Housing prices prediction with deep learning: an application for the real estate market in Taiwan

Choujun Zhan
‡*School of Information Science and Technology,*
Xiamen University Tan Kah Kee College, Fujian 363105, China

Yonglin Liu
School of Electrical and Computer Engineering
Nanfang College of Sun Yat-sen University
Guangzhou, China
nfsysuliuyonglin@gmail.com

Wangling Chen
School of Electrical and Computer Engineering
Nanfang College of Sun Yat-sen University
Guangdong 510970, China

Zeqiong Wu
School of Electrical and Computer Engineering
Nanfang College of Sun Yat-sen University
Guangzhou, China
wuzeqiong@gmail.com

Zefeng Xie
School of Electrical and Computer Engineering
Nanfang College of Sun Yat-sen University
Guangdong 510970, China

2020 IEEE 18th International Conference on Industrial Informatics (INDIN) | 978-1-7281-4964-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/INDIN45582.2020.9442244

***Abstract*—The housing market is increasing huge, predicting housing prices is not only important for a business issue, but also for people. However, housing price fluctuations have a lot of influencing factors. Also, there is a non-linear relationship between housing prices and housing factors. Most econometric or statistical models cannot capture non-linear relationships yet. Therefore, we propose housing price prediction models based on deep learning methods, which can capture non-linear relationships. In this work, we construct a dataset, including the housing attributes data and macroeconomic data in Taiwan from January 2013 to December 2018. The housing attributes data includes two types of housing transactions, which are "land + building" (Type1) and " land + building + park" (Type2). Macroeconomic data includes housing investment demand ratio, owner-occupier housing ratio, housing price to income ratio, housing loan burden ratio, and housing bargaining space ratio. Then, this dataset is utilized to evaluate the prediction methods based on deep learning algorithms BPNN and CNN to predict housing prices. Experimental results show that CNN with housing features has the best prediction effect. This study can be used to develop targetted interventions aimed at the housing market.**

***Index Terms*—Deep learning, housing prices prediction, time series**

I. Introduction

In modern society, the housing market is one of the most parts of society and the national economy. A number of studies reveals that the housing market highly associates with monetary policy [1], [2], social networks [3], [4], labor market [5], [6], stock market [7], [8], investment and consumption [9], [10]. Accurately predicting housing prices plays an important role in the housing market. Accurately forecasting housing prices can help the government stabilize the real estate market and avoid large fluctuations in housing prices.

The main obstacle in housing prices prediction is that the real estate market is influenced by many factors, including macroeconomic and its own value. The house itself is both investment and consumption goods, while The purpose of investment or consumption determines whether the house will be used for living or renting [11]. Some scholars investigate the relationship between investment demand, owner-occupier demand, and housing prices [12]. In 2001, researchers argue that bank loans had a greater impact on housing prices forecasting than interest rates [7], while mortgages would result in greater affordability for housing owners [13]. In 2016, Zhang et al. investigated the impact of income inequality in China's cities on the price-to-income ratio and housing vacancy rate [14]. Scholars have investigated the relative importance of the real estate market bubble in terms of increasing investment demand and the ratio of house prices to income [12] [15]. In a strong market, a limited supply of housing will give sellers more bargaining space [16], while bargaining space can also influence the housing prices [17].

The subprime mortgage crisis almost destroyed the world financial system in 2007-2008, leading to housing prices fell sharply, triggering the global financial crisis. However, the Taiwan housing market is nearly not influenced by this crisis, proving that the Taiwan housing market is highly resistant to the financial crisis. Therefore, we choose Taiwan as the research object to predict house prices. For time series prediction, such as oil price forecast [18], Residential load forecast [19], stock price forecast [20] and traffic flow forecast [21] Etc., They used deep learning methods to build predictive models and found that deep learning methods have better results for time-series predictions. Fu et al. proved deep learning models (LSTM and GRU) in research on predicting

978-1-7281-4964-6/20/$31.00 ©2020 IEEE

traffic flow Better than Econometric Model (ARIMA) [21]. In all the studies reviewed here, deep learning methods have certain advantages in time series prediction. There remain deep learning methods in housing prices prediction about which relatively little is adopt.Therefore, this study builds a time series dataset, using deep learning methods (BPNN, CNN) to predict housing prices.

