Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization

Case Study: Malang, East Java, Indonesia

Adyan Nur Alfiyatin
Faculty of Computer Science
Brawijaya University, Malang, Indonesia

Hilman Taufiq
Faculty of Computer Science
Brawijaya University, Malang, Indonesia

Ruth Ema Febrita
Faculty of Computer Science
Brawijaya University, Malang, Indonesia

Wayan Firdaus Mahmudy
Faculty of Computer Science
Brawijaya University, Malang, Indonesia

***Abstract*—House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. There are three factors that influence the price of a house which include physical conditions, concept and location. This research aims to predict house prices based on NJOP houses in Malang city with regression analysis and particle swarm optimization (PSO). PSO is used for selection of affect variables and regression analysis is used to determine the optimal coefficient in prediction. The result from this research proved combination regression and PSO is suitable and get the minimum prediction error obtained which is IDR 14.186.**

Keywords—House prediction; regression analysis; particle swarm optimization

I. Introduction

Investment is a business activity that most people are interested in this globalization era. There are several objects that are often used for investment, for example, gold, stocks and property. In particular, property investment has increased significantly since 2011, both on demand and property selling [1]. One of the increasing of property demand is because of high population in Indonesia. Indonesian Central Bureau of Statistics states that in East Java 50% of the population of East Java classified as a young population who have age approximately at 30 years old [2]. The result of this census indicates that the younger generation will need a house or buy a house in the future. Based on preliminary research conducted, there are two standards of house price which are valid in buying and selling transaction of a house that is house price based on the developer (market selling price) and price based on Value of Selling Tax Object (NJOP). According to Lim, et al the fundamental problem for a developer is to determine the selling price of a house [3]. In determining the price of home, the developer must calculate carefully and determine the appropriate method because property prices always increase continuously and almost never fall in the long term or short [4].

There are several approaches that can be used to determine the price of the house, one of them is the prediction analysis. The first approach is a quantitative prediction. A quantitative approach is an approach that utilizes time-series data [5]. The time-series approach is to look for the relationship between current prices and prevailing prices. The second approach is to use linear regression based on hedonic pricing [6], [7]. Previous research conducted by Gharehchopogh, et al. [7] using linear regression approach get 0,929 error with the actual price. In linear regression, determining coefficients generally using the least square method, but it takes a long time to get the best formula.

Particle swarm optimization (PSO) is proposed to find the coefficients aimed at obtaining optimal results [8]. Some previous researches such as Marini and Walzack [9], [10] show that PSO gets better results than other hybrid methods. There are several advantages of PSO, in the small search space PSO can do better solution search [11]. Although the PSO global search is less than optimal [12], but on the optimization problem the value of the variable on the regression equation can find a maximum solution using PSO [12], [13].

This research aims to create a house price prediction model using regression and PSO to obtain optimal prediction results. PSO is used for selection of affect variables in house prediction, regression is used to determine the optimal coefficient in prediction. In this study, researchers wanted to know the performance of the developed model in time series data. Prediction house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location. This research is focused in Malang City, because Malang is one of tourism and urban city in East Java.

II. Related Work

A. House Price Affecting Factors

There are several factors that affect house prices. In his research Rahadi, et al. [14] divide these factors into three main groups, there are physical condition, concept and location. Physical conditions are properties possessed by a house that

can be observed by human senses, including the size of the house, the number of bedrooms, the availability of kitchen and garage, the availability of the garden, the area of land and buildings, and the age of the house [15], while the concept is an idea offered by developers who can attract potential buyers, for example, the concept of a minimalist home, healthy and green environment, and elite environment.

Location is an important factor in shaping the price of a house. This is because the location determines the prevailing land price [16]. In addition, the location also determines the ease of access to public facilities, such as schools, campus, hospitals and health centers, as well as family recreation facilities such as malls, culinary tours, or even offer a beautiful scenery [17], [18]. In general, the factors affecting the house prices will be presented in Table 1.

TABLE I. HOUSE PRICE AFFECTING FACTORS

Literature	Physical condition							Concept	Location				
[15] (Limsombunchai, 2004)		√		√		√	√		√	√		√	
[18] (Jim and Chen, 2009)		√					√	√					√
[17] (Kisilevich, Keim and Rokach, 2013)												√	√
[16] (Zhu and Wei, 2013)									√	√	√	√	√
[14] (Rahadi, et all, 2015)	√	√	√	√	√		√	√	√	√	√	√	√
[19] (Bryant, 2016)	√	√		√	√								

B. Hedonic Pricing

Hedonic pricing is a price prediction model based on the hedonic price theory, which assumes that the value of a property is the sum of all its attributes value [20]. In the implementation, hedonic pricing can be implemented using regression model. Equation 1 will show the regression model in determining a price.

$$y = a.x_1 + b.x_2 + \cdots + n.x_i \quad (1)$$

Where, y is the predicted price, and x_1, x_2, x_i are the attributes of a house. While a, b, ... n indicate the correlation coefficients of each variables in the determination of house prices.

