2017 1st International Conference on Informatics and Computational Sciences (ICICoS)

Predicting House Sale Price Using Fuzzy Logic, Artificial Neural Network and K-Nearest Neighbor

Muhammad Fahmi Mukhlishin, Ragil Saputra, Adi Wibowo
Department of Informatics
Faculty of Science and Mathematics, Diponegoro University
Semarang, Indonesia
fahmi.mukhlish@gmail.com, ragil.saputra@undip.ac.id, bowo.adi@undip.ac.id

***Abstract*—Determining the value of land and home are regularly determined at the earliest by the seller, however determining the right price in the sales process will affect the buyer's desire to elect and bid. Special characteristics in Indonesia, tax object value (NJOP) and location parameters are high influence to the price. In this paper we proposed the prediction of land and house value using several methods. Fuzzy logic, Artificial Neural Network and K-Nearest Neighbor are compared in this paper to discover the most appropriate method that can be used as a reference for determining the price by the sellers. Google Maps is used to represent the spatial data for prediction parameter. The variables that used in the methods are NJOP of land, the locations, the age, NJOP of house, and the valuable location of the land. The experimental methods are tested by comparing between the real price transaction and the prediction using MAPE formula.**

Keywords—House Price, Fuzzy, Artificial Neural Networks, K-Nearest Neighbor.

I. INTRODUCTION

House is one of the important needs for people which has function as a place for rest and gather with family [1]. In the housing market, the initial prices are an important factor in the process of buying and selling houses and land. Determining the initial selling price of a house or land usually depends on the seller, however determining the right price in the sales process will affect the buyer's desire to bid and make selections. The initial price for each house and land are varied according to residential facilities and home geographical conditions. In Indonesia, one of the parameters to determine the selling price of the house that can be quantity calculated are the value of the tax object (NJOP) of the land and the value of the tax object of the building [2]. Both parameters are influenced by several factors such as strategic location and also age of building.

The initial house prices prediction is challenging and requires the best method to get the best prediction accuracy. In the predicting the sale price of a house that has an uncertainty parameter, fuzzy logic becomes one of the solutions that can be used in solving the problem [3-4]. Moreover, artificial neural network methods are used to predict house selling prices [5]. In addition to using fuzzy logic and artificial neural networks, predictions can also use the K-Nearest Neighbors algorithm, for estimating residential prices for the residential property market in Hong Kong [6]. Several machine learning methods are compared to get the best prediction of house pricing [7].

In advance, spatial analysis is used to determine house and land prices [8]. Jeffrey et. al discussed the spatial effects on the dynamics of house prices in the USA that have a significant influence in urban house price growth rates [9]. In addition, real estate pricing in Vienna (Austria) was investigated that requires recognition of spatial heterogeneity in housing prices [10]. However, only few studies that apply the prediction model using spatial data as analysis factor for the prediction.

In this study, several prediction methods were compared to find out the best predicted results for determining the selling price of a house compared to the real price. The methods (Fuzzy, Neural Network, k-nearest neighbor) were tested to predict house prices in Indonesia. In this study, spatial data and attribute data are employed as determinant parameters in the determination of house and land prices. Spatial data consists of the selling value of the tax object (NJOP) of the land and the strategic location of the land, while the attribute data includes the selling value of tax object (NJOP) of the building, the condition of the house and the age of the house. This study uses spatial data in the form of a strategic location of the land, called accessibility level that influence by infrastructure facilities such as public health services, education, downtown and economy. In spatial data the Geographic Information System (GIS) has three elements, namely vertices, lines (arcs), and broad (polygons) in the form of vectors or raster representing geometric topology, size, shape, position and direction.

