Jurnal Matematika UNAND
Vol. **15** No. **2** Hal. 163 – 178
Edisi April 2026
ISSN : 2303–291X
e-ISSN : 2721–9410

https://doi.org/10.25077/jmua.15.2.163-178.2026

LOCALLY WEIGHTED KNN-BASED FUZZY REGRESSION FOR PROPERTY VALUATION UNDER MARKET UNCERTAINTY

HAZMIRA YOZZA[a,b], RISWAN EFENDI[a,c,*], NOR AZAH SAMAT[a], IZZATI RAHMI[b], RIDHO SAPUTRA[b]

[a] *Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjung Malim, Perak, Malaysia,*
[b] *Department of Mathematics and Data Science, Faculty of Mathematics and Natural Sciences, Andalas University, 25163 Padang, Indonesia,*
[c] *Department of Mathematics, Faculty of Science and Technology, UIN Sultan Syarif Kasim Riau, 28293 Pekanbaru, Indonesia.*
email : riswanefendi@fsmt.upsi.edu.my

Received: February 27, 2026 Received in revised form: March 3, 2026
Accepted: March 9, 2026 Available online: April 30, 2026

Abstract. *Accurate evaluation of house valuation is crucial, as misestimation of house prices can lead to serious consequences for various stakeholders. House prices can be modeled as a function of their constituent attributes; however, they are inherently fuzzy due to negotiation processes and unpredictable market conditions. This study aims to develop predictive rules for triangular fuzzy numbers of house prices for new properties by implementing a locally weighted k-nearest neighbor (KNN)-based fuzzy regression approach and to compare its performance with possibilistic fuzzy regression. The dataset used in this study is a house valuation dataset. The analysis employs the modified Cheng and Lee k-nearest neighbor fuzzy regression and Tanaka's possibilistic fuzzy regression. The results indicate that the modified Cheng and Lee k-nearest neighbor fuzzy regression outperforms possibilistic fuzzy regression in predicting triangular fuzzy of house prices. The best predictive performance is achieved when the modified Cheng and Lee approach is implemented with $k = 29$ nearest neighbors, Minkowski distance with exponent parameter $p = 1.6$, and an unequal weighting scheme with $r = 1$.*

Keywords: Fuzzy regression, house price, triangular fuzzy number modified Cheng and Lee KNN fuzzy regression possibilistic fuzzy regression

1. Introduction

In the field of real estate economics, the valuation of residential properties is a central research concern since housing markets are heterogeneous and house prices are influenced by a range of structural, location, and economic factors [1,2,3,4,5]. The

*Corresponding author

assessment of housing prices is crucial for various stakeholders, including lenders, developers, and policymakers engaged in urban planning and risk evaluation, as well as individual buyers and sellers. A highly accurate and reliable estimation of house prices is important as it has an impact on various economic decisions and risk management for a range of stakeholders. Sellers and real estate professionals can optimize their marketing strategies and prevent market inefficiencies and mispricing by utilizing an accurate house price estimation. Buyers will be able to evaluate affordability and negotiate more effectively with the assistance of a precise house price prediction.

On the contrary, a misestimation of house prices brings serious consequences for both buyers and sellers. Sellers may incur capital losses when their house is undervalued. This circumstance may lead to rejected offers or missed opportunities for purchasers if the market price is higher than anticipated. Conversely, overestimation may cause the buyers to overpay compared to its actual market value. In addition, buyers will miss affordable opportunities due to perceived high house prices. Furthermore, erroneous valuation influences financial decisions as increased value uncertainty correlates with a heightened likelihood of mortgage rejection, elevated interest rates, and a diminished loan-to-value ratio [6].

The appraisal strategy is a conventionally recognized approach for assessing housing prices. This method depends only on the appraiser's judgement and data from comparable sold houses. Recent studies emphasize the implementation of automated valuation methods to diminish subjectivity and enhance uniformity and efficiency in property value [7]. These models offer more methodical estimations through the application of statistical approaches.