In this work, we adopt the real estate market in Taiwan to analyze whether macroeconomic factors can help predict house prices. The housing attributes and macro-economic variables of this study are derived from the open data of the real estate transactions released by the Taiwan Ministry of the Interior and the Taiwan Economic Journal [22]. First, we adopt transaction data on two types of houses ("land + building" and "land + building + park") in 9 administrative regions of Taiwan from January 2013 to December 2018. The housing transaction information of each administrative region has several different attributes, including the age of the house (Age), monthly house sales (Deals), land transfer area (Land Area), building transfer area (Building Area), and etc. Additionally, another five macroeconomic variables are adopted, namely, the ratio of housing investment demand, the ratio of housing owner-occupier demand, the ratio of housing price to income, the ratio of housing loan burden and the ratio of housing bargaining space.

Then, we consider two different class of cases: (1) only housing attributes are adopted as features for training prediction model; (2) macroeconomic data is utilized as additional information to train the model. Then, we use historical data of the last three, five, or six months to predict to predict house prices in an administrative region for the next month. Hence, there are total $2\times6 = 12$ different cases. We employed Pearson correlation coefficient to analyze the the correlation between the training features (housing attributes and macroeconomic information) and the housing price before establishing the housing price model. Finally, deep learning (BPNN, CNN) algorithms are used to build models for predicting housing prices, and six evaluation indicators (RMSE, MAE, MAPE, R^2, $R^2_{adjusted}$, RMSLE) are used to evaluate the model. The results show that the CNN of the deep learning model has the highest prediction accuracy. In addition, it turns out that five-month training dataset is suitable for predicting housing prices.

II. Data description

A. Data source

The data about housing attribution and macroeconomic information utilized in this study are derived from the open data of the real estate transactions of the Taiwan Ministry [22]. The real estate transactions website was officially launched in August 2012, announcing three aspects of house transaction signs, price information and sign information, while the Taiwan Economic Herald was established in April 1990 and release nine financial information listed on financial markets. We choose historical transaction data of two type of house in Twain released from January 2013 to December 2018. There are a total of 865,242 records of the transaction signs in the dataset. There are two main types of these house transactions: type-1 transactions include the transaction of houses and land itself, namely, "land + building"; type-2 transaction includes transactions of housing land, building and park, namely, "land + building +park".

There are a total of 29 housing attributes including monthly house sales, house age, land area, building area, park area, number of rooms, number of bathrooms, the use zoning ratio of life, the use zoning ratio of business, the ratio of having managed organization, the present building situation pattern-the ratio of having compartmented, the total housing price, the housing unit price, etc. Additionally, five macroeconomic variables, including housing investment demand ratio, owner-occupier housing ratio, housing price to income ratio, housing loan burden ratio, and housing bargaining space ratio, also influence the housing price.

Tier and region in Taiwan: Six municipalities (New north city , Kaohsiung City, Taichung City, Taipei City, Taoyuan City, Tainan City) and three towns (Keelung City, Chiayi city, Hsinchu city) in Taiwan are adopted and divided into three Tier (shown in the TableI). Finally, the region within the jurisdiction that obey the normal distribution were finally selected as experimental data. In this work, we just adopt regions, in which the housing price approximately follows the normal distribution, for developing a prediction algorithm. For example, FigureII-A shows the distribution of house prices from 2013 to 2018 Xinyi District in Taipei and Dashe District in Kaohsiung, respectively. Note that the distribution of house prices in Taipei approximately follows a normal distribution, while the price distribution of houses in Dashe District of Kaohsiung is far from a normal distribution.