III. DATA SET

In this research, we use house price data based on NJOP from Land and Building Tax (PBB) payment structure. Due to limited access to the data, this study used 9 houses data in time series scattered in Malang City area, within 2014-2017. Normalization of data is done by completing the empty data at a certain time with the assumption that land prices tend to change every 2 years, while building prices tend to be stable.

The data tabulation offer information of the houses includes: home id, address (street name), longitude-latitude, year, building area, land area, NJOP building price (IDR/m^2), NJOP land price (IDR/m^2), distance from city center(km), amount number of campuses, amount number of restaurants, amount number of health facilities, amount number of playground, amount number of schools, amount number of traditional markets or malls, amount number of worship places, and also easiness access to public transportation. The city center in this study defined as the location of the square of Malang City. The distance to city center is calculated using Google maps. Meanwhile, easy access to public transportation is calculated between radius 400 meter. The calculation of nearest objects in the certain radius using buffering techniques accessed through the site http://obeattie.github.io/.

IV. RESEARCH METHODOLOGY

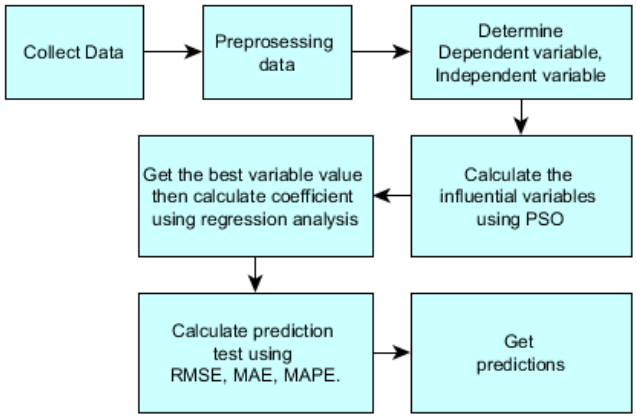

Fig. 1. Diagram flow research.

Based on Fig. 1, the process of regression analysis and particle swarm optimization methods is described in the following section:

A. Regression analysis

The prediction model used in this research is hedonic pricing, the suitable model using regression, with the standard formula as shown in (1). The dependent variable symbolized as *Y* is NJOP price and independent variables with symbol x_1- x_{14} consist of year, building area, land area, NJOP land price (IDR/m^2), NJOP building price (IDR/m^2), distance to center of the city, amount number of campuses, amount number of restaurants, amount number of health facilities, amount number of amusement parks, amount number of educational facilities, amount number of traditional markets, amount number of worship places, and easiness to public transportations is shown in (2).

$$NJOP = a.building\ area + b.land\ area + \cdots + n.public\ transportation \quad (2)$$

In this case, the public transportation variable will be 0 or 1, 0 means no public transport passes the area within 200 meters. And 1 means that there is public transports which passes through the area.

B. Particle Swarm Optimization (PSO)

PSO is a stochastic optimization method that represents solutions as particle [21]. Amount number of particles are generated randomly, where each particle consists of some dimensions of *xi* position and velocity *vi*. Each particle will measure its fitness value which shown in (3).

$$f(x) = \varepsilon \text{ from prediction} \quad (3)$$

Where, *f (x)* is the fitness value of each particle that indicates the error prediction value. Each particle will explore the solution search space to get optimal results. The displacement from one position to another is greatly influenced by the speed of each particle, to obtain the best position required a dynamic speed formulation using (4) [22].

$$v_i^{t+1} = w.v_i^t + c_1 . r_1 (p_i - x_i) + c_2 . r_2 (pg_i - x_i) \quad (4)$$

Where, *vi* shows the velocity value for the particle dimension to *i* to *n*, *t* denotes the iteration time, *w* is the value of the inertia vector whose value is obtained dynamically using (5) [23]. p_i is the best position ever obtained for each particle, while the pg_i is the best position ever achieved by the whole particle. c_1 and c_2 sequential are cognitive and social constant, which in this study is 2.5 and 0.5. r_1 and r_2 are 0.5 and 2.5. Once obtained speed will be updated position using (6).

$$W = (w\max - w\min)\frac{iterasi-t}{iterasi} + w\min, \quad (5)$$

$$x_i^{t+1} = x_i + v_i^{t+1}, \quad (6)$$

In the PSO, too fast particle displacement position can make the method fail to obtain the optimum solution. This problem can be handled by performing speed control or velocity clamping [9]. The speed control mechanism by conducting conditions for the speed of each particle uses (7).

$$\text{if } (v_{ij}^{t+1} > v_j^{max}) \text{ then } v_{ij}^{t+1} = v_j^{max}$$
$$\text{if } (v_{ij}^{t+1} < v_j^{min}) \text{ then } v_{ij}^{t+1} = v_j^{min}, \quad (7)$$

While, the value of $\mathbf{v_j^{max}}$ is generated using equation 8 and $\mathbf{v_j^{min}}$ is the negative value of $\mathbf{v_j^{max}}$.

$$Vj\max = k\ \frac{(x_{j,\max} - x_{j,\min})}{2} + k \in [0,1] \quad (8)$$

Calculation cycle of velocity values *vi* and updated position *xi* will be repeated until maximum iteration is achieved. When the iteration is over, the best particles come out as the optimum solution.

