Housing Price Prediction Using Search Engine Query Data

Qian Dong*
National Bureau of Statistics of China , Beijing, P.R.China– queenad@hotmail.com

Nana Sun
National Bureau of Statistics of China , Beijing, P.R.China –sunnana_4108@163.com

Wei Li
National Bureau of Statistics of China , Beijing, P.R.China- liwei@gj.stats.cn

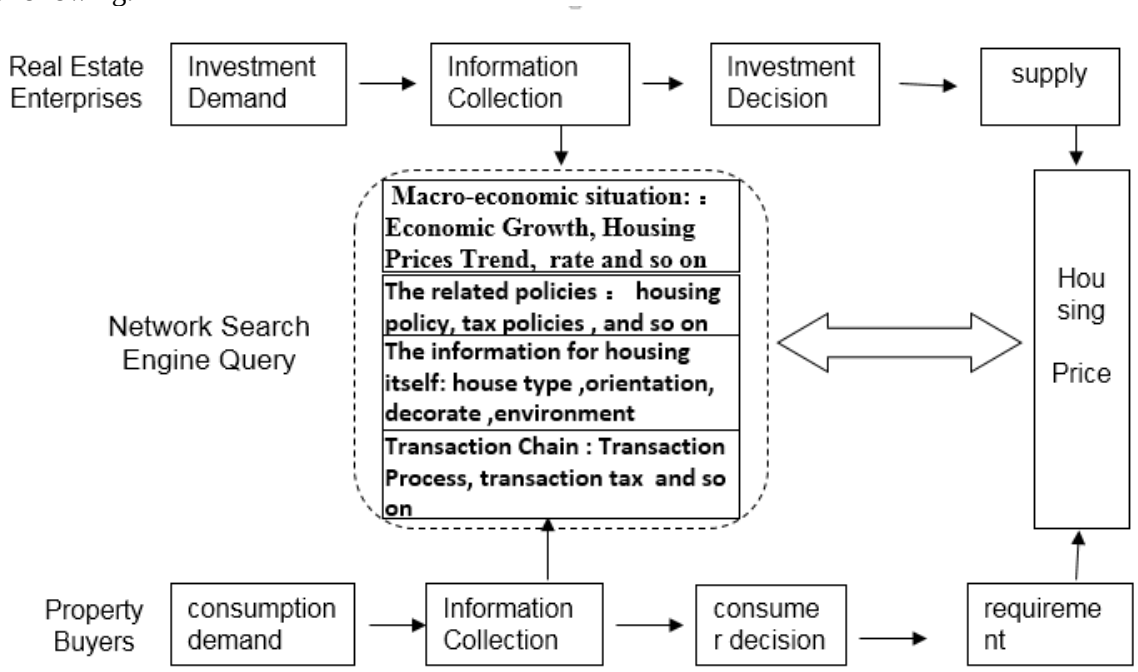
Abstract

The real estate industry is one of the economics drivers of the Chinese economy, and the housing price has been earning constant attention ever since. But the data published by government statistical

early than the official data. At same time, the prediction data can also be used as a useful supplementary and reference for the traditional housing price index.

2. Analysis of Theoretical Framework

The housing price indices are depend on supply and requirement, there are the theoretical framework as following:



3. Data Description

(1)Research Objects

Using Baidu search engine query data to predict the housing price, we should consider that people collection the real estate information may be more through advertising, friends and real estate agency at small cities or less developed, they are searching through the network for real estate information are relatively small group. Thus, we decide to choose the larger scale, a relatively developed, real estate transaction relatively active 6 cities as our research objects:

- **First-tier cities** ：Beijing, Shanghai, Guangzhou.
- **Second-tier cities** ：Nanjing, Xian, Shangyang.

(2)Variables Description

The model is usually include dependent variable and independent variables. The dependent and independent variables are summarized as following:

Dependent Variables： The New Housing Price Index and Second-hand Housing Price Index of 6 cities. Using the same month last year of data from Jan.,2012 to July, 2014, a total of 31 months data.

