we see that the impacts of many keywords on housing prices are insignificant (there are only seven variables with statistical significance), whereby R^2 is 0.89, and the value of the F test is over 22. In other words, multicollinearity is fully reflected in this outcome. To solve this question, we further exercise a stepwise regression to make all selected variables have statistical significance as in Table 4, whereby $R^2 = 0.91$, and the value of F test = 75.97.

We now further explore elastic net models by extending the generalized regression into the field of machine learning algorithms, including LASSO and ridge regression. As mentioned by Mullainathan and Spiess (2017), LASSO is very familiar to economists due to its similarity with econometrics.

Table 3. Estimation results of main keywords based on generalized regression

Keywords	Coefficients	Keywords	Coefficients
Intercept	1.60*	Rail transportation	–0.02
Urbanization	–0.09	Macro-control	0.04
Second-hand house	0.05	Monetary policy	0.16**
Housing fee	0.06	House	–0.07
Second-hand web	0.27**	Down payment	–0.19**
Housing	–0.04	Shanghai second-hand house	0.08
Housing tax	0.05	Shanghai's house web	–0.08
Mortgage calculator	0.40**	Shanghai's house price	0.04
Mortgage interest	–0.01	Shanghai housing	–0.03
Housing policy	–0.01	Shanghai rent house	–0.03
Housing price	–0.17**	Inflation	–0.07
Housing frenzies	0.23**	Property tax	–0.08
Rising prices	0.02	School-district house	–0.09
Price/income ratio	0.04	Decoration	0.25***
Pension fund	0.12	Rent house web	0.10

Note: ***, **, and * represent 1%, 5%, and 10% statistical significance levels, respectively.

Table 4. The outcome of a stepwise regression

Keywords	Coefficients	Keywords	Coefficients
Intercept	1.46***	Housing frenzies	0.25***
Urbanization	–0.11**	Monetary policy	0.15***
Second-hand web	0.36***	Down payment	–0.21***
Housing fee	0.10*	Property tax	–0.16**
Mortgage calculator	0.41***	Decoration	0.26***
Housing price	–0.18***		

Note: ***, **, and * represent 1%, 5%, and 10% statistical significance levels, respectively.

Table 5. Selected keywords based on the elastic net model

Keywords	Coefficients	Keywords	Coefficients
Urbanization	–0.026	Housing price	–0.064
Mortgage interest	0	Down payment	–0.001
Housing policy	–0.005	Pension fund	0.084
Macro-control	0.059	Monetary policy	0.099
Property tax	–0.052	Shanghai second-hand house	0.104

To select the best model, we implement 7000×100×10×80 times of traversals to get $\alpha = 0.69$, and this result is clearly close to LASSO, rather than to ridge regression. More importantly, when we use the tuning parameter (namely, $\lambda = 0.0684$), this estimation function can automatically and efficiently select critical keywords and eliminate insignificant keywords as in Table 5.

We finally apply the random forest to predict housing prices in Shanghai. Similarly, we run 50×100×10 times of traversals to obtain the best model. It is important to note that we never see all estimated parameters based on the random forest method, because thousands of trees (forests) are always hard to be explained, but we are able to clearly understand the relative importance of selected keywords to housing price prediction based on the contribution to improvements in prediction inaccuracy (mean-square errors, MSE) as shown in Table 6.[5] From this table, it is clear that Shanghai's second-hand housing prices are the most important factor to predict Shanghai's overall housing prices and mortgage interest comes in second place. In other words, the status of the second-hand housing market and the level of mortgage interest are both critical for housing price prediction in Shanghai.

[5] Varian (2014) stated that no outcome from the estimated parameters under the random forest denotes a "black box".

