Real Estate Consumption Prediction Research Based on Data Mining

Jinwen Chen[1], Hongshen Pang[2,*], Haiyun Xu[3,*], Yuenan Zhu[4],

*Corresponding author's e-mail: phs@szu.edu.cn, xuhaiyunnemo@gmail.com

The Entertainment Group Limited, Hong Kong, China[1]
Library, Shenzhen University, China[2]
Business School, Shandong University of Technology, China[3]
College of International Studies, Shenzhen University, China[4]

Abstract. With the rapid development of our economy, the real estate has developed rapidly in recent years. This paper starts with the research value of consumer characteristics of data mining in the market research of real estate industry, analyzes the application of data mining in consumer analysis research in the early stage of real estate project development by using examples and SPSS as a tool.

Keywords- Real Estate Consumption; Data Mining; Langfang City

1. Introduction

China housing market has shown rapid growth along with the glorious economic development since 2000—the wealthof housing market occupied 6.8% of the GDP in China in 2018, and the expenditure of house account for nearly 80% ofthe family assets[1][2]. The development of real estate industry is closely related to the development of national economy[3][4][5]. The relationship between the development of the real estate industry and the speed of national economic development is shown in the Table 1.

Table 1. The relationship between the speed of national economic development and the state of the real estate industry

Speed of national economic development	State of the real estate industry
＜4%	shrink
4%~5%	stagnant or retrogress
5%~8%	develop steadily
＞8%	develop rapidly
10~15%	develop more rapidly

This paper uses the data collected from a field survey in Langfang, Hebei in early 2009. The macro data is from Internet and some reference report in secondhand information collection. Competition building combined with online data and field investigation data; Consumers' characteristics was completed through the local way of questionnaire survey. Aiming at the feasibility study of this project, data mining technology is used for analysis, so as to provide a reference for the investment decision of the project.

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China

DOI 10.4108/eai.19-5-2023.2334328

2. Analytical methods and tools

There have been many applications of statistical methods in the preliminary market research of real estate projects in China, but their applications mainly remain in the ordinary statistical results, and the correlation analysis is relatively simple and shallow[6][7][8][9]. The following is an analysis of partial consumption forecast of real estate with SPSS, which is widely used now. General methods of application of SPSS technology in market survey statistical analysis:

1. Input and edit market survey data;

2. Determine the corresponding statistical functions of SPSS according to the research needs and the nature of the problem;

3. Call the menu function of SPSS to get corresponding statistical results and corresponding charts

4. Conduct relevant analysis according to statistical results and charts to provide reliable scientific basis for market investigation.

SPSS technology, which integrates data entry, data management, statistical analysis, report making and graph rendering, provides powerful support and practical methods for the statistical analysis of market survey, and is a good tool for the statistical analysis of market survey. At the same time, in the actual application, it is convenient to use Excel, so this research mainly uses SPSS and Excel to carry out.

In the market forecast, it mainly uses historical data to forecast the future price, finds out the correlation in the government work report and the historical statistical data of real estate, builds a model for predictive analysis through analysis, and obtains the trend of the future housing price and the general estimation formula.

The correlation data between housing price and economic data generally include: gross national product (GDP) and its growth rate, consumer price index and its growth rate, fixed asset investment and its growth rate, permanent resident population and its growth rate, per capita disposable income and its growth rate, Engel coefficient and its change rate. Through the derivation of these values over the years, combined with the real estate data over the years, some relationships between the housing price and these data are obtained, as shown in Figure 1, and the housing price estimation formula (1) is obtained.

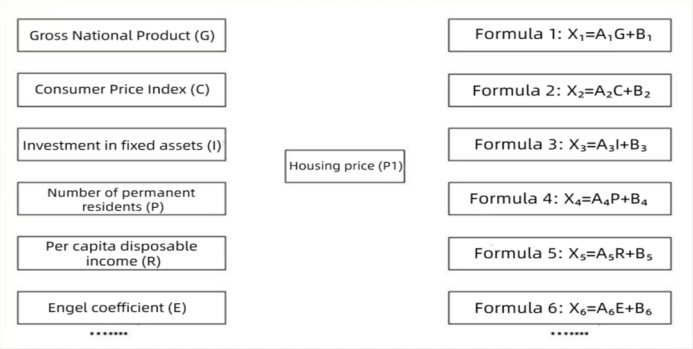

Figure 1. Calculation formulas

The formula can be obtained from Figure 1:

$$P_m = \frac{\sum X_n}{n} \tag{1}$$

3. Real estate prices and macroeconomic data empirical

We made statistics on Langfang real estate and economic data from 2002 to 2008, mined the data by SPSS software, and established the relationship between housing price and economic data to illustrate the above conclusions. Meanwhile, we also explained the first question in the market analysis module and calculated P_m. Let the transaction unit price of a primary residence be X and the gross national product be G, as shown in the Figure 2, Figure 3 and Table 2:

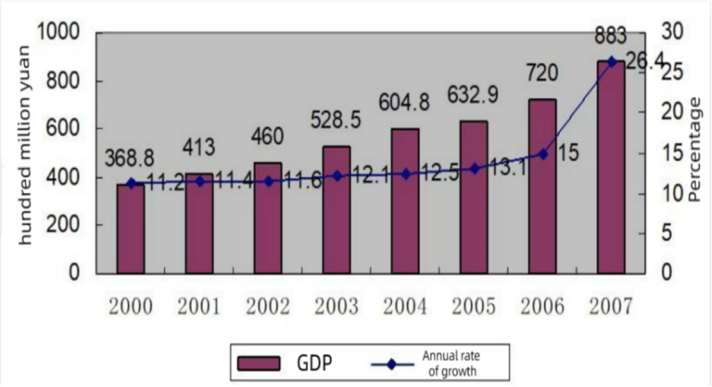

Figure 2. GDP trend of Langfang over the years

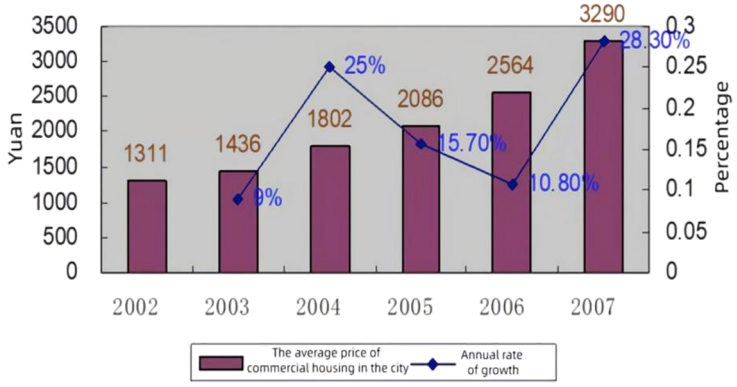

Figure 3. Average price trend of commercial housing in Langfang

Table 2. The corresponding table of Langfang real estate price and GDP

Year	Primary housing price（yuan/㎡） X	GDP (hundred million yuan) G
2002	1311	460
2003	1436	528.5
2004	1802	604.8
2005	2086	632.9
2006	2564	720
2007	3290	883

Note: Statistics for 2008 are not yet available because the survey was conducted in 2008.

3.1 Descriptive analysis

First of all, descriptive statistical analysis is carried out on the transaction price X, gross national product G, and the output result is shown in Table 3 by using SPSS Descriptives function. In this table, from left to right, are the name of the variable, frequency of the observed quantity, minimum value, maximum value, sum, mean value and standard deviation.

Table 3. Descriptive analysis of the transaction unit price of primary residence X and gross national product G

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
X	6	1311	3290	2081.50	745.890
G	6	460	883	638.20	149.366
Valid N (listwise)	6				

3.2 Correlation analysis

Second, the relationship between X variable and G variable is examined for correlation analysis. Analyze→Correlate→Bivariate analysis in SPSS is used. The output result is shown in Table 4. Table 3 shows the correlation between row variables and column variables listed in the cross cell. Top-down statistics are: Person Correlation — Pearson correlation coefficient; Sig.(1-tailed) — single-tailed t test result. The probability that the hypothesis with a correlation coefficient of 0 is true; N is the number of effective observations involved in the calculation of correlation coefficient.

Table 4. Correlation analysis results of X and G

Correlations

		X	G
X	Pearson Correlation	1	.992**
	Sig. (2-tailed)		.000
	N	6	6

G	Pearson Correlation	.992**	1
	Sig. (2-tailed)	.000	
	N	6	6

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4 shows that there is a strong positive correlation between the unit price of primary residential transaction X and the gross national product G, and the Pearson correlation coefficient is as high as 0.992. Therefore, there is a strong relationship between these two factors.

3.3 Binary linear regression prediction

The transaction price of primary housing X and gross national product G are predicted by the binary linear regression prediction method. Determine the relationship formula. The specific steps are to use Analyze→Regression→Linear analysis in SPSS.

Table 5. Binary linear regression prediction of X and G

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1080.774	202.502		-5.337	.006
	G	4.955	.310	.992	15.968	.000

a. Dependent Variable: X

Select Table 5 (regression coefficient analysis table) from the many tables output. Where, Model is the number of regression equation model, Unstandardized Coefficients are non-standardized regression Coefficients, Standardized Coefficients are standardized regression coefficients, and t is the t value of hypothesis test whose partial regression coefficient is 0, Sig. is the significance level value of the hypothesis test with a partial regression coefficient of 0.