C. Testing Methods

The model developed in this research will be tested using several methods such as Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). MAPE is calculated by making an average percentage of the absolute error of each predicted result. Thus, MAPE can indicate how much prediction error. MAPE is described in (9).

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{x_t - y_t}{x_t}\right| \quad (9)$$

MAE calculate the average of absolute error for each predicted result. MAE is useful when measuring errors in certain units. MAE values can be calculated using (10).

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \quad (10)$$

RMSE is used to calculate predicted performance by considering the prediction error of each data. RMSE formula can be seen there (11).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - p_i)^2} \quad (11)$$

V. Experiment and Result

The experimental process examines the parameters used on particle swarm optimization such as particle test, iteration test, and also inertia weight combination test.

The PSO algorithm generates population and initial velocity in the range of [0-100]. The range used has been tested from the number -1000 to 1000 and obtained that range 0-100 can provide highest fitness solutions. Particle test and iteration test for each model use a multiple of 100 in which the maximum particle test lies in 3000 particles, if the particles tested over 3000 require longer computation time. For each testing run 5 times, and the fitness value obtained from the average test results. The last test was a combination of inertia weight, performed to know the displacement velocity of each particle, inertia weight is tested in a range [0,1-0,9]. The result of each parameter testing is shown in Table 2.

TABLE II. Test Result of Parameter

M	Test Particles	Fitness	Iteration Test	Fitness	Inertia weight		Fitness
1	1800	39950.9474	700	186.704	0.8	0.4	2420.86
2	1800	825.9134	1900	45242.522	0.2	0.7	86434.266
3	500	139.68	1800	814.624	0.3	0.8	298492.2
4	2000	201506.91	500	69.38	0.2	0.7	2.126
5	2500	539040.066	1900	124.27	0.3	0.9	243.902
6	800	214060.584	600	297389.054	0.4	0.7	846.26
7	1900	236999.218	1800	581.986	0.4	0.9	388.75

M-1 represents Karang Besuki area, M-2 represents Tunggulwulung area, M-3 represents Lowokwaru area, M-4 represents Puncak Trikora area, M-5 represents Sumbersari area, M-6 represents Dinoyo area, and M-7 represents Manggar area. The experimental result shows that the fitness value based on data being tested. Furthermore, this research is better using more data.

After knowing the result of parameter testing, error values are calculated based on RMSE, MAE, and MAPE. Comparison of test values is shown in Table 3.

TABLE III. RESULT OF TESTING METHOD

Methods	Accuracy		
	MAPE	***MAE***	***RMSE***
Regression	4.84552	4.84552	2201253
Regression - PSO			
Model 1	0.73255	2837.2	14186
Model 2	0.0238	5520.95	44168
Model 3	0.02251	16635.9	99816
Model 4	5.84929	16798.2	67193
Model 5	0.42763	44950.7	179803
Model 6	0.07718	34153.1	170765
Model 7	0.0932	19830.8	79323

VI. CONCLUSION

In this paper, several tests have been performed using linear regression and particle swarm optimization methods to perform house price prediction. Based on the NJOP data of 9 houses, the system is modeling house price predictions into 7 models each of them represents one area. The area modeling includes Kelurahan Karang Besuki, Tunggulwulung, Lowokwaru, Puncak Trikora, Sumbersari, Dinoyo, Manggar. Based on the result from particle test, iteration test and inertia weight test can be concluded that M-1 represents Karang Besuki area get the best parameter for optimal prediction. Those best values of parameters obtained are 1800 particles, 700 iterations and of inertia weight 0.4 and 0.8 can get minimum prediction error RMSE as IDR 14.186. For the other model, the error prediction values are still large. Using different methods that match the time-series data will be used in the future research to obtain smaller error prediction values and using more data to get the better result.

REFERENCES

[1] R. M. A. van der Schaar, “Analysis of Indonesian Property Market; Overview and Foreign Ownership,” Investment Indonesian. 2015.

[2] The Central Bureau of Statistics, “Population Census,” 2015.

[3] W. T. Lim, L. Wang, and Y. Wang, “Singapore Housing Price Prediction Using Neural Networks,” Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov., vol. 12, pp. 518–522, 2016.