From an economic standpoint, housing prices can be modeled due to their strong correlation with numerous quantifiable characteristics. House prices are determined by a combination of structural elements, including house size, location, and environmental aspects that buyers inherently evaluate in the market. Consequently, a property's market worth is perceived as a consequence of its characteristics, including physical attributes, accessibility, and local amenities. Taking these attributes into account, numerous researchers, including [8,9,10,11,12], employed hedonic regression and various machine learning techniques, such as random forest, neural networks, and support vector regression, to model the correlation between house prices and the quantifiable characteristics of the property and its environment.

Despite its advantages, the majority of methods treat house prices as single-point or crisp observations. In reality, the house market price determination involves a negotiation process, subjective perception of environmental quality, and imperfect measurement of accessibility. Given these factors, a crisp house price can not capture the uncertainty in house prices. As a result, hedonic regression and other machine learning approaches fail to model the inherent uncertainty contained in house prices. Under these circumstances, fuzzy regression provides a more realistic representation of house price behaviour.

The purpose of this study is to build a model or identify rules to predict the TFN representing the price of a given new house by implementing locally weighted

KNN-based fuzzy regression, and to compare its performance with the possibilistic fuzzy regression.

2. Theoretical Framework

2.1. *Fuzzy Number*

Let X be a crisp set of objects, and its generic elements are represented as x. This set is usually commonly referred to as the "universe". The membership of x in a subset A of X is considered a characteristic function from X to a valuation set $\{0, 1\}$ such that:

$$\mu_A(x) = \begin{cases} 1 & x \in A, \\ 0 & x \notin A. \end{cases} \tag{2.1}$$

[13]. If the valuation set is a real interval $[0, 1]$, A is called a fuzzy set [14].

Definition 2.1. *[15,16] If X is a collection of objects denoted by x, then a fuzzy set $\widetilde{A}$ in X is a set of ordered pair*

$$\widetilde{A} = \{(x, \mu_{\widetilde{A}}(x)) | x \in X\}, \tag{2.2}$$

where $\mu_{\widetilde{A}}(x)$ is called the membership function or degree of membership of x in $\widetilde{A}$ that maps X to the membership space M. The membership degree ranges from $[0, 1]$. Degrees "0" and "1" denote the absence and complete existence of an element in a set. The intermediate values indicate that an element is only partially present in a set [15]. When M consists of only 0 or 1, then $\widetilde{A}$ is a crisp/nonfuzzy set [16].

Definition 2.2. *[13,15] A fuzzy number $\widetilde{A}$ is a fuzzy set on $\mathbb{R}$, such that:*

(1) $\widetilde{A}$ is piecewise continuous,
(2) $\widetilde{A}$ is convex,
(3) $\widetilde{A}$ is normal, namely if m is a mean value of the fuzzy number $\widetilde{A}$, then $\mu_{\widetilde{A}}(m) = 1$. It has exactly one m, where $\mu_{\widetilde{A}}(m) = 1$,
(4) $\widetilde{A}$ is monotone ascending in the interval $(-\infty, m)$ and monotone descending in the interval (m, ∞).
(5) The α-cut of the fuzzy number, $\widetilde{A}_\alpha$, is a closed interval for $\alpha \in [0, 1]$.

Definition 2.3. *[17] A fuzzy number $\widetilde{A}$ is called a triangular fuzzy number (TFN) if its membership function $\mu_{\widetilde{A}}(x)$ is given by:*

$$\mu_{\widetilde{A}}(x) = \begin{cases} 0, & x < a_l \text{ and } x > a_u, \\ \frac{x - a_l}{a - a_l}, & a_l \le x \le a, \\ \frac{a_u - x}{a_u - a}, & a \le x \le a_u. \end{cases} \tag{2.3}$$

The triplet $\widetilde{A} = (a_l, a, a_u)$ is typically used to denote the TFN $\widetilde{A}$. Let $c_L = (a - a_l)$ and $c_R = (a_u - a)$ be left and right spread of TFN $\widetilde{A}$. If $c_L = c_R = c$, TFN is said to be symmetrical, otherwise asymmetrical. A symmetrical TFN is commonly presented by $\widetilde{A} = (a, c)$ with c is the spread of symmetrical TFN.

