Proceedings of ICFTBA 2024 Workshop: Finance's Role in the Just Transition
DOI: 10.54254/2754-1169/141/2024.GA18866

# *Forecasting House Prices in Shanghai Using Common Factor Analysis and Multiple Linear Regression Analysis*

**Zhewei Deng**[1,a,*]

[1]*College of Civil Engineering, Fuzhou University, Fuzhou, Fujian, 350108, China*
*a. 052105127@fzu.edu.cn*
**corresponding author*

***Abstract:*** Predicting housing prices is critical for understanding economic trends and making decisions in real estate markets. This is accomplished by identifying important influencers and utilizing prediction models. Predicting these prices is essential for understanding broader economic trends and guiding decisions in real estate markets. The model was trained and evaluated using economic and financial data from the CEInet Statistics Database, with key variables representing major financial, demographic, and economic characteristics included. To estimate housing prices, a combination method combining common factor analysis and multiple linear regression was used. The KMO test result of 0.845 and the scores of Bartlett's Test of Sphericity ($\chi^2$=4296.29, p=0.000) confirmed the data's suitability for factor analysis. Factor analysis reduced the dimension of the dataset, revealing two important components responsible for 80.01% of the volatility in home costs. The multiple linear regression model based on these characteristics was tuned to attain good prediction accuracy. The model predicted housing prices in Shanghai with an accuracy of 94.54%.

***Keywords:*** Housing Price Prediction, Common Factor Analysis, Multiple Linear Regression, Factor Analysis.

## 1. Introduction

With its roles as a significant investment engine and a vital gauge of economic health, the housing market is vital to world economic growth. Moreover, individuals' life happiness is heavily influenced by housing prices, and the capacity to afford adequate housing has a big impact on their subjective well-being [1]. However, China has seen a considerable slowdown of increase in house prices since 2021; in several cities, prices have even slightly declined. The COVID-19 pandemic has impacted the real estate market, causing prices and transaction volumes to fall in some cities. For example, one of China's most economically advanced areas, the Greater Bay Area (GBA), has seen a particularly notable effect from COVID-19 on the housing market. Due to this worldwide epidemic, several real estate transactions were canceled or postponed as purchasers' purchasing power declined [2]. This case illustrates how the demand for housing has been impacted even in the wealthiest regions, which has resulted in a declining trend in property prices that reflects the overall drop in housing prices throughout China. As one of China's most economically developed regions, Shanghai's housing prices are particularly interesting to investigate. Its distinct economic structure and durability in the real estate sector make it an important region for understanding broader market dynamics and forecasting future developments.

DOI: 10.54254/2754-1169/141/2024.GA18866

By utilizing a method that combines multiple linear regression and common component analysis, this study, which is only focused on the Shanghai housing market, seeks to increase the accuracy of housing price forecasts. In order to minimize the dimension of the several variables influencing home prices and turn them into a set of common factors that represent the key features of the data, common factor analysis will be utilized. Then, a multivariate linear regression model will be employed to predict Shanghai home prices using these common features. By integrating these methods, the study hopes to improve the predictability and interpretability of the results, simplify the modeling process, and add a more useful housing price prediction model to the body of knowledge.

## 2. Data and Methodology

### 2.1. Data Sources

The data comes from the extensive and frequently used "CEInet Statistics Database", which offers a variety of economic, social, and financial data. Access to high-quality, trustworthy data covering different aspects of the Chinese economy, such as housing prices, macroeconomic indicators, and demographic statistics. The dataset spans from January 2006 to November 2018, with each data factor containing 155 monthly data points.

To reflect the complex influences on the housing market, a wide selection of economic, demographic, and financial variables has been used in the analysis of home price prediction. These variables are listed in Table 1. Through their reflection on the macroeconomic conditions (X1, X2, X3, X7, X8, X11), demographic transitions (X5, X10), and market-specific developments (X4, X6, X12) that impact housing supply and demand, each of these components makes a distinct contribution to the knowledge of the dynamics underlying housing pricing.