Table 5 shows that Constant and gross national product (G) have statistical significance, mainly because their significance level is not high. Therefore, the model equation can be derived from Table 4-5: X= -1080.774 + 4.955G. If the gross national product of Langfang is expected to be 100 billion yuan in 2008, it is not difficult to conclude that the transaction price of primary housing in 2008 X= -1080.774 + 4.955×1000 = 3874 (yuan /㎡), according to the model equation.

Through correlation analysis, binary linear regression forecasting method, and the use of SPSS, the real estate consumer price forecast, through more examples and historical data accumulation and analysis, will get a relatively complete calculation model, that is, the primary residential transaction price and gross national product has a relatively direct linear relationship. Of course, the above examples are relatively simple linear regression operations. In fact, multiple linear regression operations contain a lot of knowledge and content, which is relatively complex. This

example just extracts the general regularity from it to guide the establishment of a more sound model in the future.

Through the above method of analysis, deduce the relationship between the housing price and other factors and calculation formula, and then combine these economic data to get P_m.

4. Real estate consumption forecast questionnaire analysis example

4.1 Questionnaire data input and definition

After the above correlation analysis of housing price and macro economy, this paper also uses the example of mining questionnaire to analyze the influencing factors of housing price. A total of 684 valid questionnaires (sample questionnaires are shown in the appendix) were applied in this analysis. After the questionnaire results were collected, the questionnaire data results were merged, integrated and preprocessed, and input into Excel with field names as Q1, Q2, Q3..... Q17. Each field corresponds to an optional answer to the question set in the questionnaire.

4.2 Option mining

In this paper, the two questions of household population (Q15e) and affordable housing unit price (Q8) are selected for data analysis to understand the basic relationship between household population and affordable housing unit price. First, we need to know the basic information of "affordable unit price (Q8)" data, such as mean, standard difference, etc. Choose Analyze→Descriptive Statistics→Descriptives. The following dialog box is displayed(including Table 6 and Table 7):

1. Static descriptive analysis

Table 6 Description of household population and affordable housing unit price

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Q8	684	2500	6500	3665.94	1094.654
Valid N (listwise)	684				

Sample number, minimum value, maximum value, mean value and standard deviation of Q8 can be obtained from the result browsing window. It can be seen that the total mean and standard deviation of the 684 data is 3665.94 and 1094.654.

2. Conditional descriptive analysis

The above is just the description of a single data, just a statistical result, is not too significant. We have to put it into other options to mine the data. The factor "family population" group analysis, then select Organize output by family population (Q15e) and then redescribe, then see a variety of different descriptions of family population:

Table 7. The relationship between the size of various households and the price of housing they can afford

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Q8	67	2500	6500	3671.64	1217.151
Valid N (listwise)	67				

a. Q15E = 1

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Q8	43	2500	6500	3581.40	1143.981
Valid N (listwise)	43				

a. Q15E = 2

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Q8	205	2500	6500	3700.00	1065.962
Valid N (listwise)	205				

a. Q15E = 3

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Q8	198	2500	6500	3707.07	1107.222
Valid N (listwise)	198				

a. Q15E = 4

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Q8	171	2500	6500	3596.49	1059.017
Valid N (listwise)	171				

a. Q15E = 5

It can be seen from the description that the average and standard difference of the two groups are shown respectively. It is obvious that the consumption level of a family of three and a family of four is relatively high compared with other families, but the overall level is relatively average, which also indicates that the price of 3,500 yuan most meets the needs of the market. Statistical indicators can only give the general situation of the data, which is not as intuitive as the histogram. Therefore, we can output the histogram for observation, and its data features are more obvious, as shown in the Figure 4:

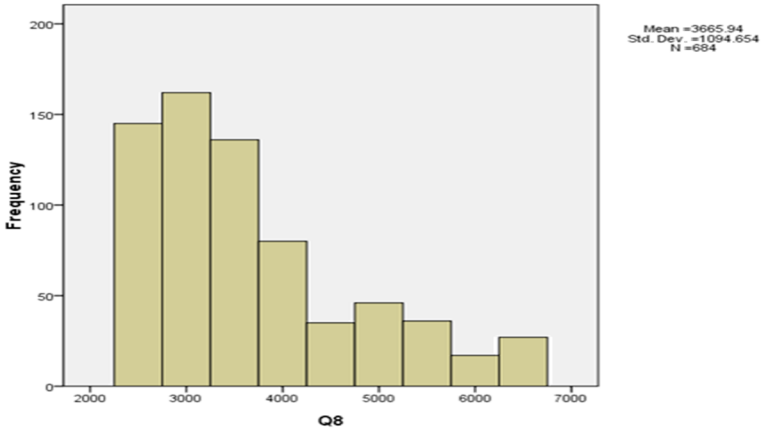

Figure 4. Histogram of the relationship between household population and affordable housing price

