A study on correlation between web search data and housing price Evidence from Beijing and Xi'an in China

Han Shichun, Wang Ping, Wang Bian

School of Management, University of Chinese Academy of Sciences, Haidian District, Beijing, China

Hanshichun@bjgzw.gov.cn, gracewsq@sina.com, wangbian10@mails.ucas.ac.cn

Keywords: Comparative prediction, economic development level, housing price index, web search data.

Abstract. The paper studied the correlation between web search data and housing price, then made a comparative analysis between Beijing and Xi'an in China. An empirical study of Beijing and Xi'an has been conducted to verify the predict ability of web search data, and found that the predict ability of web search data has a certain relationship with economic development level of the region, and web search data has a better explanation ability for the fluctuation of housing price in developed region, like Beijing.

Introduction

Controlling the rapid increase of housing price should be based on accurately grasping its law of fluctuation and mid-and-long term variation trends. Scholars once created related prediction models based on neural network, fuzzy theory, Logistic and so on, simultaneously they got some operative research results. However, these achievements generally place too much emphasis on the rigor of theoretical models, which leads to the relatively poor practical applications of these researches. Web search data has become a new resource to predict the housing price(Lynn Wu,2009). So this paper tries to establish two search indexes for predicting both Beijing and Xi'an housing price by using web search data.

The rest of the paper proceeds as follows: section 2 reviews studies on issues of housing price prediction and researches based on search data prediction; section 3 presents the theoretical analysis framework; section 4 describes the data source and the method of keywords selection; section 5 establishes two predictable models for fluctuations of Beijing and Xi'an housing price by using web search data, then tests and compares them; section 6 concludes.

Literature Review

Many scholars have paid great attention to determinants of housing price in recent years. But their works' shortcomings cannot be neglected: Firstly, the prediction accuracy is not so good; Secondly, the forecasting effect will not be good when housing price is experiencing adverse changes or severe shocks(Yang Xin, 2011; Na Li, 2012; Y Liu, 2012). The application of search engine is capturing more and more researchers' attention, such as in the field of health, society and economy.

In the area of health, some studies show that the search data can help to detect public health trends and syndrome surveillance (Doornik, J.A., 2010; Hulth, A., Rydevik, G., Linde, A., 2009). One of most influential and fundamental study is conducted by Ginsberg(2009). They use Google search data to detect influenza epidemics, and their new method can improve the timeliness compared to the influenza-like symptoms statistical data released by Center for Disease Control and Prevention of the USA. The model was also used to Google Flu Trend.In the area of society, researchers focus on the correlation between of web search data and public attention. Jurgen A. Doornik (2009) carefully studied the prediction ability of web search data, and extended the simple linear model based on web search data to a time series model with calendar variables, the new model combined empirical data with historical data, and further improved the prediction accuracy. In the area of economics,

researchers use web search data to predict consumption and unemployment. Hyunyoung Choi and Hal Varian(2009) conducted a fundamental study, they made some empirical tests on the web search data in predication of sales of retail, auto, home and travel in the United States, and they found that the prediction accuracy of all the four industries were significantly improved when the keyword search volumes were added to the traditional self-regression model as new factors. Lynn Wu and Erik Brynjolfsson(2009) found that search data had a strong predictive power on the housing transaction volume and prices through their empirical study about the real estate market in U.S..

Given recent researches on prediction based on web search, some achievements have been made in the field of certain social and economic behavior, including the fluctuations in real estate price index. However, empirical comparative studies on the real estate price index have not been found in China. It remains questionable whether it would be different to use web search data predicting real estate price fluctuations in different regions of China.