Table 6. Relative importance of main keywords

Keywords	Degree of reduction in *MSE*	Keywords	Degree of reduction in *MSE*
Urbanization	0.025	Housing fee	0.001
Mortgage interest	0.043	Rising prices	0.001
Price/income ratio	0.002	Pension fund	0.003
Macro-control	0.007	Monetary policy	0.002
Property tax	0.004	Shanghai second-hand house	0.794

Table 7. Goodness of fit of the three models

Model / Index	Generalized regression	Random forest	Elastic net
MSE	0.1021	0.0190	0.122
R^2	0.91	0.98	0.90

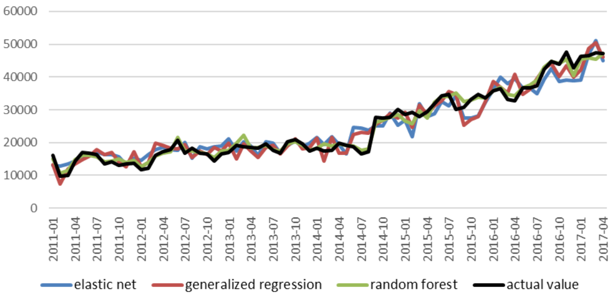

Figure 4. The trends between real and fitted housing prices

4.2. Model evaluation

After the estimation results of all three models, we must carefully evaluate their merits, especially their predictive abilities of housing prices. We plan to present them in two parts: goodness of fit for the total sample and prediction performance based on out-of-sample.

We first quote R^2 and mean-square errors $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y})^2$, where $\hat{Y}$ is the estimated housing price, to understand the degrees of goodness of fit over the total sample (2011–2017) as in Table 7. It is clear that random forest exhibits the best goodness of fit for Shanghai's housing prices with the highest value of R^2 and the lowest MSE. The worst goodness of fit is found in the elastic net model, which is even behind the generalized regression.[6]

We further apply actual housing price data in Shanghai by a comparison with the fitted housing prices through generalized regression, elastic net, and random forest models, respectively, in Figure 4. This figure aptly illustrates that the fitted value of housing prices in Shanghai based on random forest is better at capturing the actual trend in housing prices versus the other two approaches, which deviate from true housing prices at many time points. Moreover, it is noteworthy that random forest not only can fit the actual housing prices very well, but that it also fully expresses its extraordinary ability to recognize the timing of turning points.

We also want to directly compare the relative prediction abilities for the out-of-sample. Thus, we design the data from July to December 2017 (6 months) as our goal of prediction; at the same time, we use other data (namely, in-sample data for the period January 2011 to June 2017 with 78 observations) to estimate all parameters. In other words, we divide the total sample into in-sample and out-of-sample, whereby the former is used to estimate an empirical model, which is further used to examine the prediction performance of the latter.

We respectively provide two indices, mean absolute percentage error (MAPE) and root-mean-square error (RMSE), as (6) and (7) for a comparison between actual values and average prediction values as in Table 8.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100; \quad (6)$$

$$RMSE = \sqrt{\sum_{t=1}^{n}\frac{\left(F_t - A_t\right)^2}{n}} \times 100. \quad (7)$$

Here, *A* and *F* represent the actual and predicted values of time *t*, respectively. Lower values of these two indices imply better prediction performance. Based on these two indices, we easily see that random forest possesses the best prediction performance based on out-of-sample fits with the lowest values of MAPE and RMSE, the elastic net model has the second best forecasting performance, and the generalized regression has the lowest performance.

According to total sample fits, random forest is best, followed by generalized regression, and then the elastic net model. As far as out-of-sample fits are concerned, random forest is still the best prediction method in contrast with the generalized regression having the lowest prediction performance. Thus, we can more firmly state that random forest is the first choice to predict housing prices in Shanghai, and so the random forest model can be used as an early warning system of future housing prices in Shanghai. In other words, there is evidence here to show that machine learning can improve the predictive ability for Shanghai's recent housing prices. In addition, based on machine learning algorithms, such as elastic net model

[6] In Table 7, it is found that R^2 is very high, especially random forest with 0.98 on the grounds that 29 independent variables as well as their non-linear and interactions can be used to explain the variant of housing prices. In fact, although it is believed that R^2 in a cross-section estimation is far lower than a time-series estimation, the cross-sectional estimations with enormous cross-section units of housing rents, such like Chen et al. (2016) and Hu et al. (2019) still saw high values of R^2 with more than 0.7.

Table 8. Prediction performance of the three models

2017	Actual price	Prediction of generalized regression	Prediction of random forest	Prediction of elastic net
July	46183	40559	41173	40454
August	48978	39821	48410	40018
September	45764	40933	40093	41266
October	48191	41899	46188	48457
November	49105	42435	43414	44206
December	49317	44279	49893	48001
MAPE		13.05	6.89	8.95
RMSE		6432.71	3964.72	5150.05

and random forest model, we find that Shanghai's second-hand housing market is an especially important factor for predicting housing prices in Shanghai. Put differently, the second-hand housing market of this city is regarded as a critical point for price discovery of a new housing market.

