Global Journal of Business Research
Vol. 16, No. 1, 2022, pp. 1-16
ISSN: 1931-0277 (print)
ISSN: 2157-0191 (online)

IBFR
www.theIBFR.com

REAL ESTATE TREND PREDICTION USING LINEAR REGRESSION AND ARTIFICIAL NEURAL NETWORK TECHNIQUES

Sophia L. Zhou, Worcester Polytechnic Institute

ABSTRACT

An accurate assessment of future housing prices is crucial to critical decisions in resource allocation, policy formation, and investment strategies. In this work, linear regression and artificial neural network were employed to model home price indices, using datasets of the S&P/Case-Shiller home price index and twelve demographic and macroeconomic features in five metropolitan statistical areas: Boston, Dallas, New York, Chicago, and San Francisco. The data, ranging from March 2005 to December 2018, were collected from the Federal Reserve Bank, the Federal Bureau of Investigation, Macrotrends, and Freddie Mac. Three time-lagging situations were compared: no lag, a 6-month lag, and a 12-month lag. Since some data were available monthly, some quarterly, and some annually, two methods to compensate missing values, backfill and interpolation, were compared. The models were evaluated for accuracy and mean absolute error. The results showed that linear regression performed well in predicting long-term trends, while artificial neural network was suitable for short-term prediction. It was found that input factors that were statistically significant varied in different areas. The results also showed that the technique to compensate missing values and the implementation of time-lag influenced the models' performances, both of which require further investigation.

JEL: R310

KEY WORDS: Housing Price Index Prediction, Linear Regression, Artificial Intelligence, Random Forest, and Linear Regression

INTRODUCTION

The housing market is the sale and acquisition of real estate for residential or commercial purposes (Bank of England, n.d.). In the United States (US), the real estate industry accounts for 832,000 jobs and 15% of the GDP in 2018 (Stupak, 2019). As of 2020, the US housing market is worth around $33.6 trillion, a market value equal to the annual GDP of the US and China combined (Gerrity, 2020). Individual buyers, investors, businesses, and governments are all affected by the housing market. For all the stakeholders, confident assessments of the housing market in the near- and long-term future are important for critical decisions about policy making, asset allocation, and portfolio and investment management (Conway, 2018, Lyons, 2017). There is a need to develop a model that will accurately predict future housing market trends.

In this study, the S&P/Case-Shiller Home Price Indices (HPI) (S&P Dow Jones Indices LLC, n.d.), the leading measure of US residential real estate prices, were modeled by linear regression (LR) and artificial neural network (ANN) methods, using twelve demographic and macroeconomic features in five metropolitan statistical areas (MSA): Boston, Dallas, New York, Chicago, and San Francisco. The selected five areas represented different market trends. This paper adds to the knowledge about the influence of macroeconomic and demographic factors determinants for housing prices in different US real estate

markets. This work also compares the influence of different data imputation methods and different time-lag situations on the model's performance. The results call for further investigation of the effectiveness of the technique to compensate missing values and the implementation of time lag.

The remainder of the paper is organized as follows. Initially, there is a review of the literature on the influences of demographic and macroeconomic factors on house price trend; research work to build predictive models for home price index; application of artificial intelligence in real estate studies. The need for more work on the influence of demographic and macroeconomic factors on house price trend in different markets, the influence of data imputation methods on building predictive models using real world data, and the effects of time lag on building predictive models are also identified in this section. In the succeeding section, a discussion of the data and methodology utilized in the study is provided. Analysis of results are provided in the results and discussion section. The final section offers comments and suggestions for future research.

LITERATURE REVIEW

Many factors influence the housing market, including household income, wealth, metropolitan statistical area (MSA) population, age of household heads, racial composition, local tax policy, interest rates, land constraints, regulatory constraints, and construction costs (Rodda & Goodman, 2005). Current econometric research on the effect of macroeconomic determinants, demographic conditions, and policy factors on housing price is often contradictory (Trofimov et al., 2018). Most papers show a significant, positive correlation between GDP and household income and housing prices; however, factors like money supply, interest rates, and disposable income are often disputed among studies (Tripathi, 2019, Trofimov et al., 2018). A study by Renigier-Biłozor and Wiśniewski (2013) showed that the economic and financial conditions of different European countries had variable influence on the prices of real estate. Given that the housing market is heterogeneous, that is, house price dynamics vary greatly across regions (Nam, 2020), the effect of macroeconomic and demographic variables on different real estate markets over different time periods must be examined specifically. By running the same linear regression algorithm on data over the same time period across five different markets, this present research was able to compare the coefficients of different input variables and to determine significant variables in each market. It was found that, depending on the studied area, the influence of demographic and macroeconomic factors varied.

In this work, the S&P/Case-Shiller Home Price Indices and twelve demographic and macroeconomic factors were studied for five metropolitan areas -- Boston, Dallas, New York (NY), Chicago, and San Francisco (SF). The data were collected from the Federal Reserve Bank (Federal Reserve Bank, n.d.), the Federal Bureau of Investigation (FBI) (Federal Bureau of Investigation, n.d.), Macrotrends.net (Macrotrends, n.d.), and Freddie Mac (Freddie Mac, n.d.).

The time ranges of available data for twelve features were varied, limiting the range of the final dataset that contained all features. Additionally, some factors were available monthly, some quarterly, and some yearly, leading to missing values. Missing values in features is a widely known problem in data-driven modeling and can be addressed by several methods, such as eliminating the feature, imputing the data with a mean, imputing the data by last observation, or using algorithms that support the missing values. However, the methods of dealing with missing data and their impact on the model's performance are rarely discussed in real estate modeling studies. In this work, two methods of imputing missing values, i.e., backfill and interpolation, were compared and discussed. The final datasets used for modeling ranged from March 2005 to December 2018 and included all features monthly, after imputing missing data points.

The intersection of real estate and artificial intelligence technology is still in the exploratory stage (Conway, 2018). Most of the previous work has focused on using artificial intelligence (AI) and machine learning (ML) for valuation (Conway, 2018). Predicting housing price trends using artificial intelligence techniques

has attracted increasing attention in recent years (Abidoye et. al., 2019, Li, 2009, Niu & Niu, 2019, Renigier-Biłozor & Wiśniewski, 2013, Vargason, 2019). In this work, linear regression (LR) and artificial neural network (ANN) methods have been employed to create data-driven models, and the predictive accuracy of the models were compared and discussed. These algorithms have performed well in the past for prediction tasks (Abidoye et al., 2019, Gruma & Govekarb, 2016, Nisha & Sreekumar, 2017). LR has been used in establishing real estate price index models over a long time period (Bailey et. al., 1963, Malpezzi et. al., 1980). ANN has been successfully used in several recent studies to model the housing price index with high accuracy (Abidoye et al., 2019, Renigier-Biłozor & Wiśniewski, 2013).