II. METHOD

A. Fuzzy Logic

Fuzzy logic was first introduced by Lotfi Zadeh in 1965. Fuzzy logic is a mathematical framework used to represent uncertainty. The presentation of set A with the zero-one membership function, otherwise known as the crisp set can be represented by the following equation [12].

$$\mu_{(A)}(x)=\{ \_(0 \quad \text{if} \quad x \notin A)\hat{}(1 \quad \text{if} \quad x \in A) \quad (1)$$

The above equation shows that the Fuzzy set is a generalization of the crisp set by allowing the membership function to retrieve the values in the interval [0, 1]. In other words, the membership function of the crisp set can only take a

value of zero and one, while the membership function of the fuzzy set is a continuous function with the range [0, 1] [11].

B. Fuzzy Tsukamoto

The Fuzzy Tsukamoto model is one of the models in the fuzzy inference system shown in Figure 2.7. [12]. The Inference System is a process of reasoning about a given state, using all available knowledge to produce the best estimates of output. Fuzzy Inference System (FIS) or often called fuzzy expert system is a popular computing framework based on fuzzy set theory, IF-THEN fuzzy rules, and fuzzy reasoning. In the fuzzy system, the inference engine is used to adjust the fuzzy set pattern of the inputs with the antecedents of all fuzzy rules and combine all responses to produce the fuzzy set of output [12]. In the Fuzzy Tsukamoto model, any consequent IF-THEN-shaped rules should be represented by a fuzzy set with a monotonous membership function. Here are the steps of Fuzzy Tsukamoto's model [12].

1) Calculate the value of α-predicate (fire strength) of each rule during the process of evaluation of rules in the inference engine by using the MIN implication function.
2) Counting the inference results expressly (crisp) of each rule (z_1, z_2, ..., z_n). The calculation is based on the fire strength value of each rule (w_1, w_2, ..., w_n).
3) The defuzzification process uses the weighted average method .

C. Artificial Neural Network

Artificial Neural Networks are one of the artificial representations of the human brain that always simulate the learning process in the human brain. The term artificial here is used because this neural network is implemented by using a computer program capable of completing a number of calculation processes during the learning process. [13]

The back-propagation algorithm for the training process using in feed forward neural network algorithm are employed in this study. The back propagation neural network algorithm apply to do looping at two stages of propagation and repeated, until achieved acceptable results in training. Moreover, in the feed forward neural network, the information moves forward, one direction from the input to the output (via a hidden node) without the loop.

D. K-Nearest Neighbor(K-NN)

The K-NN algorithm is one of the methods used for classification analysis, but the last few decades the KNN method has also been used for prediction [14]. K-Nearest Neighbor algorithm is a method to classify objects based on learning data closest to the object. Nearest Neighbor is an approach to finding cases by calculating the proximity between the new case and the old case that is based on matching the weights of a number of existing features [14].

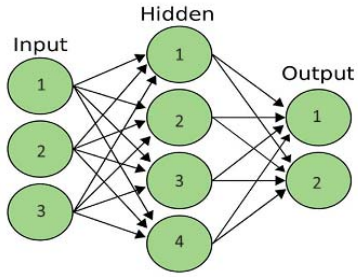

Fig. 1. Artificial Neural Networks

The working principle of K-Nearest Neighbor is to find the closest distance between the data to be evaluated with the nearest neighbor in the training data. Training data is projected into many-dimensional spaces, of which each dimension explains the features of the data. This space is divided into sections based on training data classification. A point in this space is denoted by class c, if class c is the most commonly encountered classification of the nearest neighbor of the point [14].

E. Geographic Information System

According to S. Murai, Geographic Information System (GIS) is an information system used to enter, store, recall, process, analyze and produce geo-referenced data or geospatial data, to support decision making in land use planning and management, natural resources, environment, transportation, City facilities, and other public services [15].

Geospatial data sources are digital maps, aerial photographs, satellite images, statistics tables, and other related documents. Geospatial data can be divided into graphical data called geometric data and attribute data (thematic data). Graphic data has three elements, namely the point (node), the line (arc), and the area (polygon) in the form of vectors or raster representing the geometry of topology, size, shape, position, and direction [15].