Table 1: Analyzing Variables

| Variable | Measurement Method | Symbol |
|---|---|---|
| Housing Price | Average Housing Price (Yuan/square meter) | P |
| Regional Economic Strength | Gross Domestic Product (GDP) per capita (Yuan per person) | X1 |
| Inflation Level | Cumulative Customer Price Index (CPI) (base month = 100) | X2 |
| Investment in Fixed Assets | Cumulative Fixed Asset Investment (base month = 100) | X3 |
| Long-term Loan Costs | Loan Rates for More Than Five Years (%) | X4 |
| Population Size | Registered Population (10 thousand People) | X5 |
| Currency Exchange Influence | US dollar to yuan exchange rate (Yuan) | X6 |
| Income Level of Residents | Average Disposable Income per Capita (Yuan) | X7 |
| Labor Market Health | Unemployment Rate (%) | X8 |
| Supplements of Land | Area of Transferred Land Use Rights (10000 Square Meter) | X9 |
| Household Demographics | Average Household Size (Person) | X10 |
| Investment in Real Estate | Investment in Real Estate Construction (100 million Yuan) | X11 |
| Government Bond Yield | 3-year Treasury yield (%) | X12 |

DOI: 10.54254/2754-1169/141/2024.GA18866

## 2.2. Methodology

The methodology involves several key steps. First, Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin (KMO) test is used to assess if the data are suitable for factor analysis. While Bartlett's test searches for statistically significant correlations between variables, the KMO test assesses the quality of sample collection. After these tests, the dataset's dimensionality is decreased using Principal Component Analysis (PCA), which extracts important components that account for most of the variance in the original variables. Based on the PCA results, a multiple linear regression analysis is performed using the extracted components as independent variables in order to develop a prediction model for housing prices. This method makes sure the study retains its predictive value while making complex data easier to understand.

### 2.2.1. KMO Test

The Kaiser-Meyer-Olkin (KMO) test was used to ascertain if the 12 components that were chosen were appropriate for linear regression analysis. Formula (1) gives the KMO measure of sampling adequacy. The partial covariance matrix is represented by $U_{ij}$and the correlation matrix by $R_{ij}$. The KMO value is between 0 and 1. There appears to be sufficient sampling based on the KMO values, which range from 0.8 to 1.0. If the value is less than 0.5, the factor analysis findings are most likely inadequate for the data analysis [3].

$$KMO_j = \frac{\sum_{i \neq j} R_{ij}^2}{\sum_{i \neq j} R_{ij}^2 + \sum_{i \neq j} U_{ij}^2} \tag{1}$$

With a KMO test score of 0.845 in this study, the cutoff value of 0.5 is significantly surpassed. The high KMO score indicates that there is sufficient correlation between the elements and that the data has a considerable degree of shared variance. Specifically, the factors and the data can be utilized for factor analysis and more complex statistical studies such as linear regression due to the high correlation.

### 2.2.2. Bartlett's Test of Sphericity

The null hypothesis that the variables are orthogonal is tested using Bartlett's Test of Sphericity, which implies that the variables are unrelated and unsuitable for identifying underlying structures as the original correlation matrix is an identity matrix. The formula of Bartlett's test is given by:

$$\chi = -\left(\mathrm{n} - 1 - \frac{2\mathrm{p} + 5}{6}\right) \times \ln\left|\mathrm{R}\right| \tag{2}$$

The number of the variables is denoted by p, the total sample size by n, and the correlation matrix by R. The data are suitable for factor analysis since the results of Bartlett's test of sphericity were significant ($p < 0.001$), meaning that the variables' correlations deviate considerably from zero [4].

Table 2 shows a highly significant. $\chi^2$ of 4296.29 with a p-value of 0.0. The outcome shows that the variables have significant linear correlations with one another, proving that the data structure is robust enough to facilitate factor analysis and is well-suited for building a trustworthy linear regression model.