The primary argument against active policies is that the policymaking effectiveness seriously suffers from a succession of time-lag questions by the application of public data. Even when a new policy is ready to be implemented, the condition of the economy may have changed. Thus, real-time Internet data, by directly appealing to online users regarding housing transactions, can resolve this debate. Moreover, economic forecasting is often imprecise, and so accurate and timely prediction of housing price is really essential to any proper pre-evaluation and subsequent useful political programs and implementations. In other words, as long as we can accurately predict the condition of housing prices in advance, then policymakers could look ahead when making "good" policy decisions. Machine learning algorithms – for example, the random forest model in this paper that offers the best solution of housing price prediction – are very critical for setting up real estate policies in China.

Conclusions

Using web-related services in this Internet era has become a substantial part of everday life, and the penetration rate is exhibiting non-stop growth. Thus, figuring out a way to trace every footprint from the Internet can assist us to comprehensively investigate real-time human behavior in order to get the best understanding of many economic debates. Moreover, along with the many technological advances in the Internet and their related applications, academia must start to consider how to handle and analyze big data via new methods like machine learning algorithms on the grounds that traditional econometrics cannot deal with the massive amounts of data covering a wide variety of sources and variables.

Compared to past studies, this paper offers three contributions to the housing price prediction literature. First, we select Baidu, instead of Google, to take a closer look at a Chinese version of housing price prediction. Second, we propose text mining methodologies to extract useful information from Internet search data by keywords in relation to housing prices. Finally, this paper has adopted machine learning algorithms (versus a traditional regression method) to evaluate prediction performance, finding that random forest is better at predicting Shanghai's housing prices. Thus, the authorities can introduce random forest as the basis for housing price prediction and to monitor the trend of housing prices in the future; at the same time, they can follow prediction outcomes by machine learning algorithms to establish effective and timely real estate policies.

We do note some limitations in the machine learning mechanism. Just as Mullainathan and Spiess (2017) pointed out, machine learning algorithms only target the prediction problem by discovering a very complicated and flexible structure with no need of model and variable specifications. However, the machine learning method cannot be used to estimate and infer any parameter from probability distributions of explained and explainable variables, because of no standard errors. In other words, using a machine learning algorithm cannot solve the causal relationship between independent and dependent variables in order to further show economic meanings and inferences, and this is the price of using machine learning, instead of econometric analysis. Mullainathan and Obermeyer (2017) additionally emphasized that three types of mismeasurement of independent variables can bias the prediction outcomes of machine learning: subjective, selective, and event-based; this gives rise to moral hazard and error based on various types of Internet data, such as images, languages, and others. To sum up, there are two shortcomings: an inability to estimate parameters, and taking risk at a mismeasurement of independent variables must constantly be kept in mind when applying machine learning algorithms to predict any economic variable.

China is a notable example for a high degree of Internet usage and applications, and so it is natural to apply Internet search data with machine learning methods to predict housing prices here. Moreover, China's housing frenzies have attracted more and more interests, so we adopt Internet search data via Baidu, text mining to search for useful keywords as effective predictors, and eventually machine learning to predict housing prices. Based on the estimation results, it is clear that random forest as one type of machine learning algorithm is the best prediction tool for housing prices in Shanghai, and the findings herein can help the authorities to propose useful policies to prevent possible housing bubbles in China. We expect that this study will attract academic interest in the areas of prediction via the application of machine learning algorithms.

Funding

This work was supported by the Shandong Social Science Planning Fund of China under Grant [number 16CJJJ08].

Author contributions

Jian-qiang Guo was responsible for the concept of forecasting model and economic meanings. Shu-hen Chiang was responsible for the interpretation of estimation results and the first draft of the article. Min Liu was responsible for text mining and machine learning methodologies. Chi-Chun Yang and Kai-yi Guo were responsible for the design and development of the data analysis.

Disclosure statement

The authors declare no conflict of interest.