Independent Variables： According to the factors affecting the housing price, to determine the 15 initial keywords; then, using the key words that automatic recommendation from Baidu search engine, obtain the key word database; thus, calculated the correlation coefficient for each key word and housing price index to do key words screening. After repeated comparisons and selection, Key words has been chosen as following: **Second hands housing price:** Prices trend, House source, Decoration, Real Estate Network, Public reserve funds, Mortgage interest rates, House duty, Housing rental, Real estate agency, Second hands house, Second hands housing transaction process, Second hand housing transaction taxes and fees. **New housing price:** Prices trend, House source, Decoration, Real Estate Network, Public reserve funds, Mortgage interest rates, New estate , Low-income housing.

4. The Housing Price Prediction Model

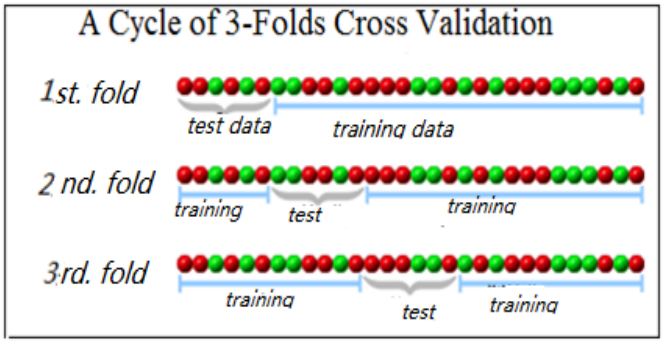
(1) Background Models

The Cross-Validation Technique

Linear Regression Model	**Mixture Linear Regression Model**	**Regression Tree Model**	**Random Forests Model**
Bagging Model	**m-Boosting Model**	**Support Vector Machine**	**Neural Network Model**

(2)The Construction of Prediction Model

With the 3-folds cross-validation technique, we fitted our prediction model by using 8 analytical models including Linear Regression, Regression Tree, Random Forests, Support Vector Machine (SVM) and so on, then compared with the predicted results of 8 models. A cycle of 3-folds cross validation shows as following:



5. Housing Price Prediction Based on Search Engine Query Data

(1) The Prediction for Second hands Housing Price Index

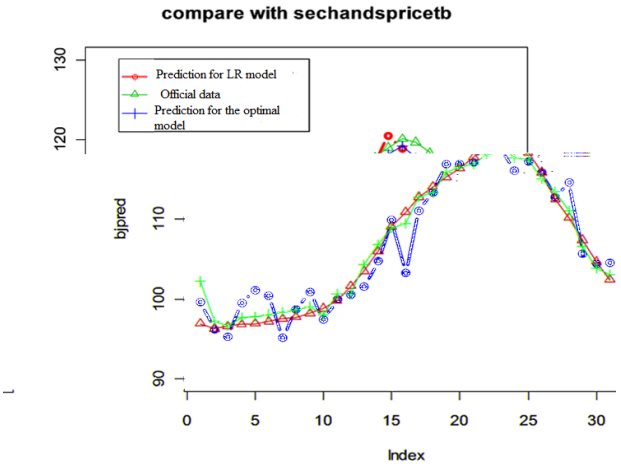
Effect of Main Keywords Search Index in Second-hands Housing prices for 6 Cities as following:

Cites	**Main Key-words Searching Indices**
Beijing	Prices trend, House source, Decoration, Public reserve funds, Second hands housing transaction process, Housing rental
Shanghai	Prices trend, House source, Decoration, Mortgage interest rates, Second hands housing transaction process, Second hands housing transaction taxes and fees, Real estate agency, Housing rental
Guangzhou	Decoration, Real Estate Network, Public reserve funds, Second hands housing transaction process, Housing rental
Nanjing	Decoration, Real Estate Network, Public reserve funds, Mortgage interest rates, Second hands house, House duty, Housing rental
Shenyang	Prices trend, Decoration, Public reserve funds, Mortgage interest rates, Second hands housing transaction taxes and fees, Second hands house, House duty
Xian	Prices trend, Decoration, Real Estate Network, Public reserve funds, Second hands housing transaction process, House duty, Housing rental