Autoregressive integrated moving average (ARIMA) and vector autoregressive (VAR) methods have been used to establish time-series models to predict housing price indices, too (Gupta et. al., 2011, Vishwakarma, 2013, Xie & Hu, 2007). ARIMA models were shown to be suitable for short-term forecasting, such as one-step-ahead forecasts (Karakozova, 2004, Stevenson & Young, 2007, Tse, 1997, Vishwakarma, 2013). In Gupta et al.'s (2011) work, eight time-lags were used, while in Abidoye et. al.'s (2019) work, five time-lags were adapted from studies of media influence on the stock market. A study suggested that the lag length for the VAR model should be two (Baffoe-Bonnie, 1998). Yet, another report did not discuss time lag (Li et. al., 2009). Time lag is an important variable in establishing predictive models, and more studies that experiment with time lags are needed.

Some studies focusing on AI modeling of the housing price index did not consider time lag (Renigier-Biłozor & Wiśniewski, 2013, Shukry et. al., 2012, Lim et. al., 2016). This work compared three sets of models with three different time-lag conditions: no lag or point prediction; a 6-month lag; and a 12-month lag, i.e., in the 6-month lag model, demographic and macroeconomics factors from six months ago were used as input parameters to model the current month's home price index, while in the no-time-lag model, factors in the current month were used to model the current month's home price index.

DATA AND METHODOLOGY

Data

The twelve macroeconomic and demographic features for each MSA chosen for this study were:

30-year fixed mortgage rate
Per-capita personal income
Resident population
Unemployment rate
Total gross domestic product (GDP)
Crime rate
Percentage of the population with mortgage debt
Median debt
Percentage of the population with severely delinquent debt
New private housing structures authorized by building permits
Index of economic conditions
Consumer price index for all urban consumers -- all items, less shelter.

Previous studies typically suggested that these factors were influential in determining housing prices (Rodda & Goodman, 2005; Tripathi, 2019; Trofimov et al., 2018).

The outputs of the models were S&P/Case-Shiller Home Price Indices, Index Jan 2000=100 (HPI) (S&P Dow Jones Indices LLC) for five different MSAs. S&P/Case-Shiller Home Price Indices are the leading

measures of US residential real-estate prices. They track the purchase price and resale value of single-family homes and are widely viewed as barometers of the US housing markets and broader economy.

Data were collected from the Federal Reserve Bank (Federal Reserve Bank, n.d.), Federal Bureau of Investigation (FBI) (Federal Bureau of Investigation, n.d.), Macrotrends.net (Macrotrends, n.d.), and Freddie Mac (Freddie Mac, n.d.). The time ranges of available data for each feature were varied, limiting the range of the final dataset containing all features. The range of the data used was from March 2005 to December 2018, and the frequency of the data was monthly, after filling in missing values. All data except mortgage rate and consumer price index for all urban consumers: all items less shelter were data for each MSA. Mortgage-rate data used were national data, due to the difficulty of locating publicly available mortgage rate data for each MSA. Data of the consumer price index for all urban consumers: all items less shelter were for core-based statistical areas (CBSA) due to availability of this data from Federal Reserve Economic Data (FRED). Table 1 summarizes the sources of each feature's data as well as the frequency of the data points, e.g., monthly, quarterly, or yearly.

Table 1: Frequency and Source of Features

Feature	Frequency	Source
S&P/Case-Shiller Home Price Index, Index Jan 2000=100	Monthly	S&P Dow Jones Indices LLC, retrieved from federal reserve economic data (FRED), Federal Reserve Bank of St. Louis
30-year fixed mortgage rate	Monthly	Freddie Mac
Unemployment rate	Monthly	U.S. Bureau of Labor Statistics, Unemployment Rate, retrieved from FRED, Federal Reserve Bank of St. Louis
New private housing structures authorized by building permits	Monthly	U.S. Census Bureau, New Private Housing Structures Authorized by Building Permits, retrieved from FRED, Federal Reserve Bank of St. Louis
Economic conditions index	Monthly	Federal Reserve Bank of St. Louis, Economic Conditions Index, retrieved from FRED, Federal Reserve Bank of St. Louis
Consumer price index for all urban consumers: all items less shelter	Monthly	U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: All Items Less Shelter, retrieved from FRED, Federal Reserve Bank of St. Louis
Percent of the population with mortgage debt	Quarterly	Federal Reserve Bank of New York: Consumer credit explorer
Median debt	Quarterly	Federal Reserve Bank of New York: Consumer credit explorer
Percent of the population with severely delinquent debt	Quarterly	Federal Reserve Bank of New York: Consumer credit explorer
Per capita personal income	Yearly	U.S. Bureau of Economic Analysis, Per Capita Personal Income, retrieved from FRED, Federal Reserve Bank of St. Louis
Resident population	Yearly	U.S. Census Bureau, Resident Population, retrieved from FRED, Federal Reserve Bank of St. Louis
Total GDP	Yearly	U.S. Bureau of Economic Analysis, Total Gross Domestic Product, retrieved from FRED, Federal Reserve Bank of St. Louis
Crime rate	Yearly	FBI: Crime in the US and Macrotrends.net

This table summarizes the sources and frequency of the twelve input variables and one output used in the study. The output, the S&P/Case-Shiller Home Price Index, was bolded and italicized. The three frequencies, monthly, quarterly, and yearly, in the second column indicate that the original data were available monthly, quarterly or yearly from the corresponding sources listed in the third column.

Missing Data Imputation

Since there were three different frequencies of available data, there were missing values in quarterly and yearly data sets. Different data imputation methods can be used to fill for missing values. In this study, missing values in quarterly and yearly data were filled by two methods: interpolation and backfilling. To fill by interpolation, values in between the two known endpoints were linearly regressed. To backfill, the

endpoint value was filled in for all the points before it, up to the previous endpoint. The accuracy of the models trained on data filled by these two methods was compared.

Time Lag

The time-lagging effect was considered, and there were three kinds of models established: no-lag, a 6-month lag, and a 12-month lag. The no-lag model used the features at a certain time point to produce the home price index for the same time point. The 6-month-lag model used features at a certain time point to produce the home price index for 6 months later. Similarly, the 12-month lag model produced the home price index for one year later.