F. Google Maps API

Google Maps uses HTML and Javascript programming languages to make it possible to place Google Maps apps on another web. The Google Maps API is a feature issued by google to facilitate the user integrating Google Maps into the user's website. In order for Google Maps to appear on the web as desired, it needs API-key. API-key is a special code generated by google for a particular website, so the Google Maps server can recognize [16].

G. Mean Absolute Percentage Error

Mean Absolute Percentage Error (MAPE) Is the average of the overall percentage error (difference) between the actual data with the data forecasting results. The accuracy measure is matched with time series data, and is shown in percentages with the MAPE formula as follows:

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|x_i - y_i|}{x_i} \times 100\%}{n}$$

MAPE = Average absolute percentage error

x_i = Value of Real Transaction
y_i = Prediction Value
n = Number of data

III. PROPOSED MODEL

The real house prices in Pedurungan Sub-district of Semarang City are used in this study. The training data consists of 7 predictor's attributes and 1 label attribute. In order to improve the data quality and eliminating duplication, preprocessing was applied.

In our proposed method, the determination of the initial selling price of the house and land are considered by the value of sales value of taxable object land (NJOP-L), sales value of taxable object building / house (NJOP-B), house age, house condition and land strategic location. The prediction system is equipped with Google Maps to facilitate in determining the strategic location of the land, get the sales value of taxable object price of land and the strategic location value of the land.

A. Fuzzy Logic

In this experiment, intuition is used to determine the membership function. Intuition is based on human intelligence and deep understanding to develop membership functions. Determination of the shape or type of curve, and the overlap between curves used adjusted based on research that has been done that is represented in Fig 2-8. The fuzzy set of input variables and output variables of the house and land sale pricing system are as follow:

1) Input Variables

a) Sales Value of Taxable Object Building (NJOP-B) Variables : input variables obtained from sales value of taxable object data building set by the Government of Semarang City

b) Sales Value of Taxable Object Land (NJOP-L) Variables: input variables obtained from sales value of taxable object data of land determined by Semarang City Government represented in map form which can be seen on Fig.9.

c) House Age Variables : input variables obtained from the calculation of the age of the house refer to the building establishment permit (IMB) as shown in Fig. 4.

d) House Condition Variables: input variables obtained from the state and structure of the house as a whole according to assumptions of homeowners and assumptions of home buyers that assessed in the percentage as shown in Fig. 5.

e) Strategic Value of Land Location Variables*:* input variables that can be represented as the level of accessibility locations as shown in Fig. 6.

2) Ouput Variables

a) Predicted House Prices Variables: the output variable that became the initial output of the first fuzzy calculation on the house price calculation as shown in Fig. 7.

b) Predicted Land Prices Variables: the output variable that becomes the output of the second fuzzy calculation on the house price calculation as shown in Fig. 8.