2.2. *Fuzzy Regression*

In many real-world problems, information is often fuzzy, imprecise, or even described linguistically. It is also often found that the relationships among variables are unclear. In this situation, single-value observations are inadequate, and the ordinary regression becomes limited. Fuzzy regression was developed to model the relationship between an output variable and a set of input variables in a fuzzy environment. It captures the fuzziness in both the data and the relationship. Two general approaches are used to estimate the coefficients of a fuzzy regression model: the possibilistic and the least-squares approach.

Possibilistic regression was first proposed by [18] and later improved by [19,20, 21,22]. This method integrates regression concepts with fuzzy set theory. The basic model is:

$$\widetilde{Y} = \widetilde{A}_0 + \widetilde{A}_1 X_1 + \widetilde{A}_2 X_2 + \cdots + \widetilde{A}_k X_k, \tag{2.4}$$

where X_j's are input variables that may be crisp (nonfuzzy) or fuzzy. $\widetilde{A}_j$'s are fuzzy regression coefficients which are expressed as a TFN. Accordingly, the output of the model, $\widetilde{Y}$, is also a TFN.

The $\widetilde{A}_j$'s are assumed symmetrical TFN, as $\widetilde{A}_j = (a_j, c_j)$. As a consequence, $\widetilde{Y}_i = (m_i, w_i)$, for $i = 1, 2, \cdots, n$ with mode, and spread are $m_i = \sum_{j=0}^{k} a_j x_{ij}$ and $w_i = c_0 + \sum_{j=1}^{k} c_j |x_{ij}|$, respectively. The fuzzy coefficients are estimated by solving a linear programming problem aimed to minimizing the total spread of all coefficients, which is formulated as:

$$Z = \min c_0 + \sum_{j=1}^{k} c_j x_{ij}, \tag{2.5}$$

subject to:

$$\begin{cases} \left(a_0 + \sum_{j=1}^{m} a_j x_{ij}\right) - (1-h)\left(c_0 + \sum_{j=1}^{m} c_j x_{ij}\right) \leq (y_i - (1-h)\, e_i, \\ \left(a_0 + \sum_{j=1}^{m} a_j x_{ij}\right) + (1-h)\left(c_0 + \sum_{j=1}^{m} c_j x_{ij}\right) \geq (y_i - (1-h)\, e_i, \end{cases} \tag{2.6}$$

where $c_j \geq 0$, for $i = 1, 2, \cdots, n$ and $j = 0, 1, \cdots, m$.
The constraints in (2.6) are usually referred to as inclusion constraints. In this formulation, h denotes the h-factor that controls the width of the estimated fuzzy band produced from this formulation.

2.3. *k-Nearest Neighbor Regression*

The k-nearest neighbor (KNN) regression is a supervised machine learning technique, typically used to predict the response of a new observation. This approach works by exploring the entire training dataset to identify the k most similar instances, which serve as the basis for predicting the output of a new observation [23]. The value of the new point is assigned based on how closely it resembles other training data examples. KNN regression has two approaches. First, calculate the average of the outputs of k selected nearest neighbors. The second is to compute an inverse-distance-weighted average of the KNN [24].

Let $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be a training dataset that is formed by N samples, each sample $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{im})$ is a vector of input that contains m features, and y denotes the output variable. If a new query sample z is given, the predicted value of the output variable for z is determined using the following steps.

(1) Compute the distance from z to $\mathbf{x}_i$ for all i;
(2) Set k and identify a set of k-nearest neighbors from the z to X;
(3) Predict the output value of z by determining a weighted average of the input values of its nearest neighbors, formulated by:

$$\hat{y}(z) = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}, \tag{2.7}$$

where w_i denotes the weight assigned to i-th observation in the training dataset that is used to estimate the desired variable.

In general, the weight assigned to the i-th observation is calculated by:

$$w_i = \frac{d_{i(rel)}}{\sum_{i=1}^{k} d_{i(rel)}}, \tag{2.8}$$

where

$$d_{i(rel)} = \left(\frac{\sum_{i=1}^{k} d_i}{d_i} \right)^p, \tag{2.9}$$

and d_i is the distance between the independent variables of z and the i-th sample, and p is the power parameter that considers different forms of weight–distance relationships [25,26].

