

# AN ANALYSIS OF REAL ESTATE PRICES AND TRANSACTIONS USING KNN ALGORITHM AND DEEP LEARNING APPROACHES

N.Gowri Sankar <sup>1</sup>, Mr.S.Muni Kumar <sup>2</sup>

<sup>1</sup>student, Mca 2<sup>nd</sup> Year Kmmips, Tirupati, Affiliated To S.V. University, Tirupati, A.P, India

<sup>2</sup>Associate Professor, Dept Of Mca, Kmmips, Tirupati, Affiliated To S.V. University, Tirupati, A.P, India

\*\*\*

**ABSTRACT** - Research on the study of houses, condominiums and buildings in Taiwan's metropolitan areas continues to be an important area of research. In real estate forecasting and analysis, methods such as statistical analysis and questionnaire data collection are widely used. However, when multidimensional data is considered, these methods are time-consuming and inadequate. This study aimed to build a real estate forecasting model that can adapt to a changing environment. Data were collected from public government databases, the collected data were standardized for accurate clustering, an appropriate data clustering algorithm was applied to the standardized data, and cross-statistical analysis was performed to verify the adopted algorithm. We used a deep learning based on autoencoder algorithm to increase the accuracy of the clustering analysis. Double-bottom map particle swarm optimization (DBM-PSO) clustering algorithm was then used to determine the optimal clustering solution. Cluster analysis and deep learning were conducted on data collected from public websites to understand the factors that led to the sustained increase in housing prices in Taiwan over the past decade. The results of this study indicate that three key factors—the number of real estate transactions, the average unit price of real estate transactions, and the building material and construction index—significantly affected real estate prices in Taiwan. Our results could help researchers and governments to focus on specific aspects of real estate development without being influenced by other related factors. In addition, the relationships between real estate trends and the aforementioned three key factors were determined to obtain valuable information that can enable the Taiwanese government to regulate the property market and prevent excessive growth. The framework proposed in this paper allows researchers and governments to focus on specific aspects of real estate development without being influenced by other related factors, and provides a new mechanism for approaching real estate-related forecasting.

**Key Words:** Machine learning, real estate, particle swarm optimization algorithm, economy, autoencoder, deep learning.

## 1. INTRODUCTION

The expression “own the land, own the wealth” is highly common in Taiwanese culture, and most families in Taiwan

aim to buy a home. Land and permanent structures—such houses, apartments, buildings, and fences—are considered to be real estate. Real estate symbolizes security and stability and serves as a form of shelter. It has become a popular investment and a reliable hedge against inflation on a global scale. The state of the economy affects housing demand, with property values and rental rates rising with increases in inflation. Open databases enable researchers to access established data for analysis. The Federal Reserve Economic Data database (<https://fred.stlouisfed.org>) of the United States contains information on various housing indicators, including housing inventory, housing starts, home sales, the housing affordability index, the housing price index, housing transactions, the mortgage-lending index, commercial property prices, interest rates, the producer price index, and foreclosed properties. Real estate data are complex and unpredictable, and most of the real estate literature focuses on the trend in housing prices and housing price forecasting. Unit price, initial property registration, and number of rooms are essential aspects in real estate research. Unexpected uncertainties such as pandemic outbreaks, market volatility, and economic crises make real estate data unpredictable. In recent years, advanced machine learning algorithms have been developed to account for such uncertainties in the long-term forecasts. Voith et al. developed a model that decomposes the variation in office vacancy rates into market-specific, time specific, and random factors. An understanding of real estate trends is crucial for the prediction of future prices. Drawing on economics, sociology, psychology, and methodological and epistemological philosophy, Mooya et al. raised three fundamental questions about the measurement and magnitude of real estate prices.

## 2. LITERATURE REVIEW

Governments aim to regulate housing prices to prevent real estate speculation and ensure that housing access is equitable. Governments can set appropriate interest rates on loans and implement other policy measures to stabilize housing prices. According to the website of the Taiwanese Ministry of the Interior (<https://pip.moi.gov.tw/Eng/EP31.aspx>), housing prices in Taiwan increased between 2000 and 2022. The trend in Taiwan's transaction index between Q1 in 2000 and Q2 in 2022 is displayed in

Figure 1. Government policies influence the behaviour of Taiwan's real estate market. Therefore, the need for government intervention to maintain stability is evident from Taiwan's real estate history. The development of property prices during key events in Taiwan is illustrated in Figure 1. Real estate has become a crucial research area because of its importance to the public.