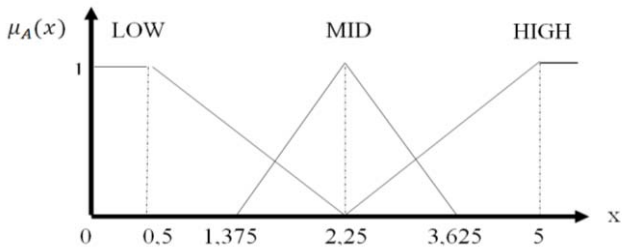

Fig. 2. Membership fuction of NJOP-L

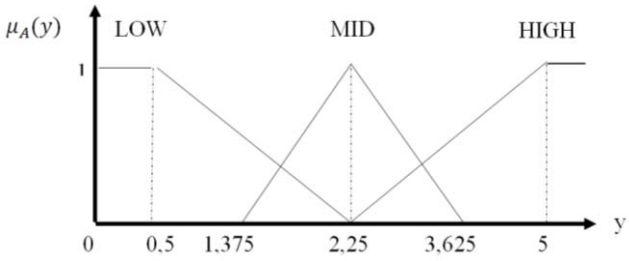

Fig. 3. Membership fuction of NJOP-B

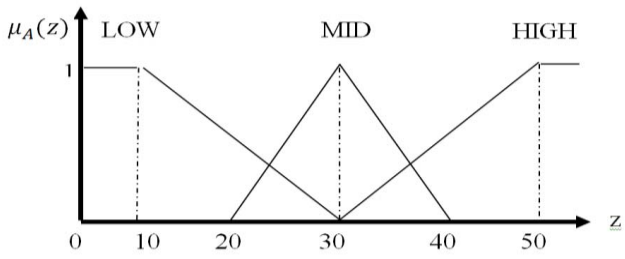

Fig. 4. Membership Function of House Age

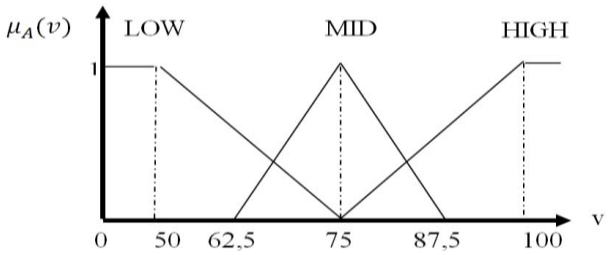

Fig. 5. Membership Function of House Condition

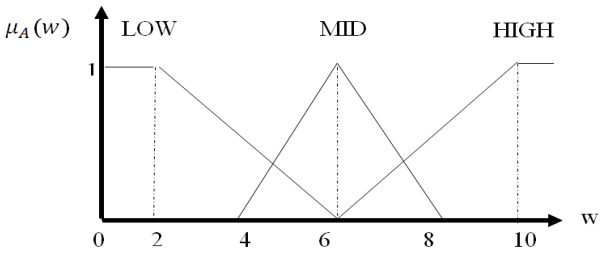

Fig. 6. Membership Function of Land Location Strategic Value

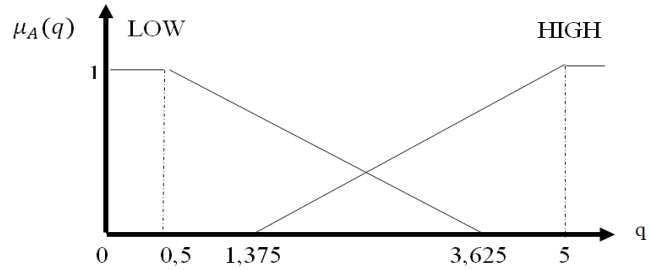

Fig. 7. Membership Function of Predicted House Prices Value

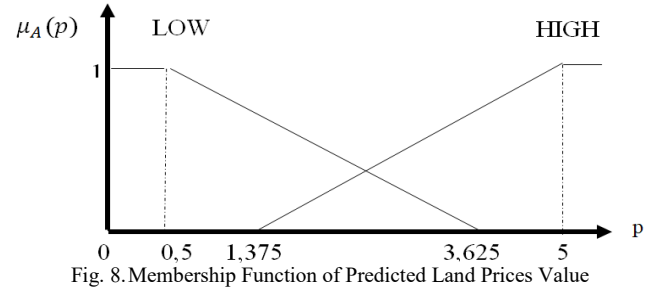

Fig. 8. Membership Function of Predicted Land Prices Value

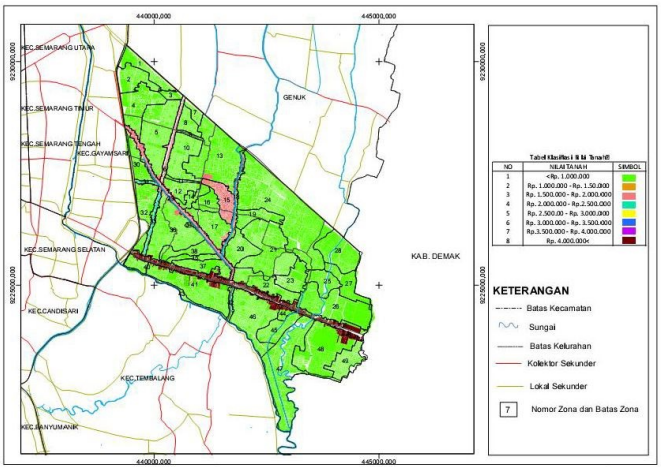

Fig. 9. Map of Land Classification Based on sales value of taxable object Price

The fuzzy rules are operated using a series of IF-THEN statements. In the first fuzzy rules there are 3 input variables; sales value of taxable objects building (NJOP-B), house age and house condition. The output variable for this fuzzy rule is the house price prediction. Moreover, two input variables are employed in the second fuzzy rules; sales value of taxable object land (NJOP-L) and the strategic value of land location. The output of the second fuzzy rule is the land price prediction. The number of statements on each rule base is the multiplication of the number of input sets. The first fuzzy rule has 27 statements formed from 3 input variables with each variable having 3 sets LOW, MID and HIGH as shown in Table I. The second fuzzy rule has 9 statements formed from 2 input variables with each variable having 3 sets LOW, MID, and HIGH as shown in Table II.

B. Artificial Neural Network

