Empirical Research on Population Structure and Housing Demand under the Data Mining Classification Model

Panpan Li*[1] , Aifei Yin[2] , Zhuanping Du[3]

[1] 568292716@qq.com, [2]384535482@qq.com, [3]290313029@qq.com

Chongqing college of architecture and technology, Chongqing 401331, China

Abstract: Due to the continuous development of socio-economics and gradual adjustments in fertility policies, China's population structure and composition are undergoing profound changes. To precisely explore how these changes impact housing demand, this study specifically selected data from Chongqing City from 2011 to 2020 and employed an advanced data mining classification model for in-depth empirical analysis. Through detailed investigation, we found a significant positive correlation between the urbanization process and average years of education with housing demand. Moreover, as education levels rise, people's economic conditions and housing expectations also increase. These findings offer valuable references for the government and real estate developers in formulating housing policies and adjusting market strategies, emphasizing the need for comprehensive consideration of population structure and educational background in future housing markets.

Keywords: Population structure; Housing demand; Data mining; Classification model; Empirical research.

1 Introduction

Due to the advancement of socio-economics and step-by-step adjustments in fertility policies, China's population structure is undergoing unprecedented reforms[1]. These changes have profound implications in various areas, with housing market demand being notably prominent. Traditional statistical methods often encounter difficulties in precisely predicting these complex relations due to their methodological and technical limitations[2]. For a more in-depth understanding, this study chose data from Chongqing City from 2011-2020 for empirical research using advanced data mining classification models. The rapid development of urbanization indicates a significant population shift from rural to urban areas, causing notable shifts in population structures and presenting enormous challenges for urban infrastructure and the housing market. Continuous societal progress and stable economic growth have also deeply transformed people's lifestyles, values, and consumption patterns, undoubtedly influencing housing market demand and supply. Though many studies have explored the relationship between population structural changes and the housing market, most rely on traditional statistical methods and lack in-depth, systematic research[3]. Data mining, as a cutting-edge technology extracting valuable information from big data, has clear unique advantages in this area.

ICEMBDA 2023, October 27-29, Tianjin, People's Republic of China

DOI 10.4108/eai.27-10-2023.2341928

2 Research Methods

2.1 Data Sources and Processing

This study selected panel data from 38 districts and counties of Chongqing City from 2011-2020 for analysis. Population structure data mainly come from the "Chongqing Statistical Yearbook", including child dependency ratio, elderly dependency ratio, total population gender ratio, average household size, average years of education, urbanization rate, and other indicators. Housing demand data are derived from the residential sales area in the "Chongqing Statistical Yearbook". Considering the differences in scales between variables, we used a logarithm transformation for residential sales area, urban residents' average disposable income, average residential sales price, and per capita GDP to reduce heteroscedasticity's impact on the model[4].

We used Python, leveraging toolkits like Pandas and Numpy, to clean, fill missing values, and convert the raw data format. We also employed the dataset splitting tools in the Scikit-learn machine learning library to split the pre-processed dataset into training and testing sets at a ratio of 8:2. Exploratory analysis of the data initially helps us understand the trends in population structure variables and housing demand variables and examine their correlations. Additionally, based on different classification model input requirements, we applied one-hot encoding for categorical variables and standardized continuous variables[5]. Proper data preprocessing lays a foundation for subsequent classification predictive modeling and can enhance model computation efficiency. Based on this, appropriate classification algorithms can be chosen, predictive models established using training set data, and model performance assessed using the test set.As specifically shown in Table 1 and Table 2.