[4] Y. Feng and K. Jones, “Comparing multilevel modelling and artificial neural networks in house price prediction,” 2015 2nd IEEE Int. Conf. Spat. Data Min. Geogr. Knowl. Serv., pp. 108–114, 2015.

[5] R. Ghodsi, “Estimation of Housing Prices by Fuzzy Regression and Artificial Neural Network,” in Fourth Asia International Conference on Mathematical/ Analytical Modelling and Comuter Simulation, 2010, no. 1.

[6] A. Azadeh, B. Ziaei, and M. Moghaddam, “A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations,” Expert Syst. Appl., vol. 39, no. 1, pp. 298–315, 2012.

[7] F. S. Gharehchopogh, T. H. Bonab, and S. R. Khaze, “A Linear Regression Approach to Prediction of Stock Market Trading Volume: A Case Study,” Int. J. Manag. Value Supply Chain., vol. 4, no. 3, pp. 25–31, 2013.

[8] H.-I. Hsieh, T.-P. Lee, and T.-S. Lee, “A Hybrid Particle Swarm Optimization and Support Vector Regression Model for Financial Time Series Forecasting,” Int. J. Bus. Adm., vol. 2, no. 2, pp. 48–56, 2011.

[9] F. Marini and B. Walczak, “Particle swarm optimization (PSO). A tutorial,” Chemom. Intell. Lab. Syst., vol. 149, pp. 153–165, 2015.

[10] A. Hayder M. Albehadili Abdurrahman and N. . Islam, “An Algorith for Time Series Prediction Using,” Int. J. Sci. Knowl. Comput. Inf. Technol., vol. 4, no. 6, pp. 26–33, 2014.

[11] Y. P. Anggodo and W. F. Mahmudy, “Automatic Clustering and Optimized Fuzzy Logical Relationship for Minimum Living Needs Forecasting,” J. Environ. Eng. Sustain. Technol., vol. 4, no. 1, pp. 1–7, 2017.

[12] Y. P. Anggodo, W. Cahyaningrum, A. N. Fauziyah, I. L. Khoiriyah, K. Oktavianis, and I. Cholissodin, “Hybrid K-means Dan Particle Swarm Optimization Untuk Clustering Nasabah Kredit,” J. Teknol. Inf. dan Ilmu Komput., vol. 4, no. 2, pp. 1–6, 2017.

[13] Y. P. Anggodo, A. K. Ariyani, M. K. Ardi, and W. F. Mahmudy, “Optimation of Multi-Trip Vehicle Routing Problem with Time Windows using Genetic Algorithm,” J. Environ. Eng. Sustain. Technol., vol. 3, no. 2, pp. 92–97, 2017.

[14] R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartotoor, and I. B. Syamwil, “Factors influencing the price of housing in Indonesia,” Int. J. Hous. Mark. Anal., vol. 8, no. 2, pp. 169–188, 2015.

[15] V. Limsombunchai, “House price prediction: Hedonic price model vs. artificial neural network,” Am. J. …, 2004.

[16] D. X. Zhu and K. L. Wei, “The Land Prices and Housing Prices —— Empirical Research Based on Panel Data of 11 Provinces and Municipalities in Eastern China,” Int. Conf. Manag. Sci. Eng., no. 2009, pp. 2118–2123, 2013.

[17] S. Kisilevich, D. Keim, and L. Rokach, “A GIS-based decision support system for hotel room rate estimation and temporal price prediction: The hotel brokers’ context,” Decis. Support Syst., vol. 54, no. 2, pp. 1119–1133, 2013.

[18] C. Y. Jim and W. Y. Chen, “Value of scenic views: Hedonic assessment of private housing in Hong Kong,” Landsc. Urban Plan., vol. 91, no. 4, pp. 226–234, 2009.

[19] L. Bryant, “Housing affordability in Australia: an empirical study of the impact of infrastructure charges,” J. Hous. Built Environ., 2016.

[20] S. Rosen, “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition,” J. Polit. Econ., vol. 82, no. 1, pp. 34–55, 1974.

[21] J. Kennedy and R. Eberhart, “Particle swarm optimization,” 1995 IEEE Int. Conf. Neural Networks (ICNN 95), vol. 4, pp. 1942–1948, 1995.

[22] R. C. Eberhart and Y. Shi, “Comparing inertia weights and constriction factors in particle swarm optimization,” IEEE Congr. Evol. Comput., vol. 1, no. 7, pp. 84–88 vol.1, 2000.

[23] A. Ratnaweera, S. K. Halgamuge, and H. C. Watson, “Self-organizing hierarchical Particle swarm optimizer with time varying acceleration coefficients,” IEEE Trans. Evol. Comput., vol. 8, no. 3, p. 240–?255, 2004