B. Dataset construction

The experiments can be mainly classified into two cases. In case one, we only use housing attribute to predict the house price. In case two, we use housing attributes and macroeconomic data to predict housing price. For case one, we adopt historical data from the last k month to predict the price of next month, namely,

$$X_k = \{x(t-k), x(t-k+1)\cdots, x(t-1)\} \tag{1}$$

is utilized to prediction $x(t)$, where $x(t)$ represents the average housing price in a region and $k = 3, 5, 6$.

$$\begin{aligned} X_k = \{&x(t-k), x(t-k+1)\cdots, x(t-1),\\ &s(t-k), s(t-k+1), \cdots, s(t-1)\} \end{aligned} \tag{2}$$

where $s(t)$ is the macroeconomic information at time t. The results are constructed to verify whether adding macroeconomic variables is helpful in predicting housing prices and which time series span is more suitable for predicting housing prices.

III. Housing prediction models

First, we need to process the data to construct the training dataset. Here, we will normalize the data and use Principal component analysis (PCA) to access the main component.

TABLE I
TIER AND REGION WITHIN THE JURISDICTION

Tier	City	Region within the jurisdiction
1	New north city	Three Gorges District,Mie District, Zhonghe District,Tucheng District, Xindian District,Xinzhuang District, Itabashi District,Linkou District, Yonghe District,Xizhi District, Freshwater area,Luzhou District
1	Kaohsiung city	Sanmin District,Zuoying District, Nanzi District,Fengshan District, Gushan District
1	Taichung city	North District,Beitun District, South District, Nantun District, Dali District,Taiping District, West District, Xitun District
1	Taipei city	Zhongshan District, Zhongzheng District, Xinyi District,Neihu District, Beitou District,Nangang District, Shilin District, Datong District, Daan District,Wenshan District, Songshan District,Wanhua District
2	Taoyuan city	Zhongli District, Pingzhen District, Taoyuan District
2	Tainan city	Central and Western District, Rende District,North District, South District, Annan District, Anping District
3	Keelung city	Zhongzheng District,Xinyi District, Anle District,Warm District
3	Chiayi city	Chiayi city
3	Hsinchu city	Hsinchu city