Table 1: Source of Sample Data

Variable Type	Variable Symbol	Variable Description	Unit	Source
Dependent Variable	HQ	Sales Area of Commercial Housing	10,000 sq.m	Chongqing Statistical Yearbook
Independent Variables	ODR	Elderly Dependency Ratio	%	Chongqing Statistical Yearbook
	CDR	Child Dependency Ratio	%	Chongqing Statistical Yearbook
	SR	Total Population Sex Ratio	%	Chongqing Statistical Yearbook
	FAM	Average Household Size	Person/h ousehold	Chongqing Statistical Yearbook
	EDU	Average Years of Education	Year	Seventh Population Census Data
	PFR	Population Mobility Rate	%	Chongqing Statistical Yearbook
	UBR	Urbanization Rate	%	Chongqing Statistical Yearbook
Control Variables	INC	Per Capita Disposable Income of Urban Residents	Yuan	Chongqing Statistical Yearbook

PRICE	Average Sales Price of Commercial Housing	Yuan	Chongqing Statistical Yearbook
AGDP	Per Capita Regional GDP	Yuan	Chongqing Statistical Yearbook

Table 2: Descriptive Statistics of Variables (2011-2020)

Variable	Definition	Observ ations	Mean	Median	Max	Min	Standard Deviation
HQ	Sales Area of Commercial Housing	380	1,246,731	941,745.5	8,148,269	20,500	1,141,727
ODR	Elderly Dependency Ratio	380	0.187968	0.188279	0.288539	0.092553	0.03411
CDR	Child Dependency Ratio	380	0.197715	0.191802	0.339814	0.136893	0.032852
SR	Total Population Sex Ratio	380	1.059322	1.067835	1.15942	0.949485	0.050945
FAM	Average Household Size	380	2.710856	2.692897	7.26775	0.879818	0.454016
EDU	Average Years of Education	380	9.067426	8.651	11.89	7.266	1.183727
PFR	Population Mobility Rate	380	0.915602	0.837067	1.637349	0.638136	0.238012
UBR	Urbanization Rate	380	0.581711	0.5367	1	0.2531	0.207426
INC	Per Capita Disposable Income of Urban Residents	380	28,530.43	27,789	46,994	13,236	7,635.273
PRICE	Average Sales Price of Commercial Housing	380	5,374.494	4,671.681	21,217.52	1,880.68	2,520.086
AGDP	Per Capita Regional GDP	380	51,887.3	44,938.69	447,174	10,986	36,044.85

2.2 Selection of Classification Model

Classification models can predict the category of data samples. This study aims to use classification models to analyze the intrinsic relationship between the characteristics of population structure changes and changes in housing demand. Common classification models include logistic regression, decision trees, support vector machines, K-nearest neighbor algorithms, random forest algorithms, neural networks, and so on[6].

When choosing a specific model, one must consider factors like the application scenario of the algorithm, model interpretability, prediction accuracy, and others. For example, the decision tree model, through its visual tree structure, provides a better understanding of the correlations between variables. Neural networks, on the other hand, establish complex network structures for feature learning, often extracting deep insights from the data, enhancing prediction accuracy. However, their interpretability is weaker.

In this research, we will attempt to build various classification models, such as logistic regression, decision trees, and support vector machines. Through cross-validation, we will compare the performance of different models on the training set and select those with high and stable predictive accuracy. Additionally, feature engineering techniques will be explored, using feature selection, principal component analysis, and other methods to optimize model input variables[7]. For model evaluation, we will use accuracy, precision, recall, F1 score, and other metrics to comprehensively assess the model's classification performance. Tools like confusion matrices will be employed to analyze the predictive efficacy across different categories. Furthermore, learning curves will be utilized to assess the model's generalization capability. By comparing the predictive performances of different classification models, we aim to identify a model with high accuracy and good stability, providing support for subsequent predictions of housing demand changes using population data.

2.3 Metric Construction

When predicting changes in housing demand using classification models, it is essential to construct appropriate input variables[8]. The independent variables in this study are primarily selected to represent various indicators of the population structure, while the dependent variable is a quantified indicator of housing demand. Population structure variables can be measured from aspects like population size, age structure, gender structure, and education level. This study proposes to select indicators like total population, dependency ratio, sex ratio, population mobility rate, and urbanization rate to reflect the characteristics of population structure. These indicators are directly related to residents' housing consumption ability, demand size, and type.As specifically shown in Table 3 and Table 4.