Theoretical Analysis and Conceptual Framework

Theoretical analysis: factors influencing real estate market and transmission delay

Participants of real estate market including purchasers of houses and real estate developers would usually have an expecting process for macro policy changes. When macro policy changes, firstly, they would analyze its economic prospects, capital cost, opportunity cost and expected benefit; Then, they would make an adjustment to their investment and purchasing behavior. So, changes of housing price would not happen immediately after the change of macroeconomic policies. A break would happen before the situation completely changes, which is to say macroeconomic policies usually have a delay. Take the lagged impact of interest rate policy on housing price for example. During a short period, the improvement of interest rates will make real estate developers transfer the cost to housing price immediately; Real estate developers would increasingly face much pressure of more inward flood of capita, and house purchasers have more and more sensitively response to higher rates. Finally, housing price would fall, and regulatory effect of interest rate policy would gradually emerge.

Changes of supply and demand have a lagged effect on housing sales prices according to price stickiness theory. Housing markets have many features, such as poor mobility and scarcity, a fixed location, large value, non-homogeneity of goods, which determine that the real estate market is in a state of information asymmetry. House purchasers can hardly judge the quality merits of houses by their outward appearance. They usually know little about the cost, as well as price and community service of similar properties. As a result, they usually spend much time gathering related information. So we can say the level of housing price does not closely follow changes of demand, besides, housing price have a lag phase.

From above theoretical analysis, we can find that macroeconomic and supply-demand relationship are both key factors affecting housing price, besides, the influence has a certain delay. Corresponding keywords of these factors can be found by using Network users' search data. Macroeconomic situation and supply-demand relationship will bring some changes in search amounts of online keywords, moreover, this kind of changes is instant. So we conclude that there is some relationship between certain web search data and housing price level, and there is an advanced-lagged relationship between them too. Further, we believe that if we choose relevant keywords and then establish a prediction model, we can get and judge the trend of housing price index.

Conceptual framework: the web search data and housing sales price index

In this section, concerning factors influencing house price fluctuation, we build a conceptual framework to illustrate correlation between web search data and house sales price index. It is shown in Figure 1.

On the one hand, macroeconomic variables including economic, political and social factors would influence the real estate developers' supply and purchasers' demand, and it is just this supply-demand level determines the housing price.

On the other hand, based on the theory of consumer behavior, we divide purchasers' behavior into consumer demand, information search and purchase decision three links; similarly real estate developers' behavior is divided into investment demand, information search and investment decision three links. It is well known that both the decision-making process of purchasers and developers requires a lot of information. Nowadays, various kinds of web platforms including search engines, blogs, as well as BBS have been becoming more and more popular, these applications have become people's best choice to collect information. At the same time, these web platforms could record searchers' behavior and provide data source for our study.

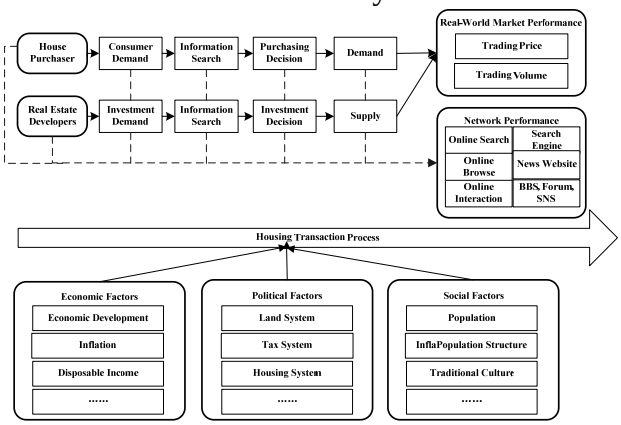

Figure 1: The conceptual framework.

In the real estate market, there is always a time lag between macroeconomic factors and housing price, while information recorded on internet tools could reflect supply-demand relation more timely, we assume that there would be a chronological order between web search data and house price index. In the next chapter, we would explore the predict ability of relative search keywords' volume on the trend of housing sales price index.

Given the previous study, the information searched by real estate and purchasers could be divided into two main parts: micro-level information and macro-level information.

(1)Micro-level information. These information reflect consumers' purchase willingness or investment intention, including the physical properties of housing, and construction materials.

(2)Macro-level information. These information reflect the influence of macro policy and economic situation on market players' expectation, including national macroeconomic policies, economic trends and market conditions, etc.

