# Housing Price Prediction Using Search Engine Query Data

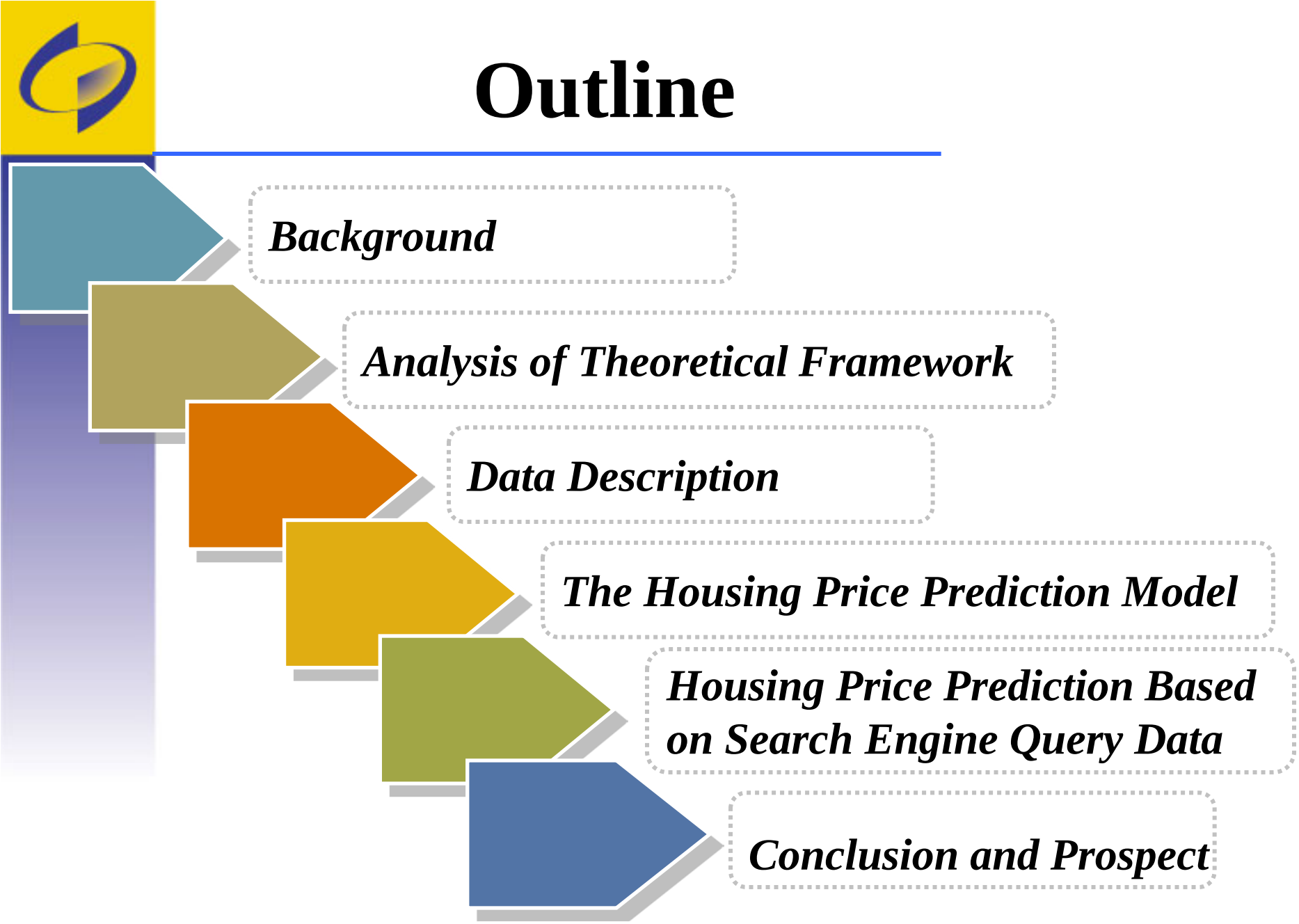
*Qian Dong*

*Research Institute of Statistical Sciences*

*of NBS*

**Oct. 29, 2014**

# Outline


Background
Analysis of Theoretical Framework
Data Description
The Housing Price Prediction Model
Housing Price Prediction Based on Search Engine Query Data
Conclusion and Prospect

# Chapter 1. Background

*Research Background*

*Overseas and Domestic of Research Status*

*Research Ideas*

# Background

- The age of big data is coming……
- Great opportunities and challenges to the government statistics
- The National Bureau of Statistics of China has started the cooperation with enterprises for the pilot research on big data.
- The real estate industry is one of the economics drivers of the Chinese economy
- Housing price is always a focus to people
- But the housing price index published by government statistical agencies are usually release at middle of each month, thus cannot to fulfill the public demand.