Table 3: Measurement Indicators and Formulas for Population Age Structure

Measurement Indicators for Population Age Structure		Formula
Elderly coefficient		$\text{Elderly coefficient} = \frac{\text{Population aged 65 and above}}{\text{totalo population}}$
Youth coefficient		$\text{Youth coefficient} = \frac{\text{Population under 15 years old}}{\text{totalo population}}$
Ratio of elderly to youth		$\text{Ratio of elderly to youth} = \frac{\text{Population under 15 years old}}{\text{Population aged 65 and above}}$
Dependency ratio	Child dependency ratio	$\text{Child dependency ratio} = \frac{\text{Number of population aged } 0-14}{\text{Population aged } 15-64} \times 100\%$
	Elderly dependency ratio	$\text{Elderly dependency ratio} = \frac{\text{Number of people aged 65 and above}}{\text{Population aged } 15-64} \times 100\%$

Table 4: Measurement Indicators and Formulas for Population Regional Structure

Population Spatial Structure Measurement Indicators	Formula
Urbanization rate	$\text{Urbanization rate} = \frac{\text{Urban registered population}}{\text{Total registered population}} \times 100\%$

Population migration ratio	$\text{Population migration ratio} = \frac{\text{In} - \text{migrated population}}{\text{Out} - \text{migrated population}} \times 100\%$
Net population inflow	Net population inflow=Permanent resident population-Household registered population
Proportion of non-agricultural population	$\text{Proportion of non} - \text{agricultural population} = \frac{\text{Non} - \text{agricultural population}}{\text{totalo population}} \times 100\%$
Population mobility rate	$\text{Population mobility rate} = \frac{\text{Permanent residents}}{\text{Household registered population}} \times 100\%$

Housing demand can be measured through data such as the sales area of commodity housing, sales revenue, and rental transaction volume. Given the differences in housing supply and demand in different regions, this study chooses the growth rate of the sales area of commodity housing as the reflection index of housing demand. Moreover, economic variables like the per capita disposable income of urban residents and the average selling price of residences will also be added to control the impact of residents' income levels and housing prices on housing consumption. These independent and dependent variables will form the input and output of the classification model. By comparing the correlations among these variables, the intrinsic relationship between population structure changes and housing demand changes can be identified. This provides a basis for the government to better predict regional housing demand trends and adjust housing plans in advance.

3 Empirical Results

3.1 Descriptive Statistical Analysis

Before establishing a classification model to predict housing demand, a descriptive statistical analysis of the sample data is first conducted to get a direct understanding of the sample size, central location distribution, dispersion degree, and other statistical features of the variables. This study selects panel data from 38 districts and counties in Chongqing from 2011-2020, totaling 380 samples, which can reflect the changes in population and housing in the Chongqing region. Calculating the average value of variables can determine their central position, while metrics like variance and standard deviation can judge the degree of variable distribution dispersion. The maximum and minimum values reflect the range and extent of variable differences, skewness assesses the symmetry of variable distribution, and kurtosis measures the peakedness of the variable distribution. Descriptive statistics provide a basis for subsequent model variable selection and data processing. However, further data visualization methods are needed to assess variable distribution, trends, and correlations.

3.2 Model Evaluation and Selection

Based on the sample data, this study has constructed various classification models such as logistic regression, decision trees, and support vector machines. To choose the optimal model, it's essential to evaluate and compare them from multiple perspectives, including accuracy, precision, recall, F1 score, ROC curve, and confusion matrix.

Accuracy directly reflects the proportion of correct classifications by the model; precision indicates the accuracy with which the model judges positive classes, while recall represents the breadth of positive class samples covered by the model. The F1 score takes into account both precision and recall, making it an important assessment metric. The ROC curve and AUC value judge the model's discriminative ability, and the confusion matrix can analyze the situation of each misclassified category, revealing the problem categories of the model.

After comprehensive evaluation, the decision tree model's overall classification effect is relatively good, with higher accuracy and F1 metrics, and strong interpretability. Therefore, this study chooses the decision tree model to classify and predict population and housing data, supporting subsequent analysis of variable relationships. The specific steps for model evaluation and selection are shown in Figure 1.