4.3 Statistical analysis

Next, SPSS is used to test the comparison of the mean of two samples in the group design. The basic function of MEANS is to group calculate description statistics for specified variables. A series of univariate descriptive statistics include MEANS, STD DEVIATION(standard deviation), SUM, COUNT(number of observed values), VARIANCE and so on. Analysis of variance table and linear test results can also be given. Select the value we want to compare in OPTION: means→number of cases→standard deviation. The following Table 8 and Table 9 appears:

Table 8. Analysis table of household population and affordable unit prices

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Q8 * Q15E	684	100.0%	0	.0%	684	100.0%

Table 9 Report table of household population and affordable unit prices

Report

Q8

Q15E	Mean	N	Std. Deviation
1	3671.64	67	1217.151
2	3581.40	43	1143.981
3	3700.00	205	1065.962
4	3707.07	198	1107.222
5	3596.49	171	1059.017
Total	3665.94	684	1094.654

The first table is a summary of our MEANS process; The second table is the report. The basic statistics and data are clear. For example: if the sample proportion extracted when we sample is consistent with the real proportion of the whole home buyer's family, it can be seen from the proportion of total visitors that a family of more than three people accounts for the largest proportion; Therefore, our target consumer group is mainly this group. The average price of 3700 yuan represents the consumption level of consumers[10].

We can excavate multiple factors again, so as to obtain a more accurate forecast of the price factor. Then, combined with industry experience, we can import the final results into the real estate consumption forecast research information system, so as to calculate the project and draw a conclusion.

5.Conclusions

The real estate industry is a complex nonlinear system with large amount of data, strong correlation and many influencing factors. This paper mainly aims at the characteristics of the real estate industry, through the introduction of data mining method, combined with the theoretical background of the real estate market research to study several problems in the real estate consumption forecast, and draws some conclusions:

(1) By using data mining and processing methods such as correlation, regression, factor and correlation analysis, SPSS software is used to establish the relationship formula between housing price and macroeconomic data, and the relationship between household population (Q8) and affordable housing price (Q15E) in the questionnaire of real estate market research is also obtained, and two influencing factors of housing price are extracted. The key points of housing price analysis in real estate consumption forecast are clarified.

(2) Data mining technology will be applied and attached importance by more and more real estate enterprises, and has broad application prospects in the feasibility of early project, consumption forecast and decision support of real estate.

(3) Data mining has high requirements on the number of data, especially in the association analysis of consumer behavior. If the number of data samples is insufficient, valuable association rules cannot be obtained.

Acknowledgments: This work was supported by the China Postdoctoral Science Foundation (No. 2019M650803, No.2020T130637). Hongshen Pang and Haiyun Xu are the corresponding authors.

References

[1] Zhang Jixuan, Deng Xiaoyu(2022).Real Estate Tax, Housing Price, and Housing Wealth Effect: An Empirical Research on China Housing Market.DISCRETE DYNAMICS IN NATURE AND SOCIETY.

[2]Cen Xi (2009). Application of data mining technology in real estate industry. Science & Technology Economic Market.

[3]Hu Ping (2009). Real estate investment and management practice. Zhejiang University Press.

[4]Xiao Xiaojun (2005). Chinese real estate dictionary real estate base volume. Silver Sound Audio-visual Publishing House.
[5]Zhang Liang (2007). The Study of Data Mining of Customer's Information of Tianjin's Real Estate Market. Tianjin Polytechnic University.
[6]Bai Zijie (1997). In a word life knowledge series: in a word real estate tips. Aviation Industry Press.
[7]Aoki, K., J. Proudman, and G. Vlieghe(2004). House prices, consumption, and monetary policy: A financial accelerator approach. Journal of Financial Intermediation.
[8]Chen, B., and R. Yang(2013). Land supply, housing price and household saving in urban China: Evidence from urban household survey. Economic Research Journal .
[9]W. Sun, S. Zheng, D. M. Geltner, and R. Wang(2017). The housing market effects of local home purchase restrictions: evidence from Beijing.The Journal of Real Estate Finance and Economics.
[10]J. Y. Campbell and J. F. Cocco(2007).How do house prices affect consumption? Evidence from micro data.Journal of Monetary Economics.