DOI: 10.54254/2754-1169/141/2024.GA18866

Table 2: The outcomes of the KMO Test and Bartlett's Test of Sphericity

| KMO Test and Bartlett's Test of Sphericity | | |
|---|---|---|
| KMO values | | 0.845 |
| Bartlett's Test of Sphericity | $\chi^2$ | 4296.29 |
| | p-value | 0.000 |

## 2.3. Common Factors

### 2.3.1. Determining Key Factors with Principal Component Analysis

By identifying common components that contribute to the majority of the dataset's variation, Principal Component Analysis (PCA) lowers the dimensionality of the data. This technique aids in the simplification of complicated data while maintaining important details. PCA works by linearly combining the original variables into new, reduced variables known as components, where the first principal component accounts for the biggest possible variance among the sample. F1 to F12 are the principal components (or factors) formed by linearly combining the original variables. The eigenvalues of the correlation matrix are computed using the factor loading matrix obtained from this procedure, and these values aid in determining the optimal number of components to extract [5]. The percentage of variance and cumulative variation explained by these components help identify the most significant elements. The eigenvalues and the cumulative variance are shown in Table 3.

Table 3: Results of the Factor Analysis

| Component | Eigenvalues | Variance | Cumulative Variance (%) |
|---|---|---|---|
| F1 | 8.5539 | 0.7128 | 71.28 |
| F2 | 1.4879 | 0.1240 | 83.68 |
| F3 | 0.8303 | 0.0692 | 90.60 |
| F4 | 0.6668 | 0.0556 | 96.16 |
| F5 | 0.2501 | 0.0208 | 98.24 |
| F6 | 0.1034 | 0.0086 | 99.10 |
| F7 | 0.0434 | 0.0036 | 99.47 |
| F8 | 0.0292 | 0.0024 | 99.71 |
| F9 | 0.0228 | 0.0019 | 99.90 |
| F10 | 0.0051 | 0.0004 | 99.94 |
| F11 | 0.0038 | 0.0003 | 99.98 |
| F12 | 0.0030 | 0.0002 | 1.00 |

Table 3 reveals that F1 and F2 were identified as being extracted from the initial solution, as indicated by the eigenvalues greater than 1. F1 and F2 account for 83.68% of the cumulative variance explained, meaning that they play a major role in explaining the variance within the initial collection of 12 variables. In other words, these two factors are capable of capturing most of the information from the original variables.

### 2.3.2. Varimax Rotation

Kaiser developed Varimax in 1958, and it's one of the most popular rotation techniques in factor analysis. With only a few big loadings and most loadings being near zero, each factor should have a basic structure according to the Varimax rotation. This method makes it easier to analyze the factors

DOI: 10.54254/2754-1169/141/2024.GA18866

since each variable is frequently related to one or a small number of factors, and each factor is linked to a limited number of variables. Furthermore, Varimax frequently permits factor interpretation according to the difference between variables with positive and negative loadings [6]. Varimax looks for a way to rotate the original factors so that the loadings' variance is maximized.

$$V = \sum \left( q_{j,l}^2 - \bar{q}_{j,l}^2 \right)^2 \quad (3)$$

The squared loading of the jth variable on the l factor is denoted by $q_{j,l}^2$, whereas the mean of the squared loadings for that factor is represented by $\bar{q}_{j,l}^2$. The factors are easier to understand because of the sparse loading matrix that the rotation produces [7].