# Domestic and Overseas of Research Status

- The prediction using **Search Engine Query Data** in business and academia has a lot of exploration and research such as
- Baidu ( Baidu Online Network Technology Co.)& Chinese Academy of Sciences:

  ***Consumer Confidence Index***
- Baidu Prediction: Baidu 2014 FIFA World Cup Prediction, College Entrance Examination Prediction and etc.

# Domestic and Overseas of Research Status

- The Research that using search engine query data to predict price index: only few papers;
- The research paper for price tendency prediction of real estate market is more less.
- *Wu L. and others(2014)*

  ***The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales***

- *Rajendra Kulkarni and others(2009)*

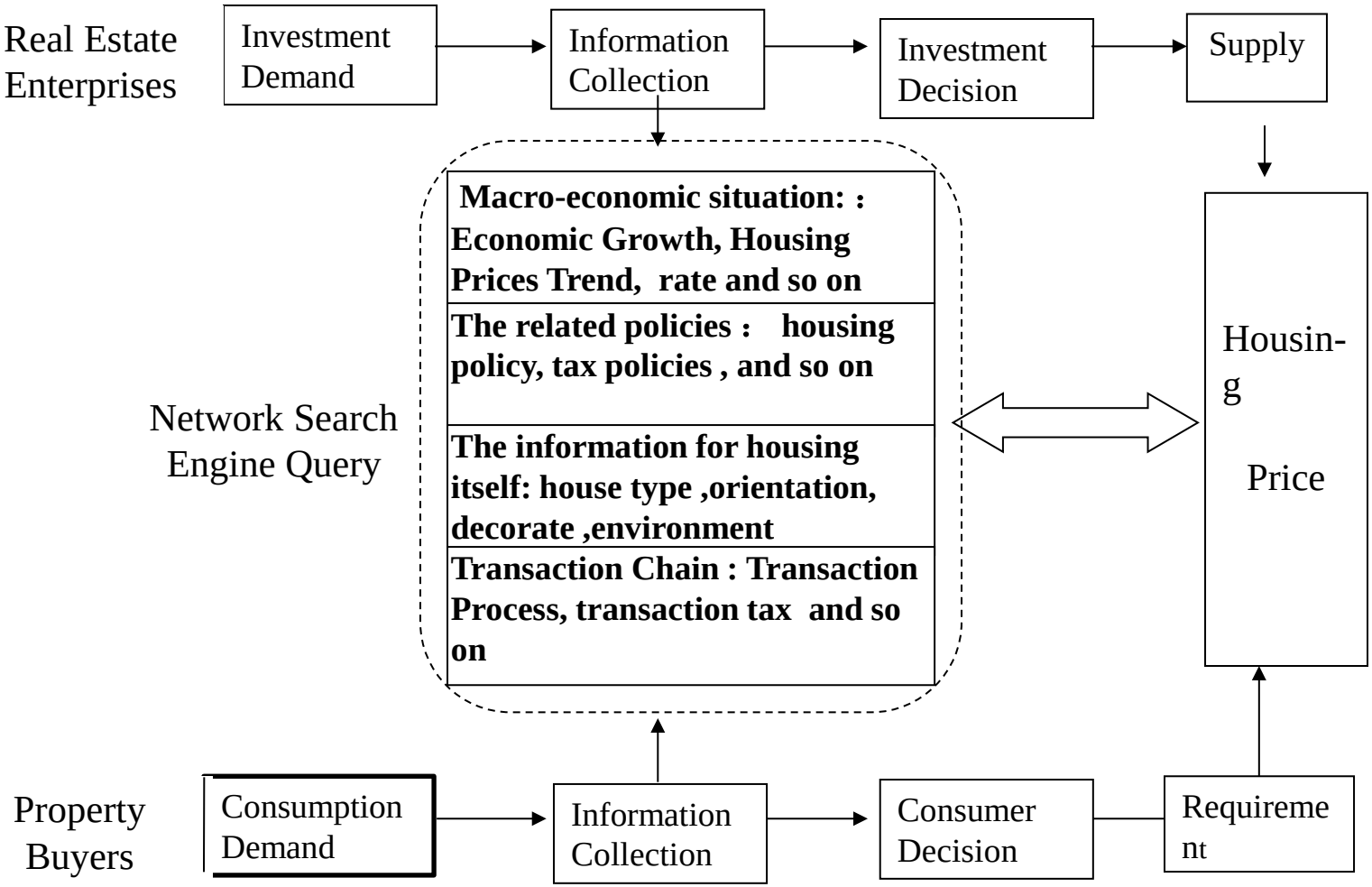
  ***Forecasting Housing Prices with Google Econometrics: A Demand Oriented Approach***