References

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, *50*, 107–120. https://doi.org/10.3790/aeq.55.2.107

Baker, S., & Fradkin, A. (2017). The impact of unemployment insurance on job search: evidence from Google search data. *Review of Economics and Statistics*, *99*, 756–768. https://doi.org/10.1162/REST_a_00674

Beracha, E., & Wintoki, M. B. (2013). Forecasting residential real estate price changes from online search activity. *Journal of Real Estate Research*, *35*, 283–312. https://aresjournals.org/doi/abs/10.5555/rees.35.3.c0ru080q45n34064

Chauvet, M., Gabriel, S. A., & Lutz, C. (2016). Mortgage default risk: new evidence from internet search queries. *Journal of Urban Economics*, *96*, 91–111. https://doi.org/10.1016/j.jue.2016.08.004

Chen, J., Guo, F., & Wu, Y. (2011). One decade of urban housing reform in China: urban housing price dynamics and the role of migration and urbanization, 1995-2005. *Habitat International*, *35*, 1–8. https://doi.org/10.1016/j.habitatint.2010.02.003

Chen, J., Ong, C., Zheng, L., & Hsu, S. (2017). Forecasting spatial dynamics of the housing market using support vector machine. *International Journal of Strategic Property Management*, *21*, 273–283. https://doi.org/10.3846/1648715X.2016.1259190

Chen, Y., Liu, X., Li, X., Liu, Y., & Xu, X. (2016). Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning. *Applied Geography*, *75*, 200–212. https://doi.org/10.1016/j.apgeog.2016.08.011

Chiang, S. (2014). Housing markets in China and policy implications: co-movement or ripple effect. *China & World Economy*, *22*, 103–120. https://doi.org/10.1111/cwe.12094

Choi, H., & Varian, H. (2012). Predicting the present with Google trends. *Economic Record*, *88*, 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *Journal of Finance*, *66*, 1461–1499. https://doi.org/10.1111/j.1540-6261.2011.01679.x

Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, *48*, 87–92. https://doi.org/10.1145/1096000.1096010

Ginsberg, J., Mohebb, M. H., Patel, R. S., Brammer, L., Smolinsky, M. S., & Brilliant, L. (2009). Detecting influence epidemics using search engine query data. *Nature*, *457*, 1012–1014. https://doi.org/10.1038/nature07634

Glaeser, E., Huang, W., Ma, Y., & Shleifer, A. (2017). A real estate boom with Chinese characteristics. *Journal of Economic Perspectives*, *31*, 93–116. https://doi.org/10.1257/jep.31.1.93

Gong, Y., Hu, J., & Boelhouwer, P. J. (2016). Spatial interrelations of Chinese housing markets: spatial causality, convergence and diffusion. *Regional Science and Urban Economics*, *59*, 103–117. https://doi.org/10.1016/j.regsciurbeco.2016.06.003

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: the case of inflation expectation. *Journal of Economic and Social Measurement*, *36*, 119–167. https://doi.org/10.3233/JEM-2011-0342

Howard, J., & Bowles, M. (2012). The two most important algorithms in predictive modeling today. In *Strata Conference: Santa Clara*.

Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media: an integrated approach of machine-learning algorithms and hedonic modelling to inform equitable housing policies. *Land Use Policy*, *82*, 657–673. https://doi.org/10.1016/j.landusepol.2018.12.030

Hui, E. C. M., & Yue, S. (2006). Housing price bubbles in Hong Kong, Beijing and Shanghai: a comparative study. *Journal of Real Estate Finance and Economics*, *33*, 299–327. https://doi.org/10.1007/s11146-006-0335-2

Jirong, G., Zhu, M., & Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine. *Expert Systems with Applications*, *38*, 3383–3386. https://doi.org/10.1016/j.eswa.2010.08.123

Lee, C., Liang, C., & Liu, Y. (2019). A comparison of the predictive powers of tenure choices between property ownership and renting. *International Journal of Strategic Property Management*, *23*, 130–141. https://doi.org/10.3846/ijspm.2019.7064

Lee, C., Lee, C., & Chiang, S. (2016). Ripple effect and regional house prices dynamics in China. *International Journal of Strategic Property Management*, *20*, 397–408. https://doi.org/10.3846/1648715X.2015.1124148