In this study, the artificial neural network architecture built consists of 3 layers, namely: the input layer, a hidden layer, and output layer. Each layer is associated with weights.

TABLE I. FIRST RULEBASE FUZZY

No	Sales Value of Taxable Object Building	House Age	House Condition	Output Predicted House Prices
1	LOW	LOW	LOW	LOW
2	LOW	MID	LOW	LOW
3	LOW	HIGH	LOW	LOW
4	LOW	LOW	MID	LOW
5	LOW	MID	MID	LOW
6	LOW	HIGH	MID	LOW
7	LOW	LOW	HIGH	HIGH
8	LOW	MID	HIGH	LOW
9	LOW	HIGH	HIGH	LOW
10	MID	LOW	LOW	HIGH
11	MID	MID	LOW	LOW
12	MID	HIGH	LOW	LOW
13	MID	LOW	MID	HIGH
14	MID	MID	MID	LOW
15	MID	HIGH	MID	LOW
16	MID	LOW	HIGH	HIGH
17	MID	MID	HIGH	HIGH
18	MID	HIGH	HIGH	LOW
19	HIGH	LOW	LOW	HIGH
20	HIGH	MID	LOW	HIGH
21	HIGH	HIGH	LOW	LOW
22	HIGH	LOW	MID	HIGH
23	HIGH	MID	MID	HIGH
24	HIGH	HIGH	MID	LOW
25	HIGH	LOW	HIGH	HIGH
26	HIGH	MID	HIGH	HIGH
27	HIGH	HIGH	HIGH	HIGH

TABLE II. SECOND RULEBASEFUZZY

No	Sales Value of Taxable Object Land	Strategic Location Land	Output Predicted Land Prices
1	LOW	LOW	LOW
2	LOW	MID	LOW
3	LOW	HIGH	HIGH
4	MID	LOW	LOW
5	MID	MID	HIGH
6	MID	HIGH	HIGH
7	HIGH	LOW	LOW
8	HIGH	MID	HIGH
9	HIGH	HIGH	HIGH

Training process used back-propagation neural network with parameter: MSE 0.00001, 0.3 learning rate, 500 epochs, and 0.3 momentum. The spatial data is converted to value in this method for perdition using neural networks.

C. K – Nearest Neigbor

In this experiment, the similar data training and testing with neural networks are employed. The k-Nearest Neighbor algorithm is based the comparing a given test example with training examples. The k-nearest neighbor is applied using k=1.

IV. RESULTS

The comparing between the real price and the prediction price are employ in the testing experiments using MAPE. The real price is collected from the real transactions from the property agent.