***……***

# Researching Ideas

- In order to solve the problem of timeliness of the Housing Price Index,
- To predict the new housing price index and second-hands housing price index for major cities in China by using Baidu Search Index (BSI)
- Because the search engine query data can be obtained in real time,
- immediate influence factors for price changes into the prediction model
- new housing price index and second-hands housing price index at the beginning of each month
- two weeks early than the official data
- at the same time the prediction data can also be used as a useful supplement and reference **FOR** the traditional housing price index

# Chapter 2. Analysis of Theoretical Framework



# Chapter 3. Data Description

*Research Objects*

*Variables Description*

# Research Objects

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

| | |
|---|---|
| ***First-tier Cities*** | *Beijing, Shanghai, Guangzhou.* |
| ***Second-tier Cities*** | *Nanjing, Xian, Shangyang.* |

# Variables Description

**Dependent Variables**

New Housing Price Index and Second-hands Housing Price Index for 6 cities. Using the same month last year of data from Jan. 2012 to July 2014, a total number of data is 31 months .

# Variables Description

## Independent Variables

*According to the Impact factors for housing price, to determine the 15 initial keywords; then, using the keywords that automatic recommendation from Baidu search engine, obtain the keywords database; thus, calculated the correlation coefficient for each key words and housing price index to do keywords screening. After repeated comparisons and selection, keywords has been choosen 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 hands 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

# Chapter 4. The Housing Price Prediction Model

*Background Models*

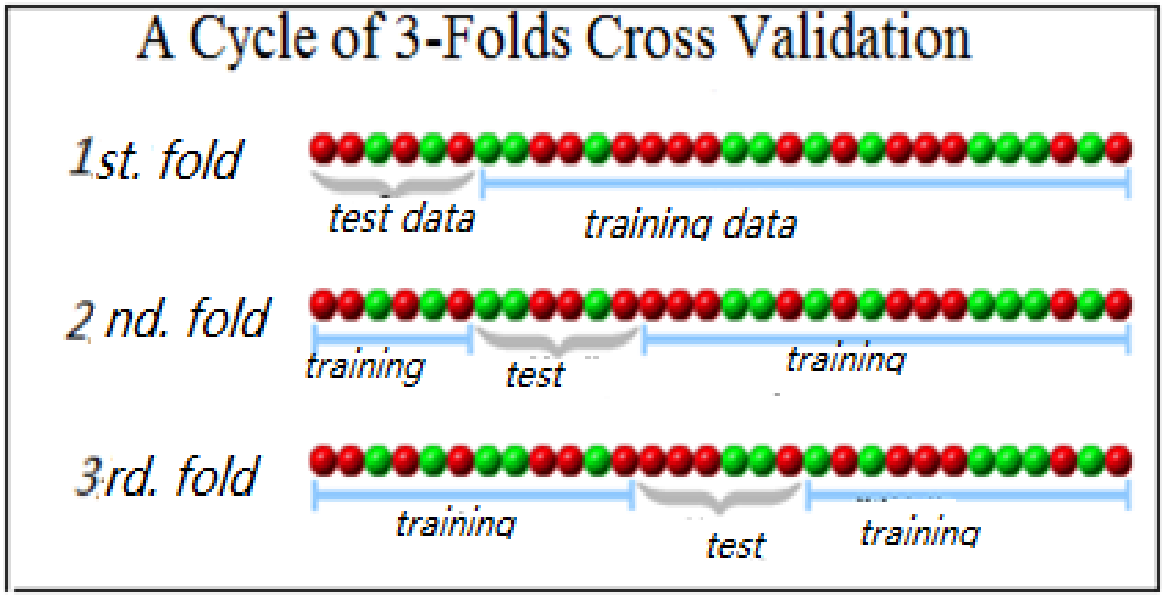
*The Construction for Prediction Model*

# Background Models

| The Cross-Validation Technique | | | |
|---|---|---|---|
| Linear Regression Model | Regression Tree Model | Bagging Model | Neural Network Model |
| Mixture Linear Regression Model | Random Forests Model | m-Boosting | Support Vector Machine |