### 3. METHODS

The proposed clustering model is outlined in this section. First, a four-layer autoencoder network minimizes the reconstruction error and the distance between the data points and their respective clusters in the code layer. Second, the data representation in the data layer is improved through nonlinear mapping. This process consists of four main steps standardizing the data for various features and attributes, reducing the multidimensional data to a low-dimensional data through dimensionality reduction based on autoencoder approach, (performing clustering in the low-dimensional space, and analyzing the obtained clusters related to real estate markets.

#### 3.1. DATA STANDARDIZATION

Data were obtained from various sources, included multiple types of information, and described various characteristics and attributes. Therefore, we used mean normalization to standardize the ranges of the independent variables ( $x$ ). The original variable file can be found in the Supplementary Files (98index.xlsx).  $x' = \frac{x - \mu}{\max(x) - \min(x)}$  where  $\mu$  is the mean value of the independent variable.

#### 3.2. REDUCING THE HIGH-DIMENSIONAL MODELS USING AUTO-ENCODER

An autoencoder is a type of neural network that aims to learn a hidden representation of its input to then reconstruct the input. Consider a single-layer autoencoder network that comprises an encoder and a decoder. The encoder uses a nonlinear mapping function  $f(x)$  to map an input  $x_i$  to the input's hidden representation  $h_i$ . The activation function used in this study is the sigmoid function, which is defined as follows:  $h_i = f(x_i) = \frac{1}{1 + \exp(-(W(1)x_i + b(1)))}$  (2) where  $W(1)$  is the encoding weight matrix and  $b(1)$  is the bias vector of the encoding. If the activation function is linear, the operation of the one-layer autoencoding network is equivalent to principal component analysis. using the decoder function  $g$ , which is expressed as follows:  $x'_i = g(h_i) = \frac{1}{1 + \exp(-(W(2)h_i + b(2)))}$  (3) where  $W(2)$  is the decoding weight matrix and  $b(2)$  is the bias vector of the decoding. The condition  $W(2) = W(1)^T$  can act as a regularize and reduce the number of parameters to optimize. The optimal parameters of  $W$  and  $b$  have to be defined during the coding process. The goal of

an autoencoder is to minimize the error of the reconstruction process. The cost function shown in equation (4) can be solved to obtain these two parameters by minimizing  $\text{cost}(\theta)$ , since it is directly related to the parameters  $W$  and  $b$ .  $\text{cost}(\theta) = \frac{1}{N} \sum_{i=1}^N \|x_i - x'_i\|^2$  (4) where  $\theta = \{W(1), W(2), b(1), b(2)\}$ ,  $x'_i$  is calculated as Equation (3). A deep autoencoder network can be formed by stacking several one-layer autoencoders. Here, the hidden representation of a previous autoencoder is used as input for the next autoencoder. However, because the gradient descent algorithm used for training is sensitive to initial weights, deep autoencoders can be difficult to train. To overcome this problem, a pre-training algorithm is used, which learns more robust features than the gradient descent algorithm before fine-tuning the entire model. Restricted Boltzmann machines and denoising autoencoders are used to construct a model that can reconstruct inputs from their distorted versions. Denoising autoencoders can be considered stochastic versions of autoencoders. The difference between a denoising autoencoder and a conventional autoencoder is that a denoising autoencoder performs a stochastic corruption process. This process randomly sets some inputs to 0. A denoiser autoencoder is trained to reconstruct the undistorted version of the input from the distorted version.

#### 3.3 CLUSTERING THE LOW-DIMENSIONAL SPACE

Kennedy and Eberhart introduced the particle swarm optimization (PSO) algorithm. PSO is an efficient learning algorithm based on evolutionary computation that mimics the movement of flock of birds to simulate social behavior. Each candidate solution is treated as a particle in a swarm, and each particle's velocity can be adjusted in accordance with its experience and the information that it exchanges with other particles in the swarm. The PSO algorithm effectively finds optimal solutions in complex search spaces through interactions between individuals in a population of particles. problem and is represented as the set of all feasible solutions. The swarm consists of  $M$  particles, which are expressed as  $S = (X_1, X_2, \dots, X_M)$ . Second, each particle is initialized with a random solution and is represented as a point in a  $D$  dimensional space, where  $X_i = (x_{i1}, x_{i2}, \dots, x_{iD})$  and  $x \in (X_{\min}, X_{\max})^D$ . A particle's velocity is represented as  $v_i = \{v_{i1}, v_{i2}, \dots, v_{iD}\}$ , where  $v \in (V_{\min}, V_{\max})^D$ . Third, each particle finds its best position by comparing its current position's fitness with that of its previous best position, which is represented as  $p_{\text{best}} = \{p_{\text{best}1}, p_{\text{best}2}, \dots, p_{\text{best}D}\}$ .