Areas

The five selected metropolitan statistical areas for this study were Boston, New York (NY), San Francisco (SF), Dallas, and Chicago. These areas exhibited different housing market trends from 2005 to 2018: growth-decline-rapid growth to new high (Boston and SF); growth-decline-growth to recovery (NY); growth-decline-slow growth (Chicago); and flat with small variances-growth (Dallas). Two datasets containing data on the selected features in each area were created, one for interpolated data and one for backfilled data. Within each of these sets, three datasets for each of the lagging patterns were created. In total, thirty datasets were created.

Algorithms

The models selected for this study were linear regression (LR) and artificial neural network (ANN). When implementing the models, all datasets were split into a train set and a test set. The train set was used to train the model, and the test set was used to validate the model's performance or to see how well it could generalize to new data. In this study, the 2005-2017 HPI and its corresponding input features data were used as the training set. The 2018 HPI and its corresponding input features data were used as the test set. For each MSA, thirty ANN and thirty LR models were created.

A data normalization process was conducted first. This function was used to change the values in different columns to a common scale, to avoid distorting differences among multiple columns, as shown in Equation 1:

$$x_{new} = \frac{x - \mu}{\sigma} \tag{1}$$

Where μ is the mean value, and σ is a standard deviation. This process is called standardization or Z-score normalization. The function produces a normally distributed dataset.

The LinearRegression package in sklearn was utilized to build the LR model. The least-squares algorithm was adopted by the package and can be explained in the following equations (Angelini, 2019, Groß, 2012):

$$f(x) = X\beta + \epsilon \tag{2}$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \tag{3}$$

$$X = \left(\begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \right) \tag{4}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} \tag{5}$$

$$\in = \begin{pmatrix} \in_1 \\ \vdots \\ \in_n \end{pmatrix} \tag{6}$$

$$min\, L\big(y, f(x)\big) = \sum_{i=1}^{n} [y_n - f(x_i)]^2 \tag{7}$$

Where y is the observed target, x is the variable, β is the coefficient for the variable, ∈ is the error term, and L is the residual sum of squares between the observed targets and the predicted targets. The task of this algorithm is to find a set of coefficients and errors that can achieve the smallest residual sum of squares between observed and predicted outputs. The quality of the model was determined by Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and accuracy. A small MAE value represents high consistency between the predictions by the model and the actual labels. A small RMSE proves that the spread of predicted errors is small. The MAE and RMSE are calculated by the two equations below:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{9}$$

Where N is the number of data points, ŷ is the predicted value, and y is the actual value.

The ANN model used in this study had two hidden layers and 64 neurons in each layer. The neurons in the input layer were determined by the input properties and parameters. There were 832 parameters in hidden layer 1, 4160 parameters in hidden layer 2, and 65 parameters in the output layer. An RMSprop function was added onto the original back-propagation algorithm to avoid overfitting, which is represented by the following equations:

$$E[g^2]_t = \beta E[g^2]_{t-1} + (1-\beta)\left(\frac{\delta C}{\delta w}\right)^2 \tag{10}$$

$$w_t = w_{t-1} - \frac{\eta}{\sqrt{E[g^2]_t}}\frac{\delta C}{\delta w} \tag{11}$$

Where E[g] is the moving average of squared gradients, δC/δw is the gradient of the cost function, β is the moving average parameter, and η is the learning rate. The feed-forward, back-propagation process was conducted 400 times (epochs) to reach a global minimum. MAE and accuracy were applied to assess and summarize the quality of the ANN model.

RESULTS AND DISCUSSION

Linear Regression

Table 2 summarizes the statistics for the performance of LR models predicting 2018 HPI of the studied areas. Six models of each area have been built for two data imputation methods, backfilling and interpolating, and three lagging conditions, i.e., no-lag, a 6-month lag, and a 12-month lag. The MAE, accuracy, RMSE, R-squared, and adjusted R-squared of the train datasets of each metropolitan area were obtained and presented in Table 2. R-squared and adjusted R-squared are defined in the equations below as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \mu)^2} \quad (12)$$

$$R_{adj}{}^2 = 1 - \frac{(1-R^2)(N-1)}{N-k-1} \quad (13)$$

Where N is the number of data points, ŷ is the predicted value, y is the actual value, μ is the mean of y, and k is the number of variables in the model excluding the constant.

The MAE, RMSE and accuracy of the test datasets were also obtained and included in Table 2. The models for each MSA with highest accuracy and least MAE, using test datasets, are italicized in Table 2. The models with high accuracy also have small MAE and RMSE. Figure 1 compares the MAE of the test sets in all the LR models.

It can be seen in Table 2 that the accuracy for prediction values of all the train sets is very high (> 95.73%), the MAE is small (<6.98), and the R-squared and adjusted R-squared are close to 1. These results show that the prediction results of train sets have a high consistency with the true values. Thus, the success of the learning process is demonstrated.

Figure 1 and Table 2 show that the method of filling in missing data makes a difference in model accuracy. Comparing Figure 2 (A) and Figure 2 (B), the MAE varies when the strategy of data imputation changed for the same dataset, even though only seven out of twelve input variables required imputation. For example, MAEs of the SF models using testing datasets vary significantly; there is a difference of ~12 points between the MAE for the backfilled 12-month lag model and the interpolated 12-month lag model. As is similar to this study, real-world datasets frequently contain missing values for different reasons. Of course, there are other methods to compensate for missing values in a dataset, such as imputation using mean/median values, imputation using deep learning, hot-deck imputation, etc. The results in this study show that the data-filling technique used on a dataset that has a significant number of missing values can impact the quality of the data-driven model. It may be worthwhile to conduct future studies to investigate more thoroughly into the impact of data-filling techniques on creating a real estate price-trend model.

Figure 1 and Table 2 also show that choice of time lag influences the model's performance for each MSA. In both the interpolated and backfilled datasets, the MAEs of the models for each MSA varied by time lag. For example, the models of SF with backfilled data decreased significantly as time lag increased. A similar trend was seen in the backfilled data for Dallas. However, the exact variation of the MAEs vs. time lag was different for different markets. The MAEs of models for Chicago increased slightly with the increase in time lag in both backfilled and interpolated datasets.