# 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 for 8 models. A cycle of 3-folds cross validation shows as following:



# Chapter 5. Housing Price Prediction Based on Search Engine Query Data

***The Prediction for Second Hands Housing Price Index***

***The Prediction for New Housing Price Index***

# Main Keywords Search Indices for Second-Hands Housing prices at 6 Cities

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

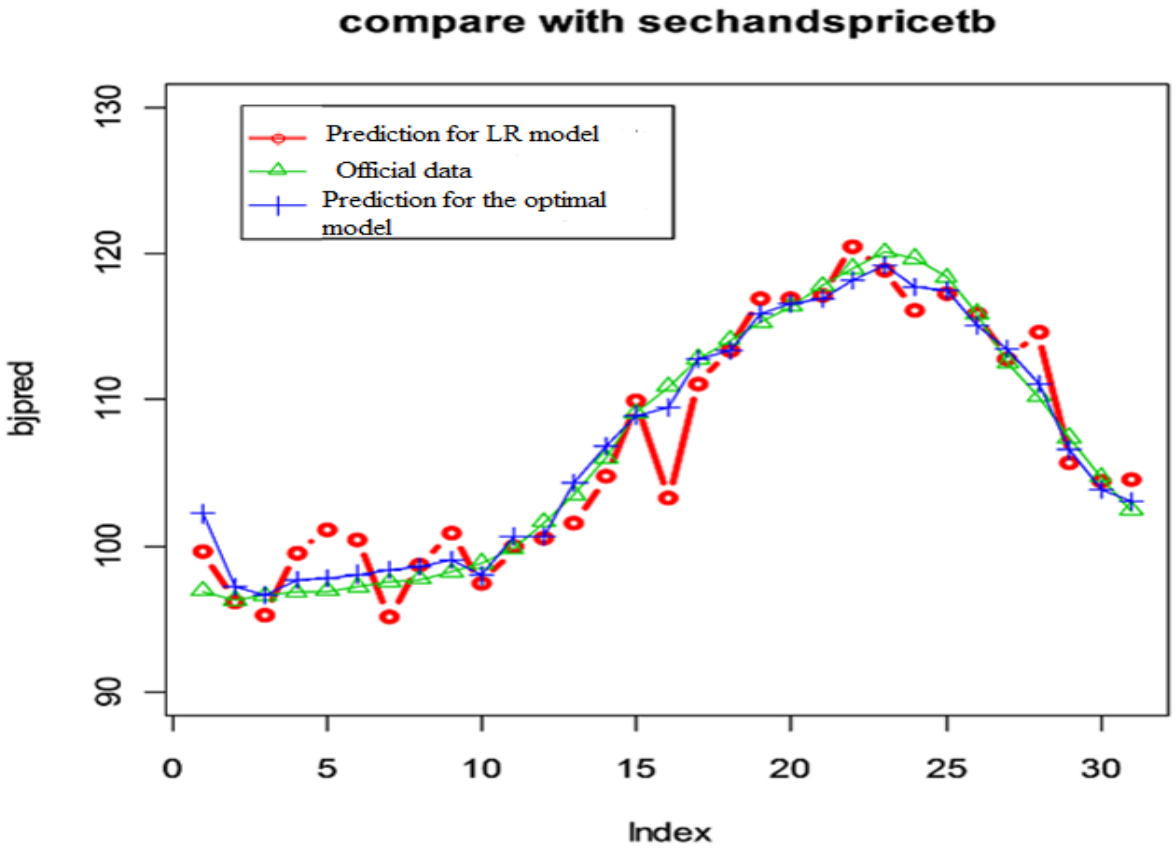
# The Prediction for Second Hands Housing Price Index