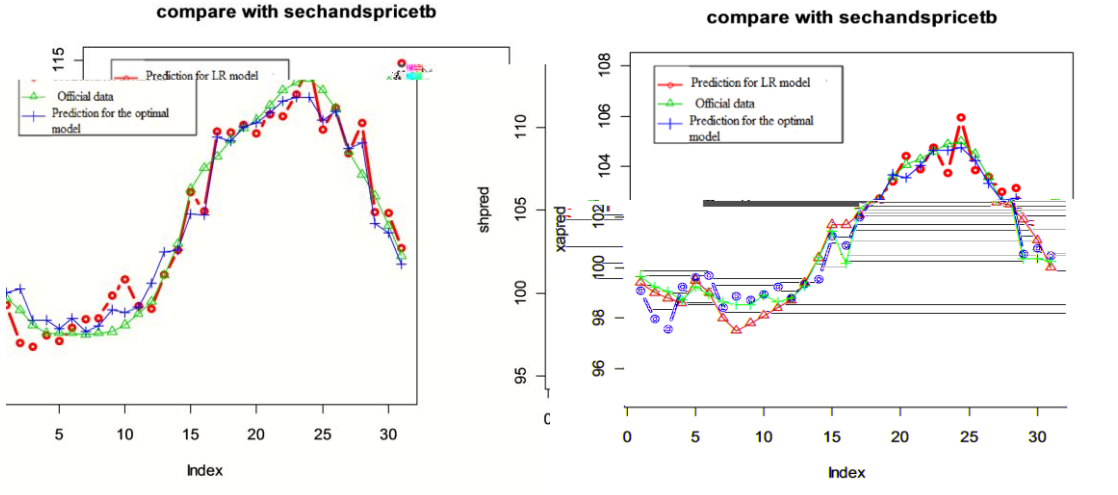
The optimal prediction models for second-hand housing prices of 6 cities shows below:

Order	Cities	Fit the optimal model	Stability of the optimal model
1	Beijing	Random Forests	Random Forests
2	Shanghai	SVM	SVM
3	Guangzhou	SVM	SVM
4	Nanjing	SVM	SVM
5	Shenyang	SVM	SVM
6	Xian	SVM	SVM

There is effect drawing for Second Hands Housing Price of the Beijing Prediction Model shows below:



Effect Drawing of Second Hands Housing Prices of the Shanghai &Xian Prediction Models as following:



(2) The Prediction of New Housing Price Index

Effect of Main Keywords Search Index in New Housing prices for 6 Cities as following:

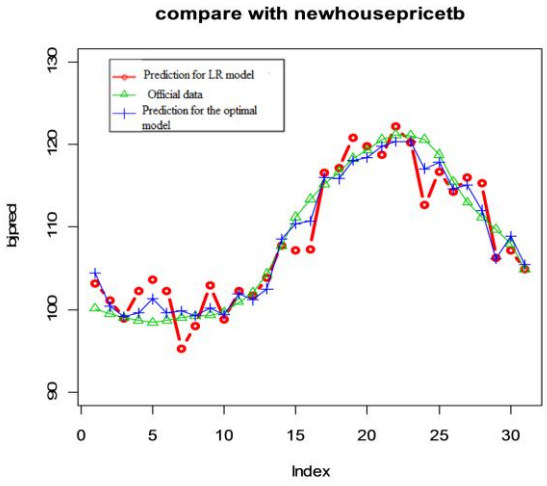
Cites	Main Key-words Searching Indices
Beijing	Prices trend, House source, Decoration
Shanghai	House source, Decoration, Low-income housing
Guangzhou	Decoration, Public reserve funds, Mortgage interest rates, Low-income housing

Nanjing	Prices trend, Real Estate Network, Public reserve funds, Mortgage interest rates
Shenyang	Prices trend, Decoration, Public reserve funds
Xian	Decoration, Real Estate Network, Public reserve funds, Mortgage interest rates