Figure 2 shows the curve of searching amount of keywords, "website of housing price" with housing sales price index (HSPI), from August 2007 to December 2012. It shows strong concordance between the tendency of housing sales price index (HSPI) and the tendency of search terms' searching volume.

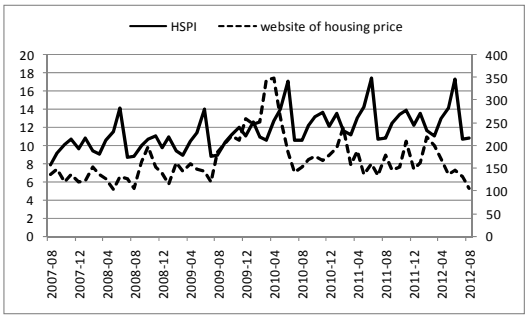

Figure 2: The search data compared with the HSPI curve.

Since different search terms represent different meanings, and show different correlation with our housing price indicator, we should develop scientific method to select the objective, reasonable and comprehensive keywords, in order to ensure our model could predict and housing sales price index scientifically.

Empirical Studies

Data description and pre-processing

In this study, the monthly housing sales price index data of the two regions (Beijing and Xi'an) is from August 2007 to December 2012 as reported by Wind Database (a total of 65 monthly year-on-year data). In the later sections, 59 monthly data will be used for modeling for each region, while the rest be used to examine prediction results. The web search data is from Google Trends (http://www.google.com/trends/), which can provide us query data as CSV files, and search volume of keywords from January 2004 till current week. Search volume downloaded from Google Trends is not absolute search volume, but rather a relative data.

Keywords selection

In this paper, we select related keywords for Beijing and Xi'an following these rules:

Firstly, set initial keywords. We choose nine same initial keywords for these two regions, including housing price, housing price network, real estate, real estate information, apartment, rent an apartment, opening quotation, construction materials, second-hand housing, etc. Secondly, build keyword database. As for Beijing, 125 new different keywords are recommended by public search engine, and 125 new different keywords are recommended for Xi'an province. These keywords would form Beijing and Xi'an keyword database respectively; Thirdly, calculate the Pearson correlation coefficient. For fitting the index better, we calculate six times correlation coefficient for each keyword from Beijing and Xi'an keywords database, including the correlation coefficient between 0-5 month lagged web search data and the index. Finally, we pick out 10 keywords from Beijing and Xi'an respectively with greater correlation coefficients, which is shown in Table 1.

Table 1: Selected keywords(translated to English)*.

Keywords(BJ)	r	Keywords(XA)	r
Website of housing price	0.70	Soufang website	0.60
Housing price	0.68	Housing price	0.58
Housing price trend	0.66	Real estate	0.55
New real estate	0.64	Xi'an housing price	0.54
Sell house	0.61	Xi'an second-hand housing	0.53
Housing price Beijing	0.59	Xi'an house network	0.51
Beijing housing price	0.57	second-hand housing	0.50
Buy house	0.52	Villa	0.50
price-fixed housing	0.51	House rental network	0.48
Housing price of Beijing	0.50	New house	0.48

* *Please see appendix 1 for Chinese search query.*

Synthesis of the search index and the determination of search lead

Based on the selected nine search terms in 4.2, we add together these nine terms' search volume in the same month, as the search index. Besides, in order to remain same magnitude order of housing price index and search index, the housing price index is multiplied by 100 to get the final housing price index. The correlation coefficient between housing sales price index and 0-7 month lagged search index for Beijing and Xi'an are both shown in Tab 2. It is can be found that, as for Beijing, the biggest correlation coefficient appears when search index leads 2 months, while for Xi'an, the biggest correlation coefficient appears when search index leads 1 months.

Therefore, we assume, purchasers would more probably search information 2 months earlier than housing price index in Beijing, while in Xi'an, that is about 1 months earlier.

Table 2: Correlation coefficients corresponding to different delay of search index.