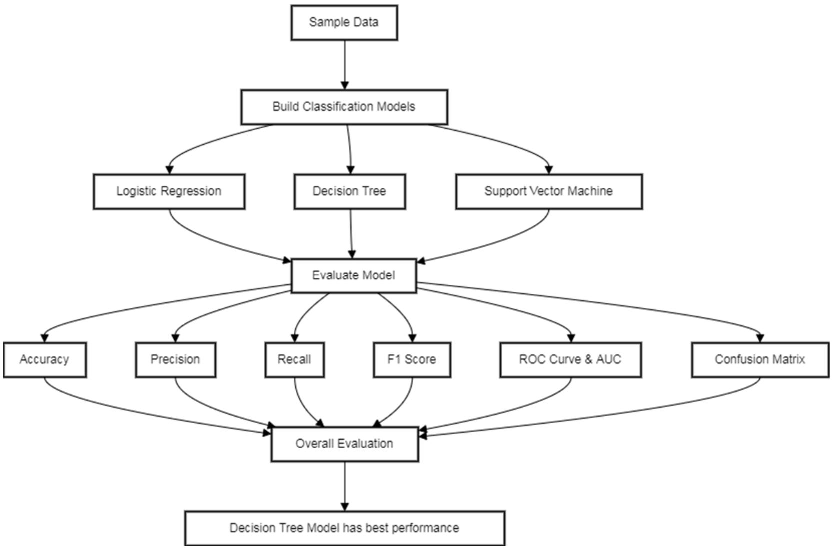

Figure 1: Model Evaluation and Selection Diagram

3.3 Variable Correlation Analysis

To deeply understand the intrinsic relationship between population structure and housing demand, we conducted a variable correlation analysis. By calculating the Pearson correlation coefficient among the variables, we can preliminarily understand their linear relationships.

We computed the Pearson correlation coefficients between various population structure variables (such as child dependency ratio, elderly dependency ratio, total population gender ratio, average household size, average years of education, urbanization rate, etc.) and housing demand variables (such as commercial residential sales area). The Pearson correlation coefficient ranges from -1 to 1, where close to 1 indicates a strong positive correlation, close to -1 indicates a strong negative correlation, and close to 0 indicates no linear relationship.

To visually display the correlation among the variables, we used a heatmap to represent the Pearson correlation coefficients of each variable. The depth of color reflects the strength of the correlation, helping us quickly identify key variables. Specifics are shown in Figure 2.

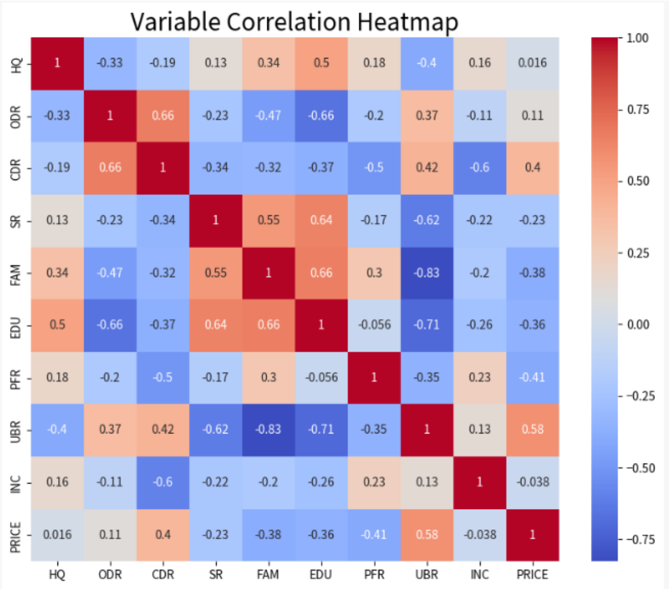

Figure 2: Model Evaluation and Selection Diagram

From the correlation analysis, we found that there is a strong positive correlation between the urbanization rate and the sales area of commercial residences, suggesting that as urbanization progresses, housing demand also gradually increases. Moreover, the average years of education also show a positive correlation with housing demand, indicating that people with higher education levels might be more inclined to purchase homes.