The metric distance is crucial to the KNN algorithm. The Minkowski distance is a widely used metric for measuring the closeness of two observations. Suppose the i-th and j-th observations are denoted as $\mathbf{X}_i$ and $\mathbf{X}_j$, respectively, where $\mathbf{X}_i = (x_{i1}, x_{i2}, \cdots, x_{im}) \in \mathbb{R}^m$ and $\mathbf{X}_j = (x_{j1}, x_{j2}, \cdots, x_{jm}) \in \mathbb{R}^m$. The Minkowski distance between $\mathbf{X}_i$ and $\mathbf{X}_j$ is defined as [27]:

$$d_{MD}(\mathbf{X}_i, \mathbf{X}_j) = \left(\sum_{k=1}^{m} |x_{ik} - x_{jk}|^p \right)^{1/p}, \ p \geq 1. \tag{2.10}$$

Different distance measures will be obtained by setting different p. The Manhattan distance is specified by setting $p = 1$. The Euclidean distance is a Minkowski distance with $p = 2$.

2.4. *Modified Cheng and Lee k-Nearest Neighbor Fuzzy Regression*

The KNN-based fuzzy regression algorithm was developed by [27] as a non-parametric approach to a fuzzy dataset whose fuzzy output variable is presented in an interval (y_l, y_u). A slight modification of this algorithm is performed by [28] by replacing the original interval-based outcome with a symmetrical TFN.

Suppose a training dataset consists of n tuples of data, $(\mathbf{X}_i, \widetilde{Y}_i); i = 1, 2, \cdots, n$ where $\mathbf{X}_i = (X_{i1}, X_{i2}, \cdots, X_{im})$ is a vector of crisp input variables and $\widetilde{Y}_i$ is a

fuzzy output variable that is presented as a symmetrical TFN, denoted as $\widetilde{Y}_i = (y_l, y, y_u) = (y - e, y, y + e)$. Suppose there is a new observation with an unknown TFN for its outcome variable, denoted as z. The modified Cheng and Lee KNN fuzzy regression determines the TFN of output for z using the following algorithm.

(1) Calculate the appropriate similarity/dissimilarity metric between the z and all observations in the training data.
(2) Set k and identify k nearest neighbors of z.
(3) Based on its k-nearest neighbors estimate the TFN of z's output by determining the weighted average of the mode and spread of TFNs of its nearest neighbors, formulated as:

$$\hat{y}_z = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}, \tag{2.11}$$

$$\hat{e}_z = \frac{\sum_{i=1}^{k} w_i e_i}{\sum_{i=1}^{k} w_i}, \tag{2.12}$$

respectively, where w_i is a weighting factor formulated as Eqs (2.8) and (2.9). Subsequently, the lower and upper boundaries of the TFN are calculated from $\hat{y}_{lz} = \hat{y}_z - \hat{e}_z$ and $\hat{y}_{uz} = \hat{y}_z + \hat{e}_z$, respectively [28].

2.5. *Performance Measure*

The accuracy of the predicted fuzzy output is calculated based on the distance between the membership functions of two outputs: the observed output FN, denoted by $\widetilde{Y} = (y_l, y, y_u)$, and the predicted output TFN, denoted by $\widehat{\widetilde{Y}} = (\hat{y}_l, \hat{y}, \hat{y}_u)$. Suppose $\mu_{\widetilde{y}_i}(x)$ and $\mu_{\widehat{\widetilde{y}}_i}(x)$ are membership functions of $\widetilde{Y}$ and $\widehat{\widetilde{Y}}$. The relative distance of fitting two fuzzy membership functions is defined as the difference between those fuzzy membership functions, which is standardized to the observed membership function. Generally, it is formulated as [29]:

$$D_r = \frac{\int_{S_{\widetilde{Y}} \cap S_{\widehat{\widetilde{Y}}}} \left| \mu_{\widetilde{y}_i}(x) - \mu_{\widehat{\widetilde{y}}_i}(x) \right| dx}{\int_{S_{\widetilde{Y}}} \mu_{\widetilde{y}_i}(x) dx}. \tag{2.13}$$

The mean relative distance (MRD) is calculated by averaging the relative distance across all observations.