The optimal prediction model for second-hands housing prices at 6 cities

| 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 |

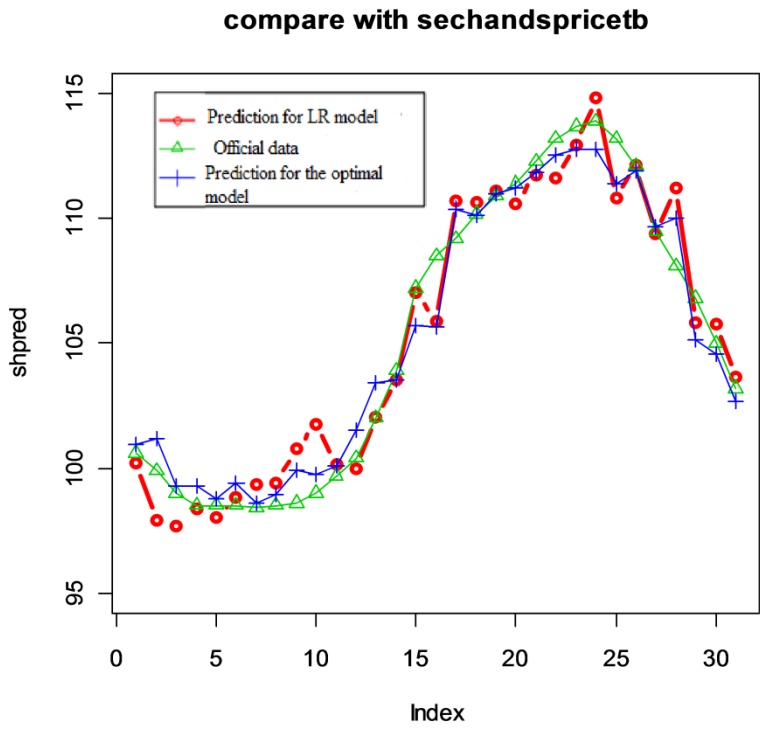
# The Prediction for Second Hands Housing Price Index

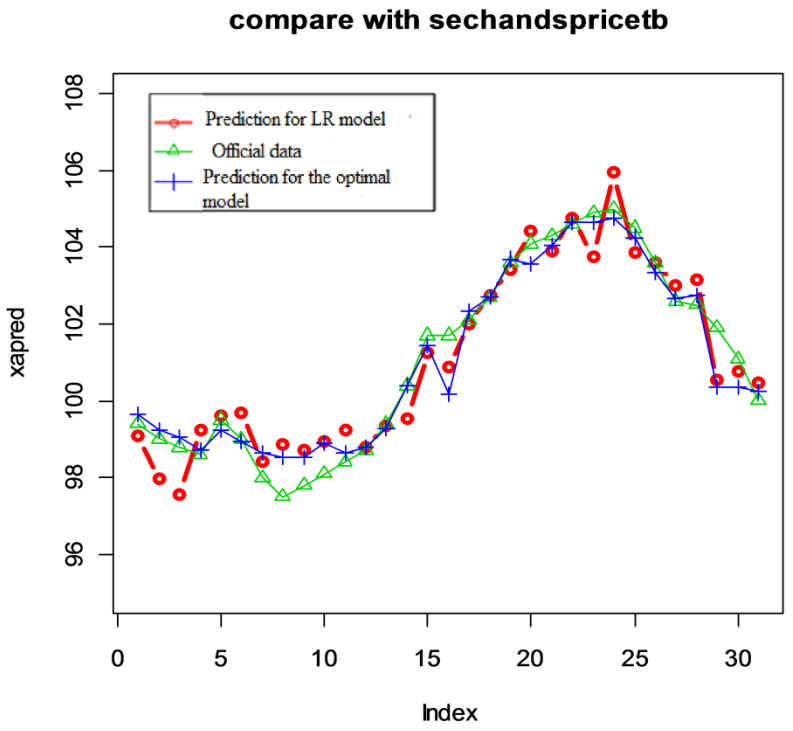
Figure for the Prediction Model of Second Hands Housing Price at Beijing