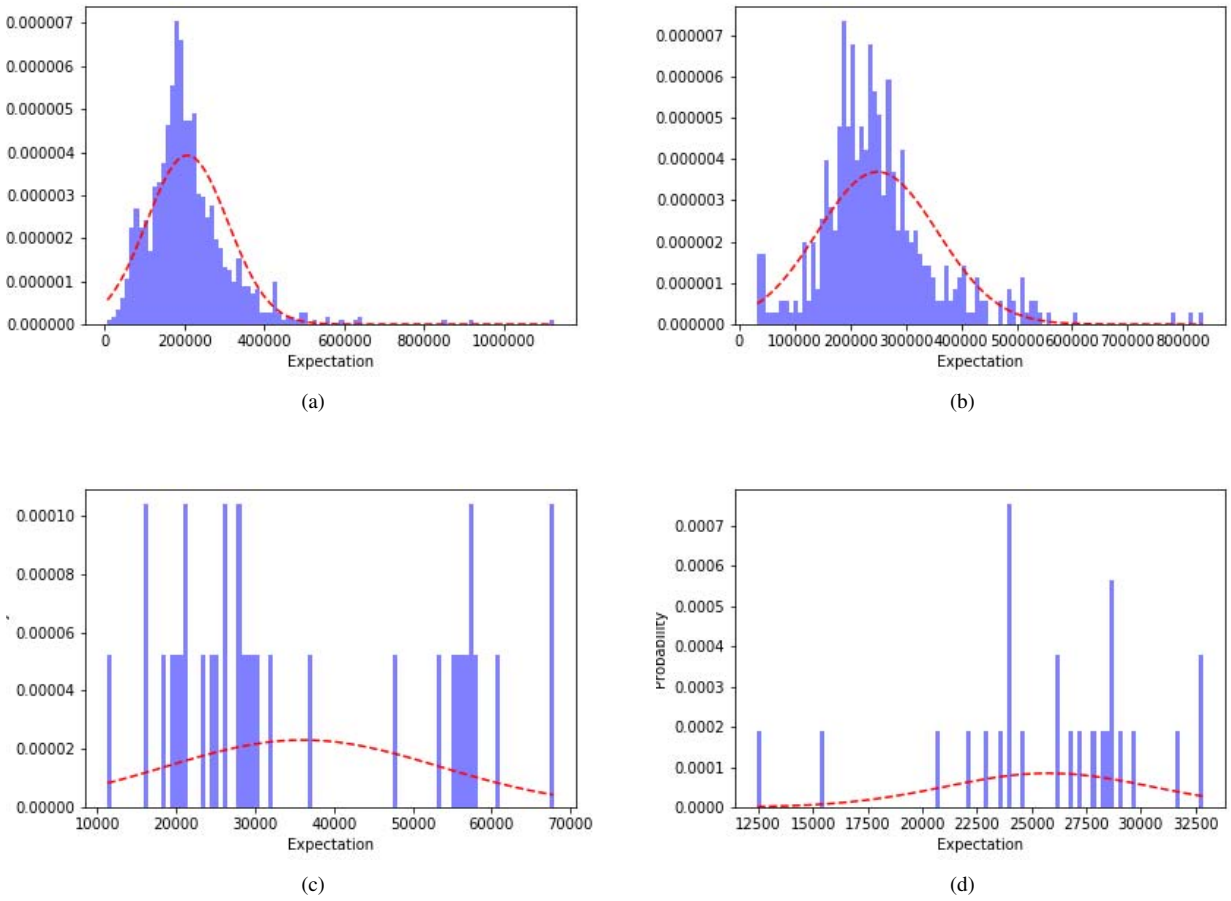

Fig. 1. (a) Distribution of house price (land+building) of Xinyi District in Taipei in 2015; (b) distribution of house price (land+building+park) of Xinyi District in Taipei in 2013; (c) distribution of housing price (land+building) of Dashe district in Kaohsiungcity in 2014; (d) distribution of housing price (land+building+park) of Dashe district in Kaohsiungcity in 2015.

- Note that the scale of different features is different. Hence, the higher-value feature will dominate and weaken the impact of lower-valued features of the model. Standardization can adjust housing attributes and macroeconomic records on different scales to a notionally common scale. The standardized formula is as follows:

$$x^{'} = \frac{x - \bar{x}}{\sigma} \tag{3}$$

where x is original feature vector, $\bar{x}$ is the mean of feature vector, σ is the standard deviation of feature vector.

- Principal component analysis (PCA) [23] is an unsupervised learning method that uses an orthogonal transforma-

tion to reduce high-dimension linearly related variables into low-dimension linearly independent variables, which are adopted as the main component. Since the number of principal components is less than the number of original variables, the principal component analysis belongs to the dimensionality reduction method. Dimension reduction plays an important role in the data preprocessing process. On the one hand, it can increase the sampling density of the data set samples, and on the other hand, it can remove the influence of noise in the data set. The formula for calculating the principal components a_k is as follows:

$$a_k = \arg\max var(a_k^T x) = a_k^T \sum a_k, \tag{4}$$

where a_k^T is unit vector, namely $a_k^T a_k = 1$.

After data processing, we adopt Back Propagation Neural Network (BPNN) [24] and Convolutional Neural Network (CNN) to predict housing price. We use the mean square error as the loss function of the neural network model,

$$E = \frac{1}{2}\sum_k (y_k - t_k)^2 \tag{5}$$

where y_k represents the predicted value of the neural network, t_k represents the true value, k represents the dimensions of the data.