Table 4: Rotated Component Matrix

| Component | Rotated Component Matrix | |
|---|---|---|
| | F1 | F2 |
| $Z_{X1}$ | 0.9210 | 0.3829 |
| $Z_{X2}$ | 0.8830 | 0.4692 |
| $Z_{X3}$ | 0.9069 | 0.3673 |
| $Z_{X4}$ | -0.9632 | 0.3292 |
| $Z_{X5}$ | 0.8641 | 0.4928 |
| $Z_{X6}$ | -0.4230 | -0.6169 |
| $Z_{X7}$ | 0.9035 | 0.3553 |
| $Z_{X8}$ | -0.3153 | -0.4622 |
| $Z_{X9}$ | -0.5893 | -0.4542 |
| $Z_{X10}$ | -0.8463 | -0.3128 |
| $Z_{X11}$ | 0.8877 | 0.4139 |
| $Z_{X12}$ | 0.0093 | 0.6940 |

Z-scores are used to show how much and in which direction each variable contributed to the two components that were taken out of the factor analysis. A negative Z-score indicates a strong negative effect, whereas a high positive Z-score indicates that the variable has a significant positive influence on the component. As Table 4 illustrates, GDP per capita (X1) has a strong positive Z-score for F1, showing a significant link with economic advancement, in contrast to Loan Rates for More Than Five Years (X4), which has a negative Z-score for F1 and may imply financial strain or limits.

Overall, factors like GDP per capita and the Cumulative CPI have significantly positive contributions, indicating that F1 likely indicates features connected to economic progress and investments. On the other hand, factors such as the US dollar to yuan exchange rate and the 3-year Treasury yield have significant contributions to F2, suggesting that it captures the dynamics of the financial markets and monetary policy.

### 2.3.3. Cumulative Variance

In PCA, cumulative variance is the total amount of variance explained by a collection of principal components. It is computed by adding together the individual variations that are explained by each constituent up to a predetermined amount [8]. The percentage of variance for each component and the cumulative variance can be expressed as follows:

DOI: 10.54254/2754-1169/141/2024.GA18866

$$\text{Percentage Var}_i = \frac{\lambda_i}{\sum \lambda_i} \times 100 \quad (4)$$

$$\text{Cumulative Var}_n = \sum_{1}^{n} \text{Percentage Var}_i \quad (5)$$

Where $\lambda_i$ is the eigenvalue associated with that component, the cumulative variance $\text{Var}_n$ for n components is the sum of the percentages of variance explained by these n components.

The main objective of PCA's cumulative variance is to ascertain how many principal components a data set should keep. Usually, a threshold (e.g., 80%) is selected, and components are kept in place until the total variance approaches or surpass this threshold. By reducing the data set and enhancing interpretability without compromising vital information, this strategy reduces data dimensionality while retaining the bulk of the information.

Table 5: Cumulative Variance

| Component | Variance | Proportional Variance | Cumulative Variance |
|---|---|---|---|
| F1 | 7.070995 | 0.589250 | 0.589250 |
| F2 | 2.530394 | 0.210866 | 0.800116 |

As seen in Table 5, F1 and F2 account for approximately 80.01% of the overall variance in the data. The figure suggests that these two components adequately represent the majority of the variability found in the dataset, allowing them to reflect the data's underlying structure on a lower dimensionality. Keeping these two elements would probably result in a more basic model that nonetheless captures most of the data.

### 2.3.4. Factor Scores

The values for every observation on the extracted factors are represented as factor scores, which helps to further comprehend the connection between the individual data points as well as the underlying latent variables that the factors represent.

Table 6: Contributions of Variables to Extracted Factors

| Variable | F1 | F2 |
|---|---|---|
| X1 | 2.1798 | -2.2035 |
| X2 | 2.3787 | -2.5497 |
| X3 | -1.5395 | 2.1775 |
| X4 | -1.5005 | 3.4875 |
| X5 | -1.9578 | 7.1008 |
| X6 | -0.1511 | 1.4554 |
| X7 | 0.0270 | 0.1553 |
| X8 | -0.2183 | 0.5308 |
| X9 | -0.0619 | 0.5323 |
| X10 | 0.9040 | -0.9277 |
| X11 | -0.9174 | -0.5298 |
| X12 | 0.4678 | -0.8767 |

DOI: 10.54254/2754-1169/141/2024.GA18866

Table 6 provides a detailed breakdown of how each variable contributed to the characteristics that were extracted. Each coefficient denotes the importance of a certain variable in creating the corresponding factor. For example, GDP per capita (X1) significantly contributes positively to F1, meaning that GDP per capita plays (X1) a key role in this factor. Conversely, its negative contribution to F2 implies a negative correlation with the attributes that characterize this factor. Similarly, Registered Population (X5) has a significant influence on F2, as evidenced by its strong positive coefficient.