.

TABLE III. PREDICTION PRICING TEST

No	Transaction Price	Fuzzy Pred.	Percent Error	Suitability	ANN. Pred	Percent Error	Suitability	KNN Pred.	Percent Error	Suitability
1	425	531	25%	75%	504	19%	81%	600	41%	59%
2	355	433	22%	78%	461	30%	70%	550	55%	45%
3	800	679	15%	85%	686	14%	86%	1500	88%	13%
4	1400	1017	27%	73%	1191	15%	85%	1300	7%	93%
5	1300	1311	1%	99%	734	44%	56%	1300	0%	100%
6	650	646	1%	99%	658	1%	99%	1500	131%	-31%
7	1000	1017	2%	98%	1244	24%	76%	1100	10%	90%
8	600	646	8%	92%	492	18%	82%	550	8%	92%
9	550	526	4%	96%	771	40%	60%	1500	173%	-73%
Average			12 %	88%		23 %	77%		57 %	43%

TABLE IV. PREDICTION PRICING TESTWITHOUT VARIABLE STRATEGIC LOCATION LAND

No	Transaction Price	Fuzzy Pred.	Percent Error	Suitability	ANN. Pred	Percent Error	Suitability	KNN Pred.	Percent Error	Suitability
1	425	428	1%	99%	730	72%	28%	600	41%	59%
2	355	306	14%	86%	565	59%	41%	550	55%	45%
3	800	554	31%	69%	833	4%	96%	1500	88%	13%
4	1400	1131	19%	81%	1314	6%	94%	1300	7%	93%
5	1300	1354	4%	96%	1858	43%	57%	1300	0%	100%
6	650	548	16%	84%	833	28%	72%	1500	131%	-31%
7	1000	900	10%	90%	1220	22%	78%	1100	10%	90%
8	600	520	13%	87%	1747	191%	-91%	550	8%	92%
9	550	520	5%	95%	1074	95%	5%	1500	173%	-73%
Average			13 %	87%		58 %	42%		57 %	43%

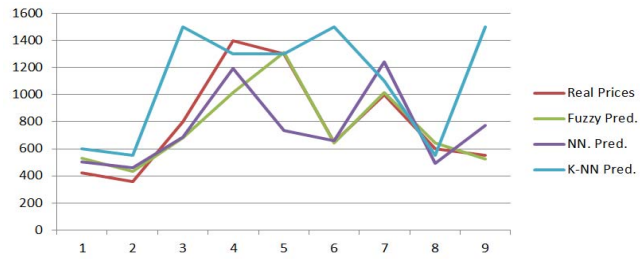

Fig. 10. Prediction Pricing Test Chart

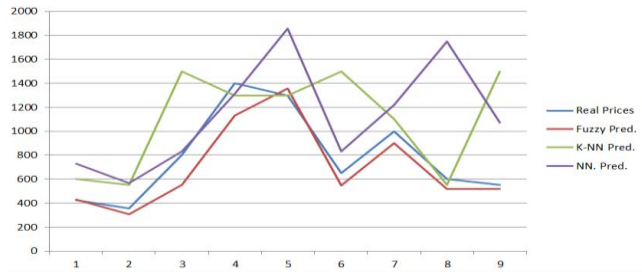

Fig. 11. Prediction Pricing TestWithout Variable Strategic Location Land

In table III, the prediction experiment using all input variables are compared with the real prices. In this experiment, the fuzzy method produces the best accuracy with 88%, a significant difference compared to the artificial neural network method with 77% and the k-nearest neighbor method with 43%. The second experiment, the strategic location of the land is removed as input parameter. Based on the experiments results as show in Table IV, the fuzzy method still has the best performance compare to ANN and k-NN. The comparison graph of prediction price shows in Fig 11 and 12. Therefore, fuzzy method has high accuracy due to un-employing training method in the method modeling. This is extremely overturned with ANN and k-NN that are heavily affected by training process. In this study we only use 18 training data that represented several conditions.