In order to evaluate the predictive performance of the model, here we use 6 evaluation indicators. They are defined as follows:

- Root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y^i - \widehat{y}^i)^2} \tag{6}$$

- Mean absolute error(MAE):

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y^i - \widehat{y}^i| \tag{7}$$

- Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}|\frac{y^i - \widehat{y}^i}{y^i}| \times 100\% \tag{8}$$

- Root Mean Square Logarithmic Error (RMSLE):

$$RMSLE = \sqrt{\frac{1}{m}\sum_{i=1}^{n}(log(\widehat{y}^i + 1) - log(y^i + 1))^2} \tag{9}$$

- R-Square (R^2):

$$R^2 = 1 - \frac{\sum(y^i - \widehat{y}^i)^2}{\sum(y^i - \bar{y})^2} \tag{10}$$

- Adjusted R-Square ($R^2_{adjusted}$):

$$R^2_{adjusted} = 1 - [\frac{(1 - R^2)(m-1)}{m-k-1}] \tag{11}$$

where m is the number of samples, y^i is the ith true housing price of the samples, $\widehat{y}^i$ is the ith predicted housing price, $\bar{y}$ is the mean of the housing price, and k is the number of samples features. The smaller the scores of the three evaluation indicators of RMSE, MAE, and MAPE, the better the prediction performance of the model. RMSLE targets outliers in the sample. If the predicted value is lower than the actual value, the RMSLE score will be higher. The closer R^2 to 1, the better the degree of fit. If the useful feature is added to the model, $R^2_{adjusted}$ will be increase and vice versa.

IV. Experimental Results

This study aims to build a model for forecasting housing prices, using data from Taiwan real estate transactions from 2013 to 2018. We use housing features and macroeconomic variables as the training data. Here, we consider two cases. In case one, we only use the housing attributes, while in the other case, macroeconomic variables are adopted as a piece of additional information. Here, transaction data from 2013 to 2017 is used as the training dataset, while data in 2018 is used as the test dataset. Before establishing the housing prices model, we firstly measured the correlation between the features in the dataset and the housing prices.

It can be seen that: Type1-Bathroom, Type2-Age, and Type2-ParkArea correlation are low, and the rest are significantly related to the target housing prices. Among the macroeconomic variables, Housing price to income ratio and Housing loan burden ratio are highly related to the target variable, which indicates that these two macroeconomic variables have a greater impact on housing prices. Housing owner-occupier ratio, Housing investment demand ratio, and Housing bargaining space ratio have less influence on housing prices.

Deep learning models, including BPNN and CNN, are utilized to predict housing prices. Fig. 2 shows the prediction result based on 5-month historical data. We compare the results of BPNN and CNN and give detailed results in Table II. The parameter settings of these models and the proposed model are shown in Table III. Results show that the prediction based on 5-month historical data shows the best performance. However, adding macroeconomic information as additional features to the trained model does not significantly increase the performance. Therefore, the macroeconomic variables show little helpful in predicting housing prices. Compared to BPNN, we find that CNN performs better in this study. It can be seen from the Table II. This shows that R^2 is higher than 0.945, RMSE, MAE, MAPE, and RMSLE are lower in the CNN model.

V. Conclusion

The purpose of the current study was to predict housing prices based on deep learning algorithms. We constructed three-time series datasets to test which time series has a better effect on housing price prediction and compare whether adding macroeconomic variables is helpful for housing price prediction. The experimental result shows that the impact of macroeconomic variables on housing prices prediction is

TABLE II
THE STATISTICS OF FORECASTING PERFORMANCE COMPARISON

Dataset	Model	RMSE	MAE	MAPE(%)	R^2	$R^2_{adjusted}$	RMSLE
$S1x(t_3)$	BPNN	12604.22	7521.03	9.541666	0.947829	0.939898	0.128375
	CNN	12888.09	8698.426	10.36601	0.945452	0.93716	0.134916
$S1x(t_5)$	BPNN	12522.26	7903.027	9.529416	0.947068	0.914355	0.127867
	CNN	10051.99	6868.091	7.944236	0.965892	0.944812	0.106395
$S1x(t_6)$	BPNN	11744.68	7432.894	8.951377	0.95272	0.914684	0.119542
	CNN	10740.29	7585.151	9.022593	0.960461	0.928652	0.113167
$S2x(t_3)$	BPNN	13581.51	8642.252	8.807352	0.939425	0.930216	0.12125
	CNN	12660.09	8192.777	8.84016	0.947365	0.939364	0.121069
$S2x(t_5)$	BPNN	12183.96	7873.267	8.649615	0.949889	0.91892	0.120603
	CNN	11086.18	7354.108	8.113859	0.958513	0.932872	0.11077
$S2x(t_6)$	BPNN	13385.44	9089.131	12.18825	0.938587	0.889181	0.15497
	CNN	11582.12	8101.182	9.094282	0.954019	0.91703	0.122817