Even though the best-performing models for different areas had different data imputation and time-lag conditions, generally, backfilled data with a 12-month lag performed very well, with a prediction accuracy

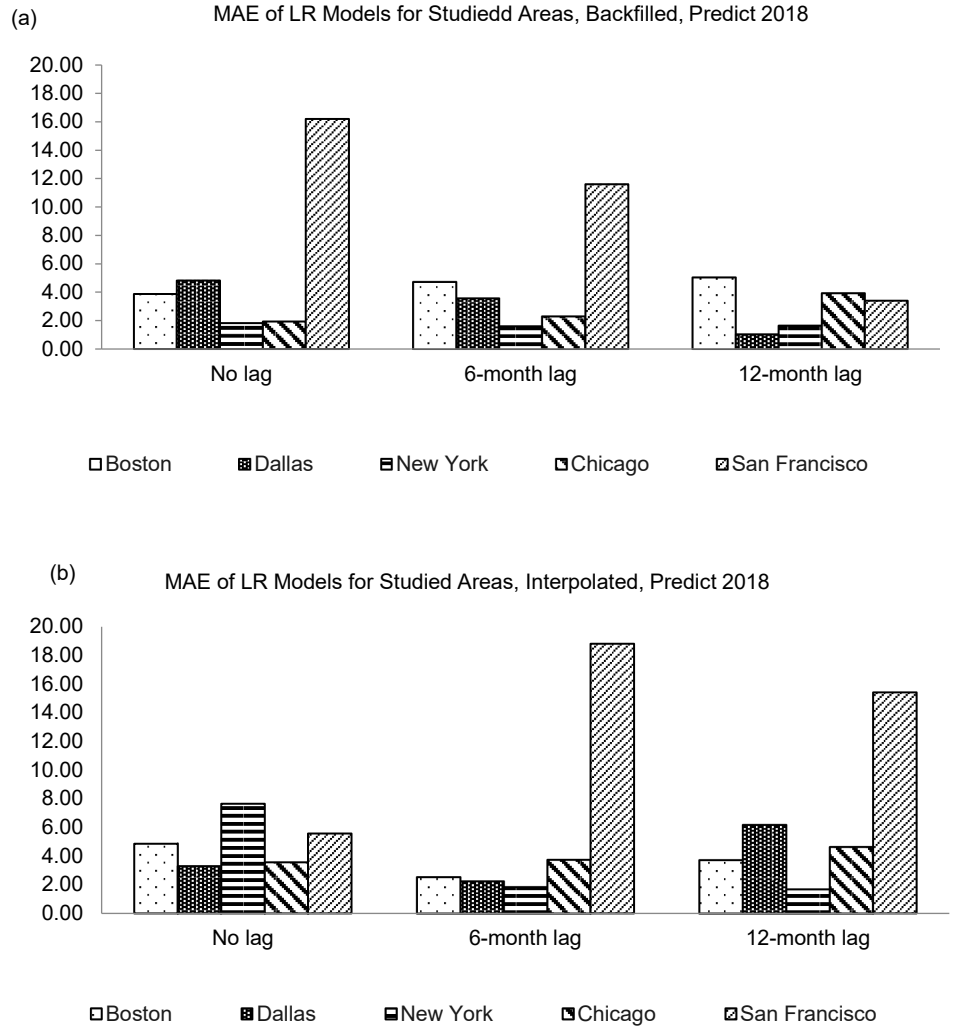
of >97.64% and MAEs of <5.04 for all five areas. Thus, the LR models showed reliable performance in predicting a housing price index one year ahead across different markets. An interesting extension for future studies would be finding the time lag that results in optimal performance of an algorithm on real-estate price prediction for specific area with selected data.

Table 2: Summary Statistics of Linear Regression Models Predicting 2018 HPI of Studied Areas

Panel A: No Lag							
			Boston	Dallas	New York	Chicago	San Francisco
Train set	Backfilled	MAE	2.1	2.1	2.15	1.81	4.72
		RMSE	2.6	2.63	2.73	2.27	6.03
		Accuracy	98.74%	98.32%	98.86%	98.62%	97.10%
		R Squared	97.19%	98.20%	97.48%	98.39%	97.64%
		Adjusted R Squared	96.95%	98.05%	97.27%	98.25%	97.44%
	Interpolated	MAE	1.85	1.45	1.51	1.73	4.55
		RMSE	2.29	1.8	1.97	2.21	5.93
		Accuracy	98.88%	98.86%	99.18%	98.67%	97.29%
		R Squared	97.81%	99.16%	98.58%	98.48%	97.72%
		Adjusted R Squared	97.62%	99.09%	98.46%	98.35%	97.53%
Test set	Backfilled	MAE	3.87	4.82	1.82	*1.93*	16.19
		RMSE	4.4	5.3	2.01	*2.08*	16.96
		Accuracy	98.19%	97.42%	99.08%	*98.65%*	93.88%
	Interpolated	MAE	4.85	3.31	7.65	3.57	5.57
		RMSE	5.38	3.57	8.96	3.74	6.31
		Accuracy	97.71%	98.23%	96.16%	97.50%	97.90%
Panel B: 6-month Lag							
			Boston	Dallas	New York	Chicago	San Francisco
Train set	Backfilled	MAE	2.11	1.79	1.96	2.74	6.62
		RMSE	2.58	2.25	2.64	3.43	8.38
		Accuracy	98.75%	98.61%	98.93%	97.91%	95.85%
		R Squared	97.27%	98.54%	97.46%	96.29%	95.54%
		Adjusted R Squared	97.04%	98.42%	97.24%	95.97%	95.16%
	Interpolated	MAE	1.83	1.29	1.65	2.8	6.27
		RMSE	2.2	1.6	2.22	3.61	7.38
		Accuracy	98.91%	99.00%	99.10%	97.88%	96.21%
		R Squared	98.03%	99.35%	98.20%	95.88%	96.55%
		Adjusted R Squared	97.86%	99.29%	98.05%	95.53%	96.26%
Test set	Backfilled	MAE	4.73	3.58	*1.63*	2.32	11.6
		RMSE	5.51	3.77	*1.88*	2.51	12.79
		Accuracy	97.78%	98.08%	*99.18%*	98.37%	95.63%
	Interpolated	MAE	*2.53*	2.25	1.84	3.74	18.81
		RMSE	*3.25*	2.34	2.2	3.91	19.5
		Accuracy	*98.82%*	98.79%	99.08%	97.38%	92.89%
Panel C: 12-month Lag							
			Boston	Dallas	New York	Chicago	San Francisco
Train set	Backfilled	MAE	2.1	1.99	1.92	2.69	6.98
		RMSE	2.6	2.46	2.27	3.27	8.44
		Accuracy	98.77%	98.46%	98.94%	97.99%	95.73%
		R Squared	97.29%	98.44%	97.95%	96.38%	95.51%
		Adjusted R Squared	97.06%	98.31%	97.78%	96.07%	95.13%
	Interpolated	MAE	1.51	1.3	1.28	2.67	5.9
		RMSE	1.93	1.75	1.62	3.28	7.02
		Accuracy	99.11%	98.98%	99.28%	97.99%	96.44%
		R Squared	98.51%	99.24%	98.96%	96.37%	96.90%
		Adjusted R Squared	98.38%	99.18%	98.87%	96.06%	96.64%
Test set	Backfilled	MAE	5.04	*1.06*	1.66	3.95	*3.4*
		RMSE	5.51	*1.18*	1.82	4.39	*5.62*
		Accuracy	97.64%	*99.43%*	99.16%	97.23%	*98.70%*
	Interpolated	MAE	3.73	6.18	1.7	4.63	15.43
		RMSE	4.41	6.99	2.38	5.43	19.65
		Accuracy	98.24%	96.69%	99.14%	96.76%	94.20%