3. Data and Method

3.1. *Data*

This study used a historical market dataset of real estate valuations collected from the Sindian District in New Taipei City, Taiwan. It comprises 288 observations and six input variables. The dataset is secondary data. It was collected by [30] and published at the UCI Machine Learning Repository. The output is the house price per unit area (in 10,000 Taiwan Dollars per 1 Ping). Originally, this variable is crisp. The crisp input variables are:

(1) X_1 = house age (in years).
(2) X_2 = distance to the nearest MRT station (in meters), measured using coordinates on Google Maps.
(3) X_3 = number of convenience stores within a 500-meter radius.
(4) House location that is represented in latitude (X_4) and longitude (X_5).

3.2. *Methods of Data Analysis*

This study was conducted in four main phases.

(1) Data preprocessing; involves:
   (a) Performing the OLS regression to select the significant input variables and to remove the outliers.
   (b) Fuzzification to transform house prices data from crisp to TFN. This study adopts the fuzzification strategy suggested by [31]. Using this strategy, the spread of a TFN is set as 10% of the output's value.
   (c) Data preparation for the range-preserved controlled 10-fold cross-validation process developed by [28] to address the drawback of the standard cross-validation technique when applied to fuzzy regression.

(2) Data analysis using possibilistic fuzzy regression:
   (a) Building a fuzzy regression model.
   (b) Predictive performance evaluation using 10 pairs of training-testing datasets generated in step 1(c). For the i-th training dataset and the i-th testing dataset ($i = 1, 2, \cdots, 10$), do:
      i. Construct a fuzzy model using possibilistic fuzzy regression based on all observations in the i-th training dataset.
      ii. Use the resulting model to predict a TFN of the house price for all observations in the i-th testing dataset.
      iii. Calculate MRD in predicting the TFN of observation in the i-th testing dataset.
      iv. The overall MRD is computed as the average of the MRD across all testing datasets.

(3) Data analysis using the modified Cheng and Lee KNN fuzzy regression.
   At this stage, the rules for predicting the TFN of house prices are derived, including the number of nearest neighbors (k), the Minkowski exponent parameter (p), and the weighting parameter (r). This study uses 5 levels of p ($p = 1.4, 1.5, 1.6, 1.7, 1.8$, and 2.0), 3 levels of r ($r = 0, 1, 2$), and 100 levels of k ($k = 1, 2, \cdots, 100$), resulting in 1500 variations. The optimal k, p, and r are determined using a controlled 10-fold cross-validation strategy similar to that used in step 2.
   Steps taken in the cross-validation process are:
   (a) Dataset preparation for the 10-fold range-preserved controlled cross-validation process using the same training and testing datasets as produced for the possibilistic approach.

(b) For each pair of training and testing datasets, apply the modified Cheng and Lee KNN fuzzy regression repeatedly for variations that are combinations of 5 levels of p, 3 levels of r, and 100 levels of k.

(c) For a given variation, calculate the overall MRD using the following steps:

i. Calculate the MRD for all pairs of training and testing datasets.

ii. Compute the overall MRD by averaging the MRDs obtained in step 3a.

(4) Comparison of predictive performance of all variations of the modified Cheng and Lee KNN fuzzy regression and possibilistic fuzzy regression.
The best-performing method is the one with the lowest MRD.

4. Result and Discussion

4.1. *Data preprocessing*

Data preprocessing is conducted to prepare the data for fuzzy regression analysis. In this step, OLS regression is performed to select the variables that significantly affect the house prices in that district. In addition, OLS regression is also performed to remove outliers from the dataset. This step generally can enhance the model's accuracy. Initially, the dataset comprises 288 observations and six input variables. The OLS regression found that only four variables significantly affect house prices, namely house age, the number of convenience stores within a 500-meter radius, the distance to the MRT station, and the house's latitude. Meanwhile, the house's longitude does not appear to influence house prices. The correlation analysis shows that this variable is highly correlated with the distance to the MRT station. Thus, once the distance to the MRT station is included, the longitude provides little additional information. Residual analysis identified 48 outliers that may degrade the model's predictive ability, distort parameter estimates, and lead to misleading conclusions. Removing outliers is expected to enhance the model's predictive ability. After removing insignificant variables and outliers, the dataset consists of 240 observations and four variables. This new dataset is then analyzed using the fuzzy regression approach.