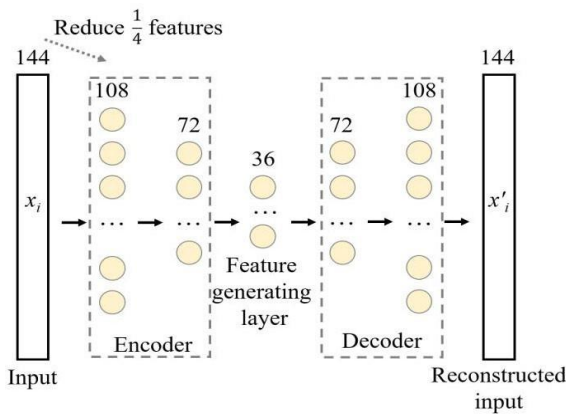


FIGURE 1. Architecture of auto encoder

#### 4. PROBLEM DEFINITION

The terms waxy and win, which represent the upper and lower limits of w, respectively, are set to 0.9 and 0.4, respectively. The terms it elation and iterations max represent the current number of iterations and the total number of iterations, respectively. The velocity vector indicates how the particle moves to reach the global best position at a given instant. However, a particle' exploration of the solution space can be limited by velocity values that are excessively low or high. The particle's velocity and position are thus limited to within  $[V_{in}, V_{max}]D$  and  $[X_{ina}, X_{mas}]D$ , respectively. These ranges define the maximum step size that the particle can use in the search DBMPSO is a Varian to Fops that uses a double bottom map (DBM) in its updating function. ADBM is a chaotic function that prevents the population becoming trapped in a local optimum, and it balances exploitation and exploration in PSO search by generating random values in three regions: regions with values of 0.0, 0.5, and 1.0. In the PSO updating function, r1 is related to updating the particle in accordance with best, and r2 is related to updating the particle in accord dance with best. In DBMPSO, the DBM function generates two sequences, namely DBMr1 and DBMr2, which replace r1 and r2, respectively, to balance global exploration and local search. The sequences DBMr1 and DBMr2 are expressed as follows:  $DBMr_{t+1, i, d} =$  where  $DBMr_{t+1} \sin 4\pi DB Mart i, d +1 2$  (8) and is the Damra sequence in the ditch element of the ith particle in the current iteration t. The update function of DBM-PSO is expressed as follows:  $v_{t+1 id} =$  wave  $i, d + c1 \times$  Debert  $1, i, d \times$  bested  $- xt i, d + c2 \times$  Debert  $2, i, d \times$  best  $- xt i$ , After the fine-tuning of the stacked autoencoder, it has learned the parameters  $\theta$ . Subsequently, low-dimensional codes can be extracted from the code layer, and the DBM-PSO algorithm is used to cluster these codes. Algorithm 1 presents the pseudocode of the DBM-PSO clues tiring algorithm. The particle encoding, fitness evaluation, and best and best updates are defined for a specific clues tiring problem.

#### 5. PROPOSED METHOD

We build a framework based clustering and deep learning to identify features of real estate. An autoencoder algorithm was employed to increase the clustering accuracy, and DBM-PSO clustering algorithm used to determine the optimal clustering solution.

2. The results of this study could serve as a reference for governments and real estate investors in developing policies, regulations, and market strategies as well as forecasting trends accurately to maintain the relationships between real estate characteristics.

The clustering results were statistically analyzed. IBM SPSS Statistics was used to perform descriptive statistical analysis, normality tests, analysis of variance (ANOVA), and linear regression. The descriptive statistics were the mean, max mum, minimum, variance, standard deviation, coefficient of skewness, and coefficient of kurtosis. The skew ness and kurtosis coefficients were employed to determine the state of deviation from normal distributions. ANOVA is appropriate when the response variables have continuous distributions and the conditions are discrete, either by nature or by design. The ANOVA results were evaluated on the basis of R2, adjusted R2, sum of squares, mean square, F value, and significance level ( $\alpha$ ). R2 represents the proportion of variance in the dependent variable that the independent variable can explain; a higher R2 value thus indicates higher explanatory power. The adjusted R2 value is a modified version of R2 that considers additional indecent dent variables that tend to bias R2 results. The mean square is the estimated variance used to determine the accuracy of the regression model, and the sum of squares is the sum of the squares of the distances between each data point and the line of best fit. The F value is used to compare the amount of variance explained by the model with the amount of error or unexplained variance to determine the test's significance. (13) The results of the linear regression included the standard Izod regression coefficient  $\beta$ , T value, tolerance, variance inflation factor (VIF), Pearson correlation coefficient, and P value. The Average value of characters in clustering results related to number of real estate transactions. standardized regression efficient is used to compare the explanatory power of the independent variables and calculate the slope of the regression equation. Moreover, the T value indicates whether a significant difference exists between the variables, with a higher T value indicating stronger evidence. The VIF is used to assess the independence of the independent variables in multiple linear regressions, and a small VIF value indicates a high degree of independence. The Pearson correlation coefficient is employed to test the linear relationship between two variables. Regression analysis was conducted to determine which relationships in the developed model had statistically significant coefficients and Pavlus Descriptive statistical analysis,