This table summarizes the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), accuracy, R squared, and adjusted R squared of the thirty linear regression models built. The MAE, RMSE and accuracy of the test sets are bolded to distinguish from those of the train sets. Panel A, B and C show three different time lag conditions studied: no lag, 6-month lag and 12-month lag, correspondingly. The second column shows the data imputation method used in building the model: backfilling and interpolating. To backfill, the endpoint value was filled in for all the points before it, up to the previous endpoint. To fill by interpolation, values in between the two known endpoints were linearly regressed. The models with the smallest MAE and highest accuracy for each of the five areas are italicized.

Figure 1: Mean Absolute Error (MAE) of Linear Regression Models with Different Time Lags and Data Imputing Conditions: (a) MAE of Models Using Backfilled Datasets, and (b) MAE of Models Using Interpolated Datasets



This figure compares the Mean Absolute Error (MAE) of the thirty linear regression models with different time lags and data imputing conditions using test datasets. Panel (a) compares the MAEs of models using backfilled datasets, and Panel (b) compares MAEs of models using interpolated datasets. Within each panel, five models with same time lag conditions for five areas are grouped for comparison. There were three different time-lag conditions: no lag, 6-month lag, and 12-month lag.

Table 3: Summary of T-test Results of The Backfilled, 12-Month Lag Linear Regression Models for Studied Areas

Panel A: Boston, Dallas, and New York									
Variables	**Boston**			**Dallas**			**New York**		
	Coeff	**T**	**P**	**Coeff**	**T**	**P**	**Coeff**	**T**	**P**
Constant	172.02	757.066	0.000	137.312	671.632	0.000	182.898	984.598	0.000
Mortgage Rate (30 years)	-1.616*	-1.818	0.071	-1.923***	-2.925	0.004	2.079***	3.009	0.003
Personal Income	-4.188	-0.730	0.467	1.374	0.663	0.508	10.812***	4.182	0.000
Resident Population	-8.262***	-2.775	0.006	15.065***	7.846	0.000	0.402	0.306	0.760
Unemployment Rate	3.2847***	3.658	0.000	2.900***	3.886	0.000	5.158**	3.459	0.001
GDP	28.271***	4.061	0.000	-0.299	-1.068	0.287	-5.008	-1.630	0.105
Crime Rate	4.191**	2.178	0.031	8.342***	6.338	0.000	-0.511	-0.924	0.357
Median Debt	-0.924*	-1.955	0.053	0.779*	1.826	0.070	-8.883***	-8.828	0.000
New Structures	-0.380**	-1.171	0.244	0.780*	1.941	0.054	0.679***	3.026	0.003
Economic Conditions Index	1.552***	4.731	0.000	-1.769***	-4.165	0.000	-0.771	-0.173	0.863
CPI less shelter	-2.338*	-1.682	0.095	-3.025**	-2.482	0.014	-4.496***	-3.873	0.000
Percent With Mortgage Debt	-5.425***	-5.783	0.000	-12.176***	-8.107	0.000	-4.414***	-4.572	0.000
Percent With Severely Delinquent	-10.094***	-7.424	0.000	-3.361***	-2.796	0.006	-8.891***	-4.428	0.000

Panel B: Chicago and San Francisco						
Variables	**Chicago**			**San Francisco**		
	Coeff	**T**	**P**	**Coeff**	**T**	**P**
Constant	134.138	484.181	0.000	185.823	267.425	0.000
Mortgage Rate (30 years)	2.420*	1.965	0.051	-0.583	-0.221	0.826
Personal Income	19.952***	4.069	0.000	52.731***	4.465	0.000
Resident Population	-1.517**	-2.249	0.026	-24.531***	-3.096	0.002
Unemployment Rate	-0.958	0.905	0.367	3.893	0.856	0.393
GDP	-19.563**	-3.250	0.001	-5.884	-0.39	0.697
Crime Rate	-0.985	-1.517	0.131	3.778**	2.355	0.020
Median Debt	-14.589***	-7.946	0.000	-18.454***	-7.739	0.000
New Structures	2.866**	3.407	0.001	0.240	0.223	0.824
Economic Conditions Index	-2.062**	-3.475	0.001	-8.851***	-6.573	0.000
CPI less shelter	-8.248***	-6.025	0.000	-13.123**	-2.298	0.023
Percent With Mortgage Debt	6.251***	3.008	0.003	-12.911***	-2.936	0.004
Percent With Severely Delinquent	-4.683***	-2.980	0.003	-12.058**	-2.170	0.032

*This table summarizes some of the regression results of the five backfilled, 12-month-lag models for five areas. As observed previously, backfilled, 12-month lag condition resulted in low MAE and high accuracy in models for all five areas. *, **, *** indicate significance at the 10, 5, and 1 percent levels, respectively. A P-value larger than the common alpha level of 0.05 was considered in this study to indicate that the variable was not statistically significant.*

Overall, LR performs well in predicting HPI. All LR prediction models for 2018 had accuracy above 92.89%. Among all the results, backfilled 12-month-lag condition generated good model performance across different markets, with the accuracy of the models ranging from 97.64% to 99.42%.

T-tests were conducted to further understand the significance of each variable. Table 3 summarizes the T-test results of the backfilled, 12-month-lag LR models for five areas, since as shown in Table 1 and Figure 1, backfilled, 12-month-lag condition resulted in low MAE and high accuracy in models for all five areas. *, **, *** indicates significance at the 10, 5, and 1 percent levels, respectively. In this study, a P-value larger than the common alpha level of 0.05 was considered to indicate that the variable was not statistically significant. It was observed that, for different areas, different variables were not statistically significant. For example, for Dallas, personal income, GDP, median debt and new structures were not statistically significant. However, for Chicago, mortgage rate, unemployment rate, and crime rate were not statistically significant. It is well recognized that the real estate market is local. Each MSA has unique demographic and economic characteristics and different real estate market characteristics. Thus, variation in the influence of different input features in different areas is expected. Table 3 suggests that the percentage of severely delinquent debt and the percentage with mortgage debt were significant for all five models and were negatively associated with HPI. Residential population, the economic conditions index, and the CPI less shelter, were significant in four out of five models. Personal income was significant in three models, and when it was significant, it was positively associated with HPI. Unemployment rate, crime rate, and median debt were also significant in three models. GDP, mortgage rate, and new structures were not significant in three out of five models. It is worth noting that multicollinearity may exist in the models and may have contributed to the results in Table 3. Multicollinearity occurs when the independent variables in the model have correlations with each other (Vatcheva et. al., 2016), which is likely the case for demographic and macroeconomic features. For example, GDP, personal income and CPI may have correlation with each other. Multicollinearity does not have a negative impact on the reliability of the model; it affects the coefficients, but it does not influence the predictions and the precision of the predictions. In addition, when there is multicollinearity, the coefficients of the model may vary with a small change of data (Vatcheva et. al., 2016).