4.2. *Possibilistic fuzzy regression result*

The following equation expresses the estimated fuzzy model obtained using possibilistic fuzzy regression.

$$\widehat{\widetilde{Y}} = -(5649.78, 0) - (0.227, 0)X_1 - (0.00349, 0)X_2 + (1.221, 0)X_3 + (227.898, 0.5636)X_4 \tag{4.1}$$

where X_1 is the house age, X_2 is the distance to the nearest MRT station, X_3 is the number of convenience stores in a 500-meter radius, and X_4 is the latitude of the house's geographic position.

This equation shows that house age negatively influences its price. Since this variable represents depreciation, the result implies that house prices decline as houses age, with older houses tending to be valued lower. Similarly, the distance from a house to the nearest MRT station also negatively influences house prices.

This result indicates that houses located farther from the MRT station are associated with lower prices. While this variable represents accessibility, the finding shows that access to public transportation is an important determinant of house prices in that district. In contrast, the regression coefficient for the number of convenient is positive. This variable reflects the availability of local amenities and services. Therefore, this finding indicates that buyers are willing to pay more for a house with better access to retail and services. The latitude position of a house also has a positive impact on the predicted house price. However, unlike the other variables, latitude is the only independent variable with a non-zero spread, indicating that it not only affects the mode estimate of house price but also its fuzziness. Latitude contributes to the heterogeneity and uncertainty in house pricing.

The finding that only latitude contributes to uncertainty suggests that its effect on house prices is more difficult to represent by a single crisp coefficient than the effects of the other explanatory variables. In this dataset, house age, distance to the nearest MRT station, number of convenience stores, and longitude appear to have relatively stable relationships with price, so their effects can still be adequately captured in a crisp form. By contrast, latitude may reflect broader spatial differences that are not directly measured in the model, such as neighborhood quality, prestige of location, environmental conditions, or access to certain local facilities. As a result, part of the uncertainty in house price formation may be absorbed by latitude, making a fuzzy representation more appropriate for this variable. Suppose five houses are located in that district. These houses are characterized by the values presented in Table 1. The predicted TFN for these house prices is obtained by replacing X_1, X_2, X_3, and X_4 in Eq. (4.1) with the corresponding house characteristics. The complete result is presented in columns 6-9 of Table 1.

Table 1. *Predicted TFN of House Price*

No	House age	Distance to MRT	Number of convenience	Latitude	Predicted TFN of house price			
					Mode	Spread	Lower bound	Upper bound
1	12	700	3	24.9794	41.47	14.08	27.39	55.55
2	21	1000	0	24.9389	25.48	14.06	11.43	39.54
3	5	960	2	24.9643	37.48	14.07	23.41	51.56
4	11	300	1	24.9663	37.66	14.07	23.59	51.73
5	4	400	5	24.9794	46.78	14.08	32.70	60.85

The performance of Tanaka's fuzzy regression to predict the TFN for given houses is evaluated using range-preserved controlled cross-validation. This process required information about the extreme values of all independent variables. The following table provides this information. In addition, this information is also useful in determining the predicted TFN. Once a fuzzy model is obtained, it can be used only to predict TFN for observations whose independent variable values are within this range.

In this strategy, all observations with these values form a boundary dataset. The remaining observations are then split into 10 folds, producing 10 pairs of training and testing datasets. The i-th testing dataset consists of observations in the i-th fold,

Table 2. *Extreme values of Independent Variables for House Prices*

Variables	Minimum value	Maximum value
House age (X_1)	0	43.80
Distance to MRT station	49.66	6488.02
Number of convenience stores	0	10
Latitude	24.93885	25.01545

while the i-th training dataset combines the boundary dataset and all observations except those in the i-th fold. For each pair, a model is built on the training data and used to predict the price TFN for houses in the testing dataset. The MRD is calculated for each pair. Table 3 presents the resulting models for each dataset pair and the MRD.