## 3. Multiple Linear Regression

In order to analyze the relationships between the factors and housing prices, a multiple linear regression analysis was performed with the scores of the two recovered factors, F1 and F2, as independent variables.

Table 7: Summary of Multiple Linear Regression Results

| Metrics | Values |
|---|---|
| Mean Squared Error (MSE) | 2,250,961.74 |
| $R^2$ Score | 0.9454 |
| Adjusted $R^2$ Score | 0.9416 |

Table 8: Regression Coefficients and Statistical Significance

| Variable | Coefficient | Standard Error | t-statistic | P-value | 95% Confidence Interval |
|---|---|---|---|---|---|
| Intercept | 15,489.79 | 175.875 | 88.073 | 0.000 | [15,100, 15,800] |
| F1 | 5271.58 | 175.178 | 30.093 | 0.000 | [4924.77, 5618.39] |
| F2 | 1708.73 | 136.620 | 12.507 | 0.000 | [1438.25, 1979.20] |

Table 7 illustrates the linear regression model's overall performance indicators, which indicate a satisfactory match with the data. With an $R^2$ score of 0.9454, the model's components can account for 94.54% of the variation in house prices. The model's robustness is further confirmed by the Adjusted $R^2$ score of 0.9416. The model's accuracy in predicting house prices is further demonstrated by the comparatively low Mean Squared Error.

Table 8 presents the regression coefficients and their statistical significance. The statistical significance and positive correlations for F1 and F2 suggest a correlation between the augmentation of these variables and the rise in home prices. Compared to F2, F1 appears to have a greater impact on house prices, as evidenced by its larger coefficient. Both factors' low p-values and tight confidence intervals support the notion that they are significant predictors in the model.

Based on the model's regression results, conclusions can be drawn: For every one-unit increase in F1, the average housing price P increases by 5271.58 units. This suggests that housing sales prices generally rise in line with economic progress. Population expansion, rising housing demand, and rising income levels are all factors that often drive the real estate market and drive up prices. Furthermore, the average housing price rises by 1708.73 units for every unit increase in F2, indicating that changes in financial indicators or price indices also raise real estate values in Shanghai. The decision model for Shanghai's average housing price can be formulated as follows:

$$P = 5271.58 \times F1 + 1708.73 \times F2 + 15489.79 \quad (6)$$

DOI: 10.54254/2754-1169/141/2024.GA18866

## 4. Limitations & Future Outlooks

This study successfully predicted housing prices in Shanghai using common component analysis and multiple linear regression. With an R2 value of 0.9454, this model is highly accurate in detecting the major factors that affect house prices, including GDP per capita, cumulative CPI, and financial indicators like loan rates. Key findings in this research were the important influence of economic advancement (F1) and financial market dynamics (F2). The combination of linear regression and common factor analysis has reduced the model's complexity without sacrificing its high level of predictive accuracy for property prices. This methodology presents a strong framework for comprehending the inducements of the real estate market in Shanghai.

Despite the model's good performance, some limits should be noted. Because the research was based on historical data, it might not have fully captured future market shocks or abrupt regulatory changes, such as those brought on by the COVID-19 epidemic or changes in the government's real estate policy. Furthermore, because the study only considered macroeconomic variables, it may have ignored microeconomic factors that also have a great impact on real estate prices, such as neighborhood-specific influences or personal preferences for dwelling. Research has shown that neighborhood qualities, such as the closeness of transit and retail services, as well as specific house characteristics like the number of bedrooms and bathrooms, strongly affect property values [9].