ANN

Table 4 summarizes the statistics for the performance of the ANN models that predict 2018 HPI of studied areas. The MAE, accuracy, RMSE and accuracy of the train datasets and test datasets of each metropolitan area are included in Table 4. The models with high accuracy also showed small MAE and RMSE. The model for each MSA with best accuracy and smallest MAE using the test datasets is italicized in Table 4. Figure 2 compares the MAE of the test sets in all the ANN models.

As shown in Table 4, like the LR method, six models of each MSA have been built for two data imputation methods and three time-lag conditions. The accuracy for prediction values of all train sets was very high (> 96.51%), and the MAE was small (<4.98). These results prove that the prediction results of train sets had high consistency with the true values. Thus, the success of the learning process was demonstrated.

Figure 2 and Table 4 again show that the method of filling in missing data makes a difference in the model's accuracy. For example, MAEs of the NY models using testing datasets varied significantly; there was a difference of ~9 points between the MAE for the backfilled 6-month-lag model and the interpolated 6-month lag model. Thus, the ANN study still suggested that an extension of this work could be a detailed study to determine the method used to compensate the missing data, to obtain reliable real-estate prediction models.

This study showed that LR performed better in long-term trend prediction in all five markets. The ANN technique was more suitable for short-term prediction. A few previous studies suggested that ANN

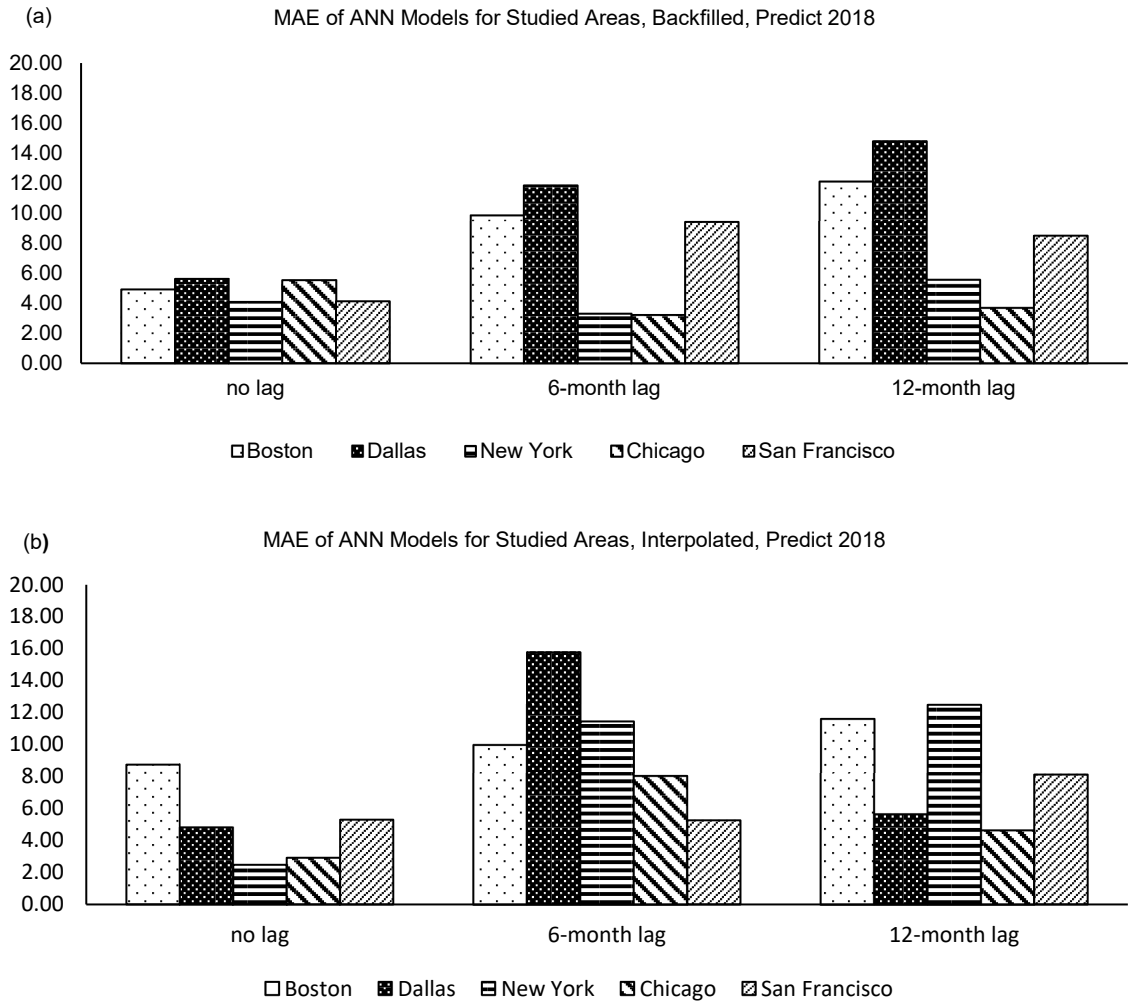
performed better than regression methods in predicting property price index (Abidoye, 2019, Lim, 2016). The difference in the characteristics of the studied areas, selection of input and output features, quality of the dataset used, and specifics in data preparation -- such as imputing missing values and time-lag choices -- in different studies may have contributed to the variation in the accuracy of obtained models by the ANN or regression methods in this study and in previous studies. This study suggests that an appropriate selection of features, data, data preparation conditions, and time lag could result in effective models by both the LR and ANN methods.