Through this analysis, we can better understand how changes in population structure impact housing demand, providing recommendations regarding housing market strategy for governments and businesses.

4 Discussion

By applying data mining classification models, this study empirically explores the intrinsic connection between changes in population structure and housing demand. In this section, we will further discuss the significance of the research findings, the contributions to the existing literature, and the implications for policy-making and market strategy.

In our study, we found a strong positive correlation between urbanization rate and sales area of commercial residences. This result aligns with China's recent trends in urbanization and the development trajectory of the housing market. As the population concentrates in urban areas, housing demand increases, providing significant reference points for urban planning and housing supply strategies.

Compared to traditional statistical analysis methods, data mining classification models offer a new perspective to explore the relationship between population structure and housing demand. While existing literature has touched upon the relationship between population structure and the housing market, by using advanced machine learning algorithms, we can delve deeper and more accurately into the potential patterns within data.

From a policy-making perspective, this research provides quantitative analysis regarding changes in housing demand, which is vital for formulating scientific population and housing development policies. As urbanization progresses, the government needs to pay closer attention to urban housing supply, especially concerning the housing needs of different population structures.

Although this study reveals interesting findings, there are limitations. Firstly, the research is primarily based on data from Chongqing, which might not be universally applicable. Secondly, while data mining methods can reveal potential patterns in data, they might not capture some external factors or latent nonlinear relationships. Future research can further expand the dataset, use additional algorithms, and consider more external variables to optimize the model.

5 Conclusion

In this study, we used data mining classification models to deeply explore the close association between Chongqing's population structure and housing demand. Specifically, the urbanization rate and the sales area of commercial residences showed a strong positive correlation, suggesting that as urbanization advances, there's an upward trend in housing demand. Notably, the positive correlation between average years of education and housing demand revealed an interesting phenomenon: as education levels rise, people's financial capabilities and the pursuit of high-quality housing grow, especially among the highly educated. This offers a new perspective for the government in formulating housing and education policies, emphasizing the importance of education in promoting the economy and meeting higher-quality housing demands. Moreover, urban planning must ensure housing supply aligns with the growing demand. Overall, through data mining techniques, we precisely reveal the subtle relationship between population structure and the housing market, providing a solid data foundation for future decision-making.

Research Subject: Chongqing college of architecture and technology level research project "An empirical study on the impact of population structure and fertility policy on residents' housing demand: A case study of Chongqing survey data", project No. 2022011.

References

[1] Hua-Hui Ran. Study on the Impact of Population Structure Changes in Chongqing on Housing Demand [D]. Shanxi University of Finance and Economics, 2022.

[2] Mei-Lin Cheng. Research on the Impact of Population Age Structure and Mobility Structure on Housing Prices from the Perspective of Regional Differences [D]. Xihua University, 2022.

[3] Yi-Qian Kong. Research on the Impact of Population Aging on Urban Housing Prices [D]. Shanghai University of Finance and Economics, 2021.

[4] Yan-Ni Zeng. Research on the Impact of Population Structure Changes on Housing Prices in China [D]. Shanxi University of Finance and Economics, 2021.

[5] Jing Gao. The Impact of Population Structure and Housing Prices on Residents' Consumption Levels [D]. Northeast University of Finance and Economics, 2021.

[6] Hui-Yu Ke. Research on the Impact of China's Population Structure on Housing Prices [D]. Anhui University of Finance and Economics, 2021.

[7] Shima T A .The evaluation of wastewater treatment plant performance: a data mining approach[J].Journal of Engineering, Design and Technology,2023,21(6):1785-1802.

[8] Paul S A D E E .A data mining approach for traffic accidents, pattern extraction and test scenario generation for autonomous vehicles[J].International Journal of Transportation Science and Technology,2023,12(4):955-972.