Search lead(month)	0	1	2	3
Correlation(BJ)	0.67	0.61	0.71	0.64
Correlation(XA)	0.52	0.67	0.54	0.33
Search lead(month)	4	5	6	7
Correlation(BJ)	0.66	0.66	0.56	0.66
Correlation(XA)	0.49	0.31	0.33	0.46

Forecasting Model for Real Estate Price Index Base on Web Search Data

Model construction

Denote housing sales price index in the tth month as {yt : t=1,2,⋯,T}(the housing sales price index of every period) and the web search data in the kth month as {qt : t=1,2, ⋯,T }. In order to reduce rounding errors, we use their natural logarithm in the model, that is to say we regard lnyt as explained variable, and lnqt as explanatory variable. Moreover, in view of the fact that housing sales price index is year-on-year data, and often affected by history data, lnBJyt-k and lnXAyt-k are also brought into the model. In order to compare the prediction results of the two regions, as well as the effect of web search data, we establish four models shown below.

$$\ln \mathrm{BJ}y_t = \beta_0 + \beta_1 \ln(\mathrm{BJ}y_{t-\mathrm{k}}) + \mu_t \quad (1)$$

$$\ln \mathrm{BJ}y_t = \beta_0 + \beta_1 \ln(\mathrm{BJ}y_{t-\mathrm{k}}) + \beta_2 \ln(\mathrm{BJ}q_{t\text{-}2}) + \mu_t \quad (2)$$

$$\ln \mathrm{XA}y_t = \beta_0 + \beta_1 \ln(\mathrm{XA}y_{t\text{-k}}) + \mu_t \quad (3)$$

$$\ln \mathrm{XA}y_t = \beta_0 + \beta_1 \ln(\mathrm{XA}y_{t\text{-k}}) + \beta_2 \ln(\mathrm{XA}q_{t-1}) + \mu_t \quad (4)$$

The least squares estimates for this model are shown in Table 3.

$$\ln \mathrm{BJ}y_t = 0.490 + 0.516 \ln(\mathrm{BJ}y_{t-7}) + 0.102 \ln(\mathrm{BJ}q_{t\text{-}2}) \quad (5)$$

$$\ln \mathrm{XA}y_t = 1.717\text{-}0.263 \ln(\mathrm{XA}y_{t-4}) + 0.145 \ln(\mathrm{XA}q_{t\text{-}1}) \quad (6)$$

Table 3 indicates that the coefficients before every explanatory variables are significant and the models fit well. Comparing Model (1) with Model (3), the goodness of fitness has been improved in both (2) and (4). The coefficient on the web search variable in (5) implies that 1% increase in search volume is associated with roughly a 10.15% increase in housing sales price index, in other words, the price index will be higher if the attention paid to search index increases. The positive coefficient before lnyt-7 is less than one, and it reflects the existence of price stickiness in the real estate market. So as to Xi'an, 1% increase in search volume is associated with roughly a 14.46% increase in housing sales price index, and the negative number before lnyt-4 is consistent with the fact that the price index is year-on-year. And we can also find that the prediction result of Beijing is better than results of Xi'an.

Table 3: The models' regression results.

Explanatory variable	Model 1		Model 2		
	C_1	$y_{t\text{-}7}$	C_1	$y_{t\text{-}7}$	$q_{t\text{-}1}$
Coefficient	0.75	0.49	0.49	0.52	0.10
P-value	0.03	0.01	0.02	0.00	0.00
R^2	0.74		0.85		
Log-likelihood	48.21		66.69		
AIC	-1.32		-1.71		
SC	-1.17		-1.56		
MAPE	0.23		0.13		
TIC	0.06		0.03		
Explanatory variable	Model 3		Model 4		
	C_2	$y_{t\text{-}4}$	C_2	$y_{t\text{-}4}$	$q_{t\text{-}1}$
T-statistic	1.69	-0.21	1.72	-0.26	0.14
P-value	0.02	0.01	0.00	0.02	0.00
R^2	0.72		0.79		
Log-likelihood	38.75		47.55		
AIC	-1.36		-1.50		
SC	-1.18		-1.36		
MAPE	0.22		0.16		
TIC	0.07		0.04		

Model checking and forecasting

During the former process, we also exclude the search index and fit two new seasonal AR models respectively for each region to test whether the web search data can decrease the prediction error. Then we use the four models to forecast the price index from July 2007 to December 2012. Table 5 reports their forecasting comparison results. It can be clearly seen that the model containing search index is of low absolute deviation and is more ideal and acceptable, and the prediction results of Beijing based on web search data is much better than that of Xi'an.