Table 4: Summary Statistics of ANN Models Predicting 2018 Home Price Index of Studied Areas

Panel A: No lag							
			Boston	**Dallas**	**New York**	**Chicago**	**San Francisco**
Train set	Backfilled	MAE	2.8	2.07	3.21	2.02	3.42
		RMSE	3.58	2.7	4.5	2.61	4.47
		Accuracy	98.35%	98.42%	98.24%	98.48%	98.17%
	Interpolate	MAE	2.17	1.63	2.4	1.89	2.51
		RMSE	2.77	2.13	3.42	2.48	3.26
		Accuracy	98.74%	98.71%	98.72%	98.63%	98.68%
Test set	Backfilled	MAE	*4.91*	5.62	4.09	5.53	*4.14*
		RMSE	*5.84*	6.44	5.09	5.87	*5.04*
		Accuracy	*97.69%*	96.98%	97.94%	96.12%	*98.43%*
	Interpolated	MAE	8.74	*4.83*	*2.48*	*2.92*	5.3
		RMSE	9.13	*5.49*	*3.22*	*3.36*	6.69
		Accuracy	95.91%	*97.41%*	*98.75%*	*97.96%*	98.00%
Panel B: 6-month Lag							
			Boston	Dallas	New York	Chicago	San Francisco
Train set	Backfilled	MAE	3.11	3.12	3.01	1.95	3.31
		RMSE	4.07	3.84	4.33	2.6	4.39
		Accuracy	98.17%	97.69%	98.35%	98.57%	98.13%
	Interpolated	MAE	2.39	1.88	2.4	1.66	2.34
		RMSE	3.11	2.46	3.52	2.26	3.05
		Accuracy	98.61%	98.58%	98.71%	98.81%	98.70%
Test set	Backfilled	MAE	9.84	11.84	3.33	3.24	9.42
		RMSE	10.3	14.69	3.99	3.58	10.14
		Accuracy	95.40%	93.64%	98.32%	97.72%	96.42%
	Interpolated	MAE	9.96	15.78	11.45	8	5.27
		RMSE	10.92	16.42	12.73	8.34	6.5
		Accuracy	95.35%	91.54%	94.22%	94.39%	98.00%
Panel C: 12-month Lag							
			Boston	Dallas	New York	Chicago	San Francisco
Train set	Backfilled	MAE	3.36	2.07	3.34	2.3	3.53
		RMSE	4.45	2.7	4.48	2.87	4.45
		Accuracy	98.05%	98.42%	98.17%	98.31%	98.06%
	Interpolated	MAE	2.39	1.95	2.41	1.91	3.01
		RMSE	3.11	2.51	3.15	2.53	3.84
		Accuracy	98.61%	98.49%	98.69%	98.57%	98.37%
Test set	Backfilled	MAE	12.09	14.78	5.56	3.7	8.5
		RMSE	14.3	15.24	6.45	4.45	10.55
		Accuracy	94.37%	92.07%	97.20%	97.41%	96.78%
	Interpolated	MAE	11.59	5.62	12.49	4.63	8.13
		RMSE	14.38	6.44	13.91	5.16	9.67
		Accuracy	94.61%	96.98%	93.73%	96.76%	96.91%

This table summarizes the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and accuracy, of the thirty artificial neural network (ANN) models built. The MAE, RMSE and accuracy of the test sets are bolded to distinguish from those of the train sets. Panel A, B and C show three different time lag conditions studied: no lag, 6-month lag and 12-month lag, correspondingly. The second column shows the data imputation method used in building the model: backfilling and interpolating. To backfill, the endpoint value was filled in for all the points before it, up to the previous endpoint. To fill by interpolation, values in between the two known endpoints were linearly regressed. The models with the smallest MAE and highest accuracy for each of the five areas are italicized.

Figure 2: MAE of ANN Models with Different Lags and Imputing Conditions (a) MAE of Models Using Backfilled Datasets, and (b) MAE of Models Using Interpolated Datasets



This figure compared the Mean Absolute Error (MAE) of the thirty artificial neural network (ANN) models with different time lags and data imputing conditions using test datasets. Panel (a) compares the MAEs of models using backfilled datasets, and Panel (b) compares MAEs of models using interpolated datasets. Within each panel, five models with same time lag conditions for five areas are grouped for comparison. There were three different time-lag conditions: no lag, 6-month lag and 12-month lag.

CONCLUDING COMMENTS

This work employed the LR and ANN techniques to achieve an accurate and reliable property price index prediction that could aid important, strategic planning and decision-making. Buying a house is often one of the most important personal financial decisions. Predicting real estate price trends will help buyers make cost-efficient decisions at an optimum time or location for them. Portfolio managers and investors will also have much to gain from accurate predictions of long-term real estate trends. Real estate is a key part of any diversified investment portfolio. By lessening managers' research workload and providing key insights about capital appreciation trends, a model that can predict property trends will allow portfolios to perform optimally. Such a model is useful to local and national governments as well. In the US, the federal government is heavily involved in real estate through mortgage institutions like Fannie Mae and Freddie Mac. Local governments often rely on property taxes to gather resources. Also, real estate constitutes a large portion of the American economy. Accounting for movements in the housing market gives

governments a better idea of projected property tax income, helps financial planning with mortgage programs, and may guide fiscal policy (Vargason, 2019).

In this work, data comprising twelve demographic and macroeconomic features and HPI that covered the period between March 2005 and December 2018 in five different metropolitan statistical areas were collected from institutes such as Federal Reserve Bank and the FBI. The five geographic areas represent four different home-price trends in the time period studied. Two methods to compensate the missing values in the data and three different time-lag situations have been analyzed, resulting in sixty total models established for the ANN and LR methods.

Evaluation of the forecasts generated by the models shows that ANN was suitable for short-term predictions and that LR performed better than ANN for long-term predictions. This study also shows that the technique to compensate missing values in the dataset and the implementation of time lag could have significant influence on the model's performance and requires further investigation.

Finally, even though real estate markets are local, this study shows that certain combinations of conditions resulted in high-performance models in all five areas, such as the five LR models with backfilled, 12-month-lag conditions, and the five ANN models using interpolated no-lag conditions. Future studies on multiple populated areas will be needed to generalize these conditions as starting points to creating a data-driven model, using different algorithms, for real estate price index prediction.

REFERENCES

Abidoye, R. B., Chan, A. P. C., Abidoye, F. D., & Oshodi, O. S. (2019). Predicting property price index using artificial intelligence techniques: Evidence from Hong Kong. International Journal of Housing Markets and Analysis, 12 (5): 1072-1092.

Angelini, C. (2019). Regression Analysis. Elsevier Inc. https://doi.org/10.1016/B978-0-12-809633-8.20360-9.

Baffoe-Bonnie, J. (1998). The dynamic impact of macroeconomic aggregates on housing prices and stock of houses: A national and regional analysis. The Journal of Real Estate Finance and Economics 17: 179–197. https://doi.org/10.1023/A:1007753421236.

Bailey, M. J., Muth, R. F., & Nourse. H. O. (1963). A regression method for real estate price index construction. Journal of the American Statistical Association, 58: 933-42.