Table 3. *Predicted Fuzzy Model and MRD*

Pair	Model	MRD
1	$\widehat{\widetilde{Y}} = -(5463.211, 0) - (0.2247, 0.013)X_1 - (0.00390, 0)X_2 + (1.2903, 0)X_3 + (220.4217, 0.4652)X_4$	2.279
2	$\widehat{\widetilde{Y}} = -(565.984, 0) - (0.227, 0)X_1 - (0.00349, 0)X_2 + (1.221, 0)X_3 + (227.986, 0.476)X_4$	2.639
3	$\widehat{\widetilde{Y}} = -(565.984, 0) - (0.227, 0)X_1 - (0.00349, 0)X_2 + (1.221, 0)X_3 + (227.986, 0.476)X_4$	2.350
4	$\widehat{\widetilde{Y}} = -(565.984, 0) - (0.227, 0)X_1 - (0.00349, 0)X_2 + (1.221, 0)X_3 + (227.986, 0.476)X_4$	2.483
5	$\widehat{\widetilde{Y}} = -(565.984, 0) - (0.227, 0)X_1 - (0.00349, 0)X_2 + (1.221, 0)X_3 + (227.986, 0.476)X_4$	2.80
6	$\widehat{\widetilde{Y}} = -(5995.304, 11.7004) - (0.2561, 8.9 \times 10^{-6})X_1 - (0.00360, 0)X_2 + (1.2208, 0)X_3 + (227.9863, 0.4763)X_4$	2.486
7	$\widehat{\widetilde{Y}} = -(5848.536, 11.5715) - (0.2441, 0)X_1 - (0.00389, 0)X_2 + (1.1230, 0.0269)X_3 + (235.8968, 0)X_4$	2.734
8	$\widehat{\widetilde{Y}} = -(4624.97, 0) - (0.233, 0)X_1 - (0.00362, 0)X_2 + (1.111, 0)X_3 + (186.8704, 0.4759)X_4$	2.400
9	$\widehat{\widetilde{Y}} = -(5719.99, 0) - (0.227, 0)X_1 - (0.00346, 0)X_2 + (1.216, 0)X_3 + (230.7094, 0.4761)X_4$	2.388
10	$\widehat{\widetilde{Y}} = -(565.984, 0) - (0.227, 0)X_1 - (0.00349, 0)X_2 + (1.221, 0)X_3 + (227.986, 0.476)X_4$	2.508

Note: X_1 = house age, X_2 = distance to the nearest MRT station, X_3 = number of convenience stores within a 500-meter radius, X_4 = latitude of the house's geographic position.

The overall MRD is the average across all iterations, which is 2.524678.

4.3. *The modified Cheng and Lee KNN fuzzy regression result*

In contrast to possibilistic fuzzy regression, which uses a mathematical equation, this method identifies the best rule for predicting the TFN of house prices in that district. This rule includes the number of nearest neighbors (k), the Minkowski distance exponent (p), and the weight parameter (r). The best rule provides the lowest MRD. Figures 1 – Figure 3 display charts illustrating the change in MRD as the number of nearest neighbors increases, with k ranging from 1 to 100. Each graph corresponds to a different value of the weight parameter r (0, 1, and 3). The results are depicted in line charts arranged in seven panels, each for a specific value of the distance exponent parameter (1.4 to 2). MRDs are obtained by averaging

across 10 iterations in the cross-validation procedure.