TABLE III
THE PARAMETERS SETTING OF MODELS

Dataset	Model	Learning Rate	Dropout	Hidden layer number	Neurons
$S1x(t_3)$	BPNN	[0.001 1e-05 1e-07]	0.5	2	[128 16]
	CNN	[0.01 0.0001 1e-06]	0	2	[64 16]
$S1x(t_5)$	BPNN	[0.001 1e-05 1e-07]	0.2	2	[32 16]
	CNN	[0.001 1e-05 1e-06]	0	3	[32 64 128]
$S1x(t_6)$	BPNN	[0.001 1e-05 1e-07]	0.5	2	[32 16]
	CNN	[0.01 0.0001 1e-06]	0.1	2	[64 32]
$S2x(t_3)$	BPNN	[0.001 1e-05 1e-07]	0.5	2	[64 32]
	CNN	[0.001 1e-05 1e-07]	0	3	[256 64 16]
$S2x(t_5)$	BPNN	[0.001 1e-05 1e-07]	0.5	2	[64 32]
	CNN	[0.001 1e-05 1e-07]	0	3	[256 64 16]
$S2x(t_6)$	BPNN	[0.001 1e-05 1e-07]	0.5	2	[128 16]
	CNN	[0.01 0.0001 1e-06]	0	3	[256 64 16]

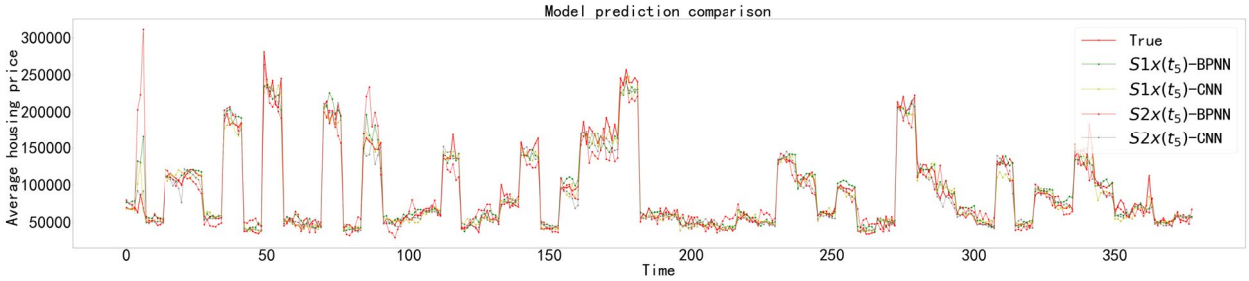

Fig. 2. The 5months time series prediction result

slightly smaller. In addition, our experiments confirmed that the five months time-series span is more suitable for predicting housing prices. For the proposed methods, CNN is considered to be the best model, R2 is higher than 0.945. Although CNN is usually applied in the field of image processing, this work confirms that it can also perform well in time series prediction. Therefore, our research shows that convolutional neural networks are suitable for regression prediction of housing prices and can effectively learn time-series information. Further research should be carried out to establish other state-of-art deep learning algorithms LSTM and GRU, to predict housing prices.

ACKNOWLEDGMENT

This work was supported by National Science Foundation of China Project 61703355, Science and Technology Program of Guangzhou, China 201904010224 and 201804010292.

REFERENCES

[1] M. Iacoviello and R. Minetti, "The credit channel of monetary policy: Evidence from the housing market," *Journal of Macroeconomics*, vol. 30, no. 1, pp. 69–96, 2008.