Table 4: The models' prediction error.

Date		12/07	12/08	12/09	12/10
BJ	Model 1	5.08%	5.08%	7.89%	7.66%
	Model 2	5.69%	5.42%	0.10%	3.15%
XA	Model 3	3.77%	3.98%	7.66%	6.98%
	Model 4	4.91%	3.99%	4.69%	4.26%
Date		12/11	12/12	Average	

BJ	Model 1	-2.44%	9.98%	5.54%
	Model 2	5.28%	1.33%	3.50%
XA	Model 3	5.31%	3.18%	5.15%
	Model 4	5.03%	4.26%	4.52%

There are two explanatory variables in the Model (2) and (4), including seven-month-lagged and four-month-lagged housing sales price index, so we can predict the housing sales price index of the next month when we get the data of current, ahead of the State Bureau of Statistics publishing the data. Based on the above analysis, we affirm that this method can overcome the delay of traditional monitoring methods, and can be used for indirect and timely monitoring of housing sales price.

Conclusion

This paper not only explores the correlation between web search data and housing sales price index, but also compares the predict ability of web search data between Beijing and Xi'an of China.

Firstly, a theoretical framework illustrating the relationship between web search data and housing price index has been established.Secondly, empirical studies of Beijing and Xi'an has been conducted to test the predict ability web search data. From both of two AR models with web search data, it can be found both coefficients of web search data are significant. In addition, comparing the prediction results of models containing web search variable with models without web search variable, we find model with web search data can reduce the percentage error of prediction, confirming that the web search data can improve prediction result.Finally, by comparing Beijing and Xi'an predict results, we find that the predict ability has a certain relationship with economic development level of the region. That is also easy to understand, cause the more developed the economy, the more possibility people have access to advance technology, and the more users could surf the internet to get information, as well as the more accurate web search data reflect real estate developers' and purchasers' attention.

Acknowledgements

Thank the support of the National Natural Science Foundation of China under Grant 71202115、71172199 and 71203218, Foundation of Dean of Graduate University of Chinese Academy of Sciences under Grant Y15101QY00, and Postdoctoral Science Foundation under Grant 2011M500434 for this study.

References

[1] Yang Xin, Peng Geng, Yuan Qinyu, Lv Benfu. A prediction study on tourist amount based on web search data a case from Hainan[C], *Proceedings 2011 International Conference on Business Management and Electronic Information.*

[2] Y Liu, B Lv, G Peng, Q Yuan, 2012. A Preprocessing Method of Internet Search Data for Prediction Improvement[C], KDD conference, Workshop on Data Mining and Intelligent Knowledge Management.

[3] Na Li, Geng Peng, Hang Chen, 2012. A Prediction Study on the Daily Number of E-commerce Orders Based on Site Search Data[C], Proceedings of 2012 IEEE 5th International Conference on Management Engineering & Technology of Statistics.

[4] Doornik J.A., 2010. Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data. 8th OxMetrics User Conference.

[5] Hulth, A. , 2009, Rydevik, G., Linde, A.: Web Queries as a Source for Syndromic Surveillance. PLoS ONE.

[6] Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, L. Brilliant, 2009. Detecting influenza epidemics using search engine Query data[J]. Nature.

[7] Varian, H., Choi, H., 2009: Predicting the Present with Google Trends. Google Research Blog; Available at SSRN, http://ssrn.com/abstract=1659302

[8] Lynn Wu，Erik Brynjolfsson, 2009. The Future of Prediction—how google searches foreshadow housing price and sales, Working paper.