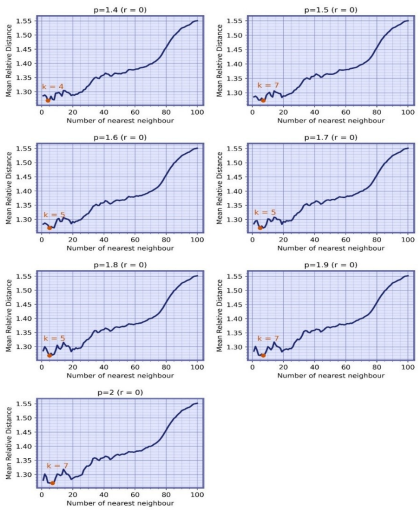

Figure. 1. *Relationship Between k and MRD for Various p* ($r = 0$)

In particular, Figure 1 shows the result for $r = 0$. The cases with $r = 0$ correspond to the equal weighting scheme; in other words, all selected nearest neighbors contribute equally to predict the mode and spread of TFN. This figure shows that the lowest MRD is achieved when the modified Cheng and Lee k-nearest neighbor is applied to data with 4 nearest neighbors for $p = 1.4$; 7 nearest neighbors for $p = 1.5$, $p = 1.9$, and $p = 2$; and 5 nearest neighbors for $p = 1.6$, $p = 1.7$, and $p = 1.8$. The MRD for all mentioned conditions is presented in Table 4.

Table 4. *MRD Across Values of Distance Exponent Parameter* ($r = 0$)

Minkowski exponent (p)	**Number of nearest neighbors** (k)	**MRD**
1.4	4	1.26885
1.5	7	1.27303
1.6	5	1.26955
1.7	5	1.26996
1.8	5	1.26808
1.9	7	1.26967
2.0	7	1.26971

The findings in Table 4 show that the MRD values are consistent across all p values studied. It varies within a narrow interval (1.26808, 1.27303). It indicates that, under equal weighted averaging, the performance of the M-CL-KNNFR in predicting the house TFN is not strongly sensitive to changes in p, once the optimum k is determined. However, the best performance is achieved at $p = 1.8$ with a MRD

of 1.26808. Figure 2 illustrates how the MRD changes as k increases for $r = 1$. It should be noted that in the case of $r \neq 0$, the selected nearest neighbors contribute differently in predicting the mode and the spread of TFN, with weights inversely proportional to the distance.

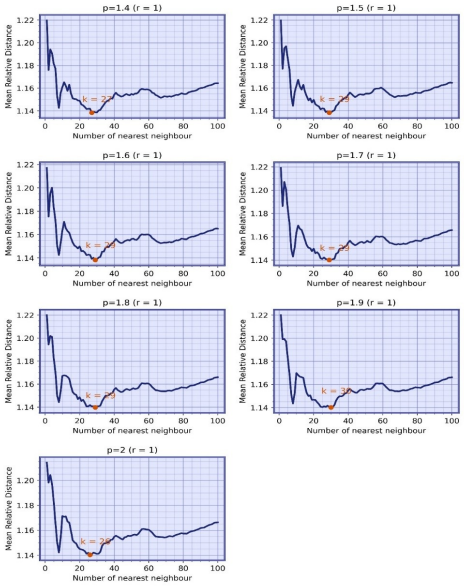

Figure. 2. *Relationship Between k and MRD for Various p (r = 1)*

This figure shows that the lowest MRD is achieved when the modified Cheng and Lee k-nearest neighbor is applied to data with 27 neighbors for $p = 1.4$; 29 neighbors for $p = 1.5$, $p = 1.6$, $p = 1.7$, $p = 1.8$; 30 neighbors for $p = 1.9$; and 26 neighbors for $p = 2$. Compared with the optimal k values found for $r = 0$, the optimal k values found for $r = 1$ are substantially larger. The MRD for the optimal combination of (p, k) selected in this step is summarized in Table 5.

Table 5. *MRD Across Values of Distance Exponent Parameter (r = 1)*

Minkowski exponent (p)	**Number of nearest neighbors (k)**	**MRD**
1.4	27	1.13845
1.5	29	1.13826
1.6	29	1.13819
1.7	29	1.14004
1.8	29	1.13973
1.9	30	1.14010
2.0	26	1.14038

As previously found in case $r = 0$, this table also indicates that the MRD is highly consistent, ranging from 1.13819 to 1.14038. It indicates that, once the k value is optimized, predictive performance is relatively robust to variation in the Minkowski distance exponent. The best predictive performance of the modified Cheng and Lee k-nearest neighbor regression is achieved at $p = 1.6$ and $k = 29$ with an MRD of 1.13819. The result for $r = 2$ is depicted in Figure 3.