.

V. CONCLUSIONS

In this study, fuzzy, artificial neural network and k-nearest neighbor are used to predict the selling price of a house. We tested the performance of this method by measuring how accurate the results of the predictions were generated compared to the real house selling price using MAPE. The spatial data and attribute data are used in the prediction system. Based on the experiment results show that the fuzzy method is superior to neural networks as well as k-nearest neighbor for the house price prediction in limited data training. In future, the optimization of fuzzy rule and increasing data training are required for improving the prediction accuracy.

REFERENCES

[1] Raharjo, N. P. “Dinamika pemenuhan Kebutuhan Perumahan Masyarakat Berpenghasilan rendah”. Semarang: Magister Teknik Pembangunan wilayah dan Kota Universitas Diponegoro, p. 30. 2010.

[2] Fahirah, F., Basong, A. & Tagala, H. H. “Identifikasi Faktor yang Mempengaruhi Nilai Jual Lahan dan Bangunan pada Perumahan Tipe Sederhana”. Jurnal Smartek, Volume 4, pp. 251 - 269 . 2010.

[3] Kuşan, Hakan, Osman Aytekin, and İlker Özdemir. "The use of fuzzy logic in predicting house selling price." Expert systems with Applications 37, no. 3, pp. 1808-1813. 2010.

[4] Gerek, Ibrahim Halil. "House selling price assessment using two different adaptive neuro-fuzzy techniques." Automation in Construction 41, pp. 33-39. 2014.

[5] Peterson, Steven, and Albert Flanagan. "Neural network hedonic pricing models in mass real estate appraisal." Journal of Real Estate Research 31, no. 2, pp. 147-164. 2009

[6] Cheung, Simon KC, and Sahminan Sahminan. "A Localized Model for Residential Property Valuation: Nearest Neighbor with Attribute Differences." International Real Estate Review 20, no. 2, pp. 221-250. 2017.

[7] Byeonghwa Park and Jae Kwon Bae, "Using Machine Learning Algorithms For Housing Price Prediction: The Case Of Fairfax County, Virginia Housing Data," Expert Systems with Applications, vol. XLII, pp. 2928-2934, 2015.

[8] Basu, Sabyasachi, and Thomas G. Thibodeau. "Analysis of spatial autocorrelation in house prices." The Journal of Real Estate Finance and Economics 17, no. 1, pp. 61-85. 1998

[9] Cohen, Jeffrey P., Yannis M. Ioannides, and Win Wirathip Thanapisitikul. "Spatial effects and house price dynamics in the USA." Journal of Housing Economics 31, pp. 1-13. 2016.

[10] Helbich, Marco, and Daniel A. Griffith. "Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches." Computers, Environment and Urban Systems 57, pp. 1-11. 2016

[11] Klir, George, and Bo Yuan. Fuzzy sets and fuzzy logic. Vol. 4. New Jersey: Prentice hall, 1995.

[12] Sivanandam, S., Sumanthi, S. & Deepa, S. *Introduction to Fuzzy Logic Using Matlab.* Heidelberg: Springer-Verlag Berlin. 2007.

[13] Sayyed Mohsen Vazirizade, Saeed Nozhati, and Mostafa Allameh Zadeh, "Seismic Reliability Assessment Of Structure Using Artificial Neural Network," *Journal of Building Engineering*, vol. XI, pp. 230-235, 2017.

[14] Larose, Daniel T. "K-nearest neighbor algorithm." Discovering Knowledge in Data: An Introduction to Data Mining, pp. 90-106. 2005

[15] Murai,*GIS Work Book.* Tokyo: Institute of Industrial Science. 1999.

[16] Chtiara, C. Implementasi sistem sistem informasi geografis (SIG) Universitas Indonesia (UI) berbasis Web dengan menggunakan google maps api. Universita Indonesia. Jakarta: Jurnal UI. 2008