Bank of England. (n.d.). How does the housing market affect the economy? Retrieved October 26, 2020, from http://www.bankofengland.co.uk/knowledgebank/how-does-the-housing-market-affect-the-economy.

Conway, J. (2018). Artificial Intelligence and Machine Learning: Current Applications in Real Estate [Master's thesis, MIT]. Dspace@MIT.

Federal Reserve Bank (n.d.). Consumer credit explorer. https://www.philadelphiafed.org/surveys-and-data/community-development-data/consumer-credit-explorer.

Federal Reserve Bank (n.d.). Federal reserve economic data. https://fred.stlouisfed.org.

Freddie Mac (n.d.). Mortgage rate. http://www.freddiemac.com/pmms/.

Gerrity, M. (2020, January 20). U.S. Housing Market's Combined Value Hits $33.6 Trillion in 2020. World Property Journal. Retrieved October 26, 2020, from https://www.worldpropertyjournal.com/real-estate-news/united-states/los-angeles-real-estate-news/real-estate-news-zillow-housing-data-for-2020-combined-housing-market-value-in-2020-us-gdp-china-gdp-rising-home-value-data-11769.php.

Groß, J. (2012). Linear Regression. Vol. 175. 2012: Springer Science & Business Media.

Gruma, B., & Govekarb D. K. (2016). Influence of macroeconomic factors on prices of real estate in various cultural environments: Case of Slovenia, Greece, France, Poland and Norway. 3rd Global Conference on Business, Economics, Management and Tourism, Rome, Italy. Procedia Economics and Finance, 39: 597 – 604.

Gupta, R., Kabundi, A., & Miller, S. M. (2011). Forecasting the US real house price index: Structural and non-structural models with and without fundamentals. Economic Modelling, 28 (4): 2013-2021. https://doi.org/10.1016/j.econmod.2011.04.005.

Karakozova, O. (2004). Modelling and forecasting office returns in the Helsinki area. Journal of Property Research, 21(1): 51-73.

Li, D. Y., Xu, W., Zhao, H., Chen, R. Q. (2009). A SVR based forecasting approach for real estate price prediction. Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, China, July 12-15.

Lim, W.T., Wang, L., Wang, Y. & Chang, Q. (2016). Housing price prediction using neural networks. 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2016), Changsha, China.

Lyons, W. (2017). Real estate investment system and method of controlling a commercial system by generating key investment indicators (Patent No. 20170243296A1). https://patents.google.com/patent/US20170243296A1/en.

Macrotrends (n.d.).www.macrotrends.net.

Malpezzi, S., Ozanne, L., & Thibodeau, T. (1980). Characteristic Prices of Housing in 59 SMSAs. The Urban Institute. https://www.huduser.gov/publications/pdf/hud-50814.pdf.

Nam, Ty. (2020). Geographic Heterogeneity in Housing Market Risk and Portfolio Choice. J. Real Estate Finan. Econ. (2020). https://doi.org/10.1007/s11146-020-09762-9.

Nisha, K.G. & Sreekumar, K. (2017). A review and analysis of machine learning and statistical approaches for prediction. 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT): 135–139. https://doi.org/10.1109/ICICCT.2017.7975174.

Niu, J. & Niu, P. (2019). An intelligent automatic valuation system for real estate based on machine learning. Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing: 1–6. https://doi.org/10.1145/3371425.3371454.

Renigier-Biłozor, M., & Wiśniewski, R. (2013). The impact of macroeconomic factors on residential property price indices in Europe. Folia Oeconomica Stetinensia, 12(2): 103-125. DOI: https://doi.org/10.2478/v10031-012-0036-3.

Rodda, D. T. & Goodman, J. (2005). Recent House Price Trends and Homeownership Affordability. U.S. Department of Housing and Urban Development Office of Policy Development & Research. https://www.huduser.gov/Publications/pdf/RecentHousePrice_P1.pdf.

Shukry, M., Radzi, M., & Muthuveerappan, C. (2012). Forecasting house price index using artificial neural network. International Journal of Real Estate Studies, 7 (1): 43-48.

S&P Dow Jones Indices LLC, S&P/Case-Shiller Home Price Index, retrieved from FRED, Federal Reserve Bank of St. Louis; June 14, 2021.

Stevenson, S., &Young, J. (2007). Forecasting housing supply: Empirical evidence from the Irish market. European Journal of Housing Policy, 7: 1-17.

Stupak, J. M. (2019). Introduction to U.S. Economy: Housing Market. Congressional Research Service. https://fas.org/sgp/crs/misc/IF11327.pdf.

Tripathi, S. (2019). Macroeconomic determinants of housing prices: A cross country level analysis. In MPRA Paper (No. 98089; MPRA Paper). University Library of Munich, Germany. https://ideas.repec.org/p/pra/mprapa/98089.html.

Trofimov, I. D., Nazaria, A., & Xuan, D. C. D. (2018). Macroeconomic and demographic determinants of residential property prices in Malaysia. Zagreb International Review of Economics & Business, 21(2):71–96. Business Premium Collection; Publicly Available Content Database. https://doi.org/10.2478/zireb-2018-0015.

Tse, R.Y.C. (1997). An application of the ARIMA model to real-estate prices in Hong Kong. Journal of Property Finance, 8(2): 155-163.

Vargason, D. (2019). Data Mining Techniques for Predicting Real Estate Trends [La Salle University]. https://digitalcommons.lasalle.edu/mathcompcapstones/44.

Vatcheva, K. P., Lee, M., McCormick, J. B., & Rahbar, M. H. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. Epidemiology (Sunnyvale, Calif.), 6(2), 227. https://doi.org/10.4172/2161-1165.1000227

Vishwakarma, V. K. (2013). Forecasting real estate business: Empirical evidence from the Canadian market. Global Journal of Business Research, 7(3): 1-14. https://ssrn.com/abstract=2148507.

Xie, X. & Hu, G. (2007) A comparison of Shanghai housing price index forecasting. 3rd International Conference on Natural Computation, Haikou, China.

BIOGRAPHY

Sophia Zhou is a student researcher at the Massachusetts Academy of Math and Science at Worcester Polytechnic Institute. She is also currently a summer research intern at Fermi National Accelerator Laboratory. Her research work has been published in the Journal of Safety Science and Resilience and presented at the 4th International Conference on Business, Management, and Finance; IEEE Global Humanitarian Technology Conference 2021; 10th International Conference on Data Science, Technology and Applications (DATA 2021); AsianNetwork 2021 Annual Conference; and ASEE NE 2020 Annual Conference. She can be reached at 100 Institute Road, Worcester, MA 01609.