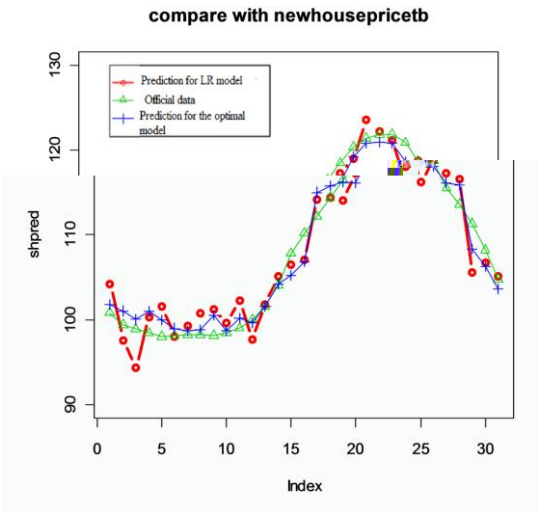
The optimal prediction models for second hand housing prices of 6 cities show below:

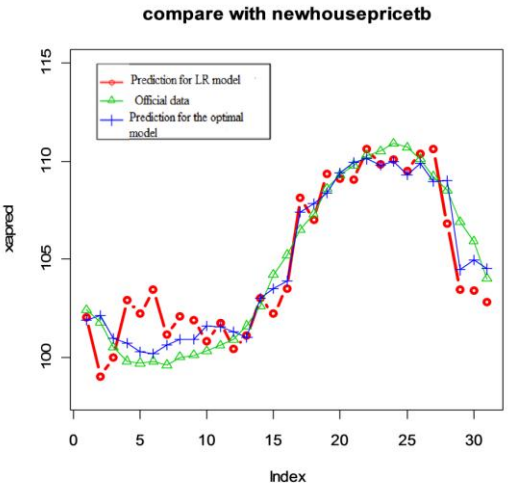
Order	Cities	Fit the optimal model	Stability of the optimal model
1	Beijing	Random Forests	Random Forests
2	Shanghai	SVM	SVM
3	Guangzhou	Random Forests	Random Forests
4	Nanjing	SVM	SVM
5	Shenyang	SVM	SVM
6	Xian	Random Forests	Random Forests

There is effect drawing for New Housing Price of the Beijing Prediction Model shows below:



Effect Drawing of New Housing Price of the Shanghai &Xian Prediction Models as following:





6. Conclusion

Based on *Baidu Search Index*, using the cross validation technique and 8 models were successfully fitted and predicted for new houses and second-hand housing price index in 6 cities, and the prediction of *NMSE* and *MSE* are reached 0.0232. Since the search engine query data can be obtained in real time, can take immediate influence factors for price changes into the prediction model, we can get the last month of new and second-hand housing price index at the beginning of every month, issued about two weeks early than the official data, make up for the traditional statistics information release lag issues.

In this paper, there are three Innovation: first of all, using *Baidu* search engine query data to predict the housing price, this types of domestic researches is rarely. Using search engine query data to predict is not only has good prediction effect, and compared with the traditional survey data, it has strong timeliness. Second**,** using the cross validation technique and 8 analytical models, and they were successfully fitted and predicted for new houses and second-hand housing price in 6 cities. Overall, the predicting trend of linear regression model and optimal model are basically same with the official data, but values of the optimal prediction model are closer with the actual value. Third, since we only have a small amount of data, In order to compensate for deviation of the small data, using 3-folds cross validation technique, ensure the accuracy and reliability of the final prediction result.

For the future works, this Idea and method can be extended to the monthly data indices such as CPI, Household Income Index, and Household Consumption Expenditure Index etc. According to the accumulation of search engine query data, the prediction value for Indices will be more accuracy.

References

[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. Statistical modeling: The two cultures [J]. Statistical Science, 2001, 16:199-215.

[4] Breiman, L., J. H. Friedman, R. A. Olshen, et al. Classification and Regression Trees [M]. Chapman and Hall, New York, 1984.

[5] Cho H I, Varian H. Predicting the Present with Google Trends[C]. Technical Report，2009，Google Inc·

[6] Iverson, L. R., A. M. Prasad, S. N. Matthews, et al. 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] Kulkarni1 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 Indicatorsvs. 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.