# The Prediction for Second hands Housing Price Index

Figures for the Prediction Model of Second Hands Housing Price at Shanghai &Xian



# Main Keywords Search Indices for New Housing Prices at 6 Cities

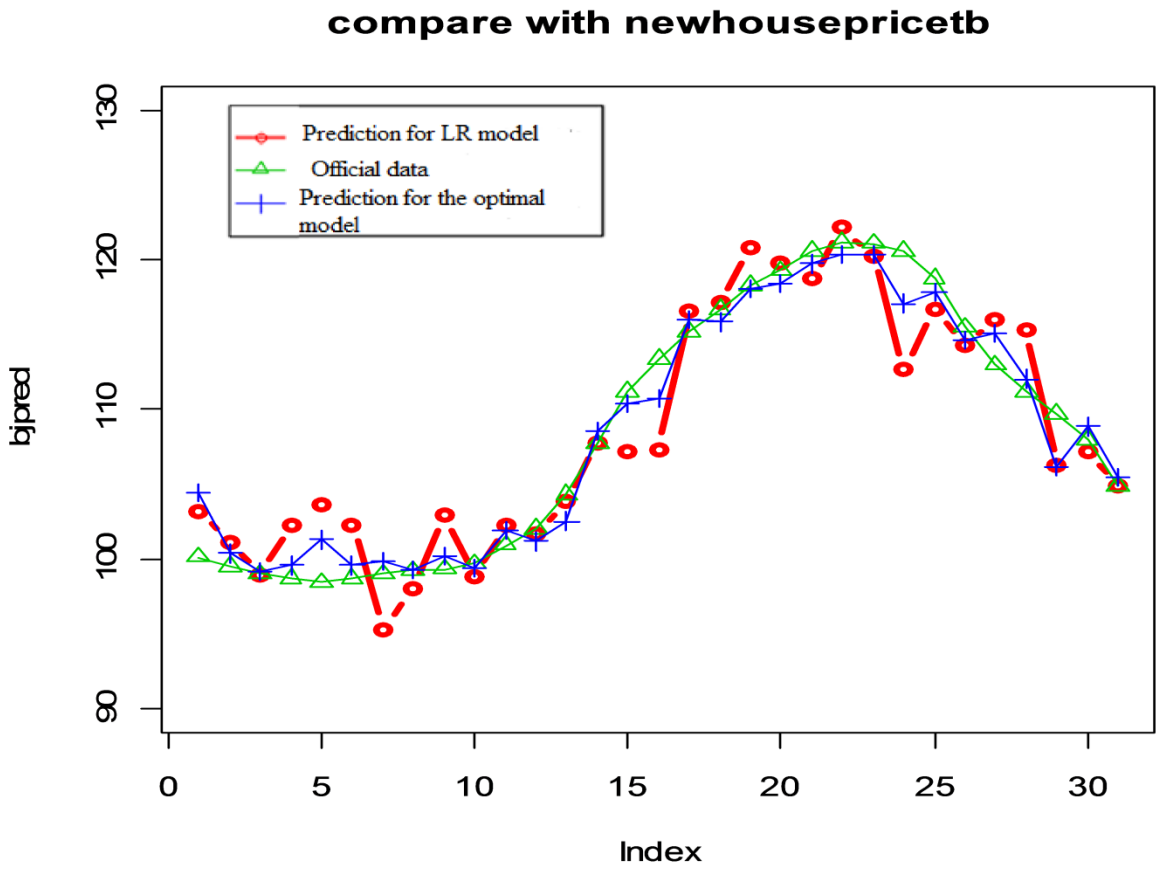
| Cites | Main Keywords Searching Indices |
|---|---|
| Beijing | Prices trend, House source, Decoration |
| Shanghai | House source, Decoration, Low-income housing |
| Guangzhou | Decoration, Public reserve funds, Mortgage interest rates, Low-income housing |
| Nanjing | Prices trend, Real Estate Network, 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 model for New Housing Prices at 6 cities