Lee, K. O., & Mori, M. (2016). Do conspicuous consumers pay higher housing premiums? Spatial and temporal variation in the United States. *Real Estate Economics*, *44*, 726–728. https://doi.org/10.1111/1540-6229.12115

Liu, T., Chang, H., Su, C., & Jiang, X. (2016). China's housing bubble burst? *Economics of Transition*, *24*, 361–389. https://doi.org/10.1111/ecot.12093

Maclennan, D., & O'Sullivan, A. (2012). Housing markets, signals and search. *Journal of Property Research*, *29*, 324–340. https://doi.org/10.1080/09599916.2012.717102

Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review*, *107*, 476–480. https://doi.org/10.1257/aer.p20171084

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*, 87–106. https://doi.org/10.1257/jep.31.2.87

Nardo, M., Petrcco-Giudici, M., & Naltsidis, M. (2015). Walking down Wall Street with a tablet: a survey of stock market predictions using the Web. *Journal of Economic Survey*, *30*, 356–369. https://doi.org/10.1111/joes.12102

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: the case of Fairfax county, Virginia housing data. *Expert Systems with Applications*, *42*, 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040

Piazzesi, M., Schneider, M., & Stroebel, J. (2020). Segmented housing search. *American Economic Review*, *110*, 720–759. https://*doi*.org/10.1257/aer.20141772

Plakandaras, V., Gupta, R. Gogas, P., & Papadimitriou, T. (2015). Forecasting the U.S. real house price index. *Economic Modelling*, *45*, 259–267. https://doi.org/10.1016/j.econmod.2014.10.050

Rae, A. (2015). Online housing search and the geography of submarkets. *Housing Studies*, *30*, 453–472. https://doi.org/10.1080/02673037.2014.974142

Rae, A., & Sener, E. (2016). How website users segment a city: the geography of housing search in London. *Cities*, *52*, 140–147. https://doi.org/10.1016/j.cities.2015.12.002

Ren, Y., Xiong, C., & Yuan, Y. (2012). House price bubbles in China. *China Economic Review*, *23*, 786–800. https://doi.org/10.1016/j.chieco.2012.04.001

Tan, Y., Xu, H., & Hui, E. C. M. (2017). Forecasting property price indices in Hong Kong based on a grey model. *International Journal of Strategic Property Management*, *21*, 256–272. https://doi.org/10.3846/1648715X.2016.1249535

Tsai, I., & Chiang, S. (2019). Exuberance and spillovers in housing markets: evidence from first- and second-tier cities in China. *Regional Science and Urban Economics*, *77*, 75–86. https://doi.org/10.1016/j.regsciurbeco.2019.02.005

Van Dijk, D. W., & Francke, M. K. (2018). Internet search behavior, liquidity and prices in the housing market. *Real Estate Economics*, *46*, 368–403. https://doi.org/10.1111/1540-6229.12187

Van Veldhuizen, S., Vogt, B., & Vogt, B. (2016). Internet searches and transactions on the Dutch housing market. *Applied Economics Letters*, *23*, 1321–1324. https://doi.org/10.1080/13504851.2016.1153785

Varian, H. R. (2014). "Big data": new tricks for econometrics. *Journal of Economic Perspectives*, *28*, 3–28. https://doi.org/10.1257/jep.28.2.3

Weng, Y., & Gong, P. (2017). On price co-movement and volatility spillover effects in China's housing markets. *International Journal of Strategic Property Management*, *21*, 240–255. https://doi.org/10.3846/1648715X.2016.1271369

Wu, J., & Deng, Y. (2015). Intercity information diffusion and price discovery in housing markets: evidence from Google searches. *Journal of Real Estate Finance and Economics*, *50*, 289–306. https://doi.org/10.1007/s11146-014-9493-9

Wu, L., & Brynjolfsson, E. (2015). *The future of prediction: how Google searches foreshadow housing prices and sales* (Working Paper). National Bureau for Economic Research. https://doi.org/10.7208/chicago/9780226206981.003.0003

Zheng, S., Sun, W., & Kahn, M. E. (2016). Investor confidence as a determinant of China's urban housing market dynamics. *Real Estate Economics*, *44*, 814–845. https://doi.org/10.1111/1540-6229.12119