In order to increase model precision, future research can concentrate on broadening the area of analysis by adding micro-level data, such as trends in urban growth or housing preferences. Furthermore, by capturing nonlinear correlations among variables, the implementation of sophisticated machine learning techniques like random forest regression or neural networks may provide deeper insights and possibly boost forecast accuracy. Studies have shown that a simple neural network with just four delays and three hidden neurons was able to produce consistent performance with an average relative root mean square error (RRMSE) of 1% across one hundred cities across the training, validation, and testing stages [10]. Finally, taking into account other Chinese cities or regions for comparison could confirm the model's suitability for other real estate markets and provide more comprehensive economic insights.

## 5. Conclusion

In this work, common factor analysis and multiple linear regression are combined in order to forecast house prices in Shanghai. Through comprehensive data analysis from the CEInet Statistics Database, the major financial, demographic, and economic factors affecting housing prices were identified. Following the confirmation of the data's appropriateness for factor analysis by the KMO (score of 0.845) and Bartlett's tests ($\chi^2$= 4296.29, p = 0.000), two primary common factors that accounted for 80.01% of the variance were identified. A multiple linear regression model with these parameters included produced a high R2 score of 0.9454, suggesting significant predictive power. The model showed that property prices are highly impacted by economic advancement (F1) and financial market dynamics (F2), with F1 having a bigger impact. Although the costs of housing in Shanghai were correctly forecasted by this study, its ability to capture future shocks and microeconomic influences may be limited by its reliance on historical data and macroeconomic factors. Subsequent studies could improve model accuracy by adding micro-level data, housing preferences, urban growth trends, and sophisticated machine learning algorithms for wider regional applications.

## References

[1] *Chen, J., Qi, X., Lin, Z., & Wu, Y. (2022). Impact of Governments' Commitment to Housing Affordability Policy on People's Happiness: Evidence from China. Housing Policy Debate, 32(4-5), 622-641.*

DOI: 10.54254/2754-1169/141/2024.GA18866

[2] *Kong, J., & Kepili, E. I. B. Z. (2023). A Survey Analysis: The Current Real Estate Marketing Situation in the China Greater Bay Area in the Context of the COVID-19 Epidemic. Real Estate Management and Valuation, 31(3), 1-19.*

[3] *Shrestha, N. (2021). Factor Analysis as a Tool for Survey Analysis. American Journal of Applied Mathematics and Statistics, 9(1), 4-11.*

[4] *Effendi, M., Matore, E. M., Khairani, A. Z., & Adnan, R. (2019). Exploratory Factor Analysis (EFA) for Adversity Quotient (AQ) Instrument among Youth. Journal of Critical Reviews, 6(6), 234-242.*

[5] *Schreiber, J. B. (2021). Issues and Recommendations for Exploratory Factor Analysis and Principal Component Analysis. Research in Social and Administrative Pharmacy, 17(5), 1004-1011.*

[6] *Abdi, H. (2003). Factor Rotations in Factor Analyses. Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA, 792-795.*

[7] *Carrizosa, E., Guerrero, V., Romero Morales, D., & Satorra, A. (2020). Enhancing Interpretability in Factor Analysis by Means of Mathematical Optimization. Multivariate Behavioral Research, 55(5), 748-762.*

[8] *Kherif, F., & Latypova, A. (2020). Principal Component Analysis. In Machine Learning (pp. 209-225). Academic Press.*

[9] *Hong, J., Choi, H., & Kim, W. S. (2020). A House Price Valuation Based on the Random Forest Approach: the Mass Appraisal of Residential Property in South Korea. International Journal of Strategic Property Management, 24(3), 140-152.*

[10] *Xu, X., & Zhang, Y. (2021). House Price Forecasting with Neural Networks. Intelligent Systems with Applications, 12, 200052.*