| 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 |

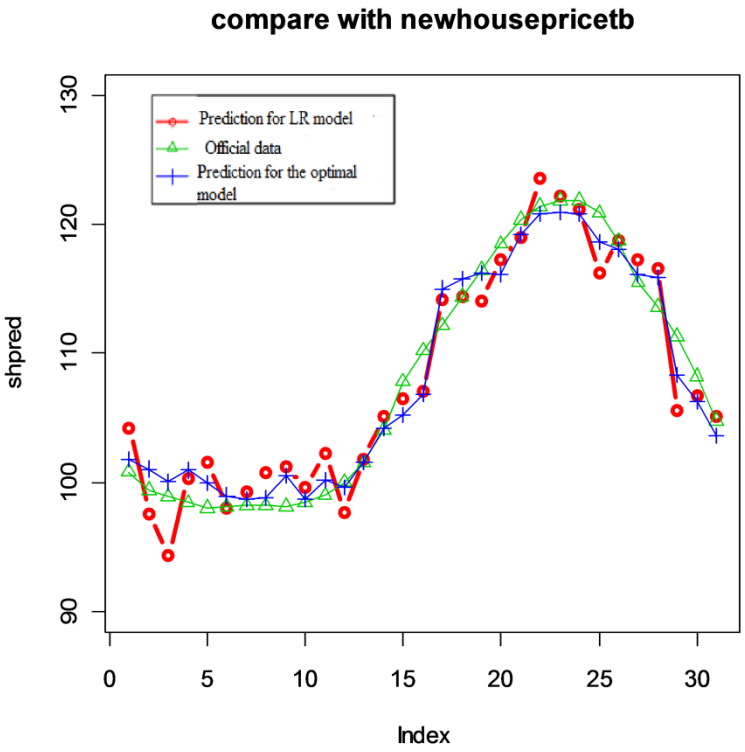
# The Prediction for New Housing Price Index

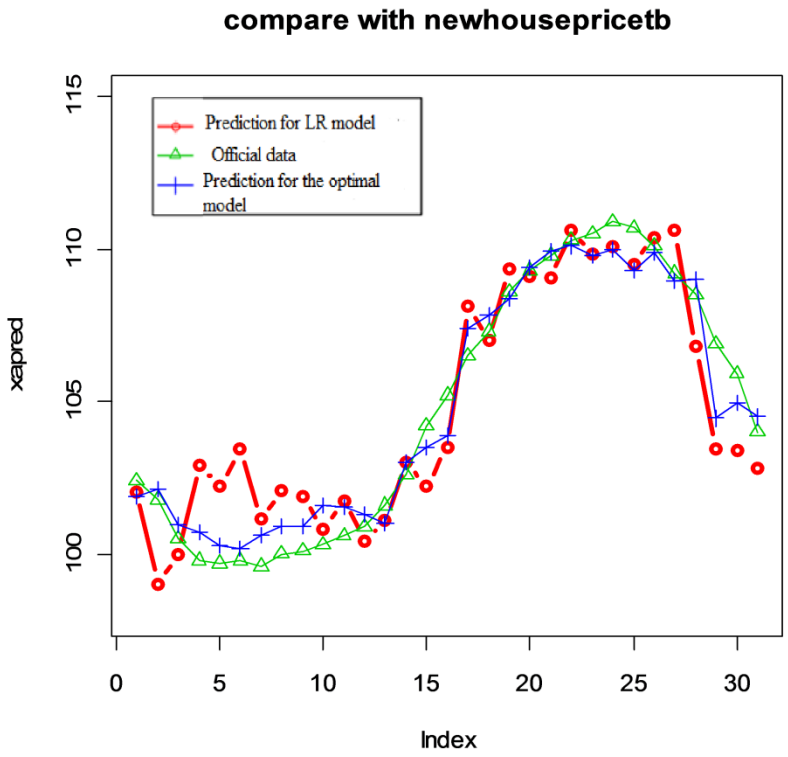
Figure for the Prediction Model of New Housing Price at Beijing



# The Prediction for New Housing Price Index

Figures for the Prediction Model of
New Housing Price at Shanghai &Xian



# Chapter 6. Conclusion and Prospect

***Results***

***Innovation***

***Future Works***

# Results

Based on *Baidu Search Index*, using the cross validation technique and 8 models were successfully fitted and predicted for new housing price index and second-hands housing price index at 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 obtain the last month of new housing price index and second-hands housing price index at the beginning of each month, issued about two weeks early than the official data, solve lag issues for release of traditional housing price index.

# 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.

# Innovation

**Secondly,** using the cross validation technique and 8 analytical models, and they were successfully fitted and predicted for new houses and second-hands 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 more close with the actual value.

# Innovation

**Thirdly,** 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 results.

# Future Works

- This Idea and method can be extended to the monthly data indices such as **CPI**, **Household Income Index**, **Household Consumption Expenditure Index** etc.
- According to **the accumulation of Search Engine Query Data**, the prediction value for Indices will be **more accuracy** in the future.


Thank You