normality tests, ANOVA, and linear regression were performed using IBM SPSS Statistics. After data clustering, the descriptive statistics were analyzed (Table S3). The high average R<sup>2</sup> value of the aforementioned 13 independent variables (0.8822) in the ANOVA indicated that these variables had high power to explain the building material and construction index (Table 3). The regression analysis results for the aforementioned 13 variables are shown in Figure indices were extracted using techniques such as data standardization, high-dimensional model reduction by using autoencoders, and data clustering.

### 5.1. FEATURE EXTRACTION AND ENGINEERING

We attempted to identify the factors that influence home prices by analyzing indices that are strongly correlated with the average unit price of real estate transactions. The unit price of a home is a critical consideration for prospective buyers, and understanding the

### 6. CONCLUSION

In this study, cluster analysis and deep-learning techniques were used to identify the key factors influencing Taiwan's sustained housing price increases over the past decade. Data collection, data standardization, autoencoder-based dimensionality reduction, and data clustering were performed in this study, and three data clusters were obtained. The first cluster contained independent variables that influence the number of real estate transactions; the second cluster contained independent variables that influence the average unit price of real estate transactions; and the third cluster comprised independent variable indices that affect the building materials and construction index when controlling for relevant variables. The framework proposed in this paper provides a new mechanism for approaching real-estate-related forecasting, thereby enabling researchers and governments to focus on specific aspects of real estate development without being influenced by other related factors. Future research should focus on expanding data collection efforts, exploring different clustering algorithms for subgroup analyses, and optimizing parameter settings. Our finding might be useful for governmental regulation of the relationship between the rental market and the real estate market; thus, considering people's housing equity when formulating real-estate-related laws and regulations is essential. Although this study only focused on Taiwan, the proposed framework can be applied globally to analyse housing markets in different regions.

### 7. REFERENCES

[1] J. Ratcliffe, M. Stubbs, and M. Keeping, *Urban Planning and Real Estate Development*. New York, NY, USA: Routledge, 2021.

[2] Y. Kang, F. Zhang, W. Peng, S. Gao, J. Rao, F. Duarte, and C. Ratti, "Understanding house price appreciation using multi-source big geo data and machine learning," *Land Use Policy*, vol. 111, Dec. 2021, Art. no. 104919.

[3] P.-F. Pai and W.-C. Wang, "Using machine learning models and actual transaction data for predicting real estate prices," *Appl. Sci.*, vol. 10, no. 17, p. 5832, Aug. 2020.

[4] O. Tatsey and C. Tariqul, "A machine learning-based 10 years ahead prediction of departing foreign visitors by reasons: A case on Türkiye," *Appl. Sci.*, vol. 12, no. 21, p. 11163, Nov. 2022.

[5] R. Voith and T. Crone, "National vacancy rates and the persistence of shocks in U.S. office markets," *Real Estate Econ.*, vol. 16, no. 4, pp. 437–458, Dec. 1988.

[6] X. Wang, J. Wen, Y. Zhang, and Y. Wang, "Real estate price forecasting based on SVM optimized by PSO," *Optik*, vol. 125, no. 3, pp. 1439–1443, Feb. 2014.

[7] Mummery, *Real Estate Valuation Theory: A Critical Appraisal*, 1st ed., Berlin, Germany: Springer, 2016.

[8] H. Usman, M. Lizam, and M. U. Adekunle, "Property price modelling, market segmentation and sub market classifications: A review," *Real Estate Manage. Valuation*, vol. 28, no. 3, pp. 24–35, Sep. 2020.

[9] J. Sun, X. Yang, and X. Zhao, "Understanding commercial real estate indices," *J. Real Estate Portfolio Manage.*, vol. 18, no. 3, pp. 289–303, Jan. 2012.