[2] R. Gupta, M. Jurgilas, and A. Kabundi, "The effect of monetary policy on real house price growth in south africa: A factor-augmented vector autoregression (favar) approach," *Economic modelling*, vol. 27, no. 1, pp. 315–323, 2010.

[3] D. Ettema, "A multi-agent model of urban processes: Modelling relocation processes and price setting in housing markets," *Computers, environment and urban systems*, vol. 35, no. 1, pp. 1–11, 2011.
[4] M. Bailey, R. Cao, T. Kuchler, and J. Stroebel, "The economic effects of social networks: Evidence from the housing market," *Journal of Political Economy*, vol. 126, no. 6, pp. 2224–2276, 2018.
[5] J. S. Zax, "Compensation for commutes in labor and housing markets," *Journal of urban Economics*, vol. 30, no. 2, pp. 192–207, 1991.
[6] G. Johnes and T. Hyclak, "House prices and regional labor markets," *The Annals of Regional Science*, vol. 33, no. 1, pp. 33–49, 1999.
[7] N.-K. Chen, "Asset price fluctuations in taiwan: Evidence from stock and real estate prices 1973 to 1992," *Journal of Asian Economics*, vol. 12, no. 2, pp. 215–232, 2001.
[8] L. Boone and N. Girouard, "The stock market, the housing market and consumer behaviour," *OECD economic studies*, vol. 2002, no. 2, pp. 175–200, 2003.
[9] E. A. Hanushek and J. M. Quigley, "The dynamics of the housing market: A stock adjustment model of housing consumption," *Journal of Urban Economics*, vol. 6, no. 1, pp. 90–111, 1979.
[10] J. Henneberry, "Transport investment and house prices," *Journal of Property Valuation and Investment*, 1998.
[11] C.-C. Lin, S.-J. Lin *et al.*, "An estimation of elasticities of consumption demand and investment demand for owner-occupied housing in taiwan: a two-period model," *International Real Estate Review*, vol. 2, no. 1, pp. 110–125, 1999.
[12] M.-C. Chen, C.-O. Chang, C.-Y. Yang, and B.-M. Hsieh, "Investment demand and housing prices in an emerging economy," *Journal of Real Estate Research*, vol. 34, no. 3, pp. 345–373, 2012.
[13] Z. A. Hashim *et al.*, "House price and affordability in housing in malaysia," *Akademika*, vol. 78, no. 1, 2010.
[14] C. Zhang, S. Jia, and R. Yang, "Housing affordability and housing vacancy in china: The role of income inequality," *Journal of housing Economics*, vol. 33, pp. 4–14, 2016.
[15] Y. Hu and L. Oxley, "Bubbles in us regional house prices: evidence from house price–income ratios at the state level," *Applied Economics*, vol. 50, no. 29, pp. 3196–3229, 2018.
[16] S. C. Bourassa, D. R. Haurin, J. L. Haurin, M. Hoesli, and J. Sun, "House price changes and idiosyncratic risk: the impact of property characteristics," *Real Estate Economics*, vol. 37, no. 2, pp. 259–278, 2009.
[17] J. P. Harding, J. R. Knight, and C. Sirmans, "Estimating bargaining effects in hedonic models: Evidence from the housing market," *Real estate economics*, vol. 31, no. 4, pp. 601–622, 2003.
[18] Y.-X. Wu, Q.-B. Wu, and J.-Q. Zhu, "Improved eemd-based crude oil price forecasting using lstm networks," *Physica A: Statistical Mechanics and its Applications*, vol. 516, pp. 114–124, 2019.
[19] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on lstm recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2017.
[20] A. Borovykh, S. Bohte, and C. W. Oosterlee, "Conditional time series forecasting with convolutional neural networks," *arXiv preprint arXiv:1703.04691*, 2017.
[21] R. Fu, Z. Zhang, and L. Li, "Using lstm and gru neural network methods for traffic flow prediction," in *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*. IEEE, 2016, pp. 324–328.
[22] I. ministry. (2020) Real estate transaction net price inquiry service website. [Online]. Available: https://lvr.land.moi.gov.tw/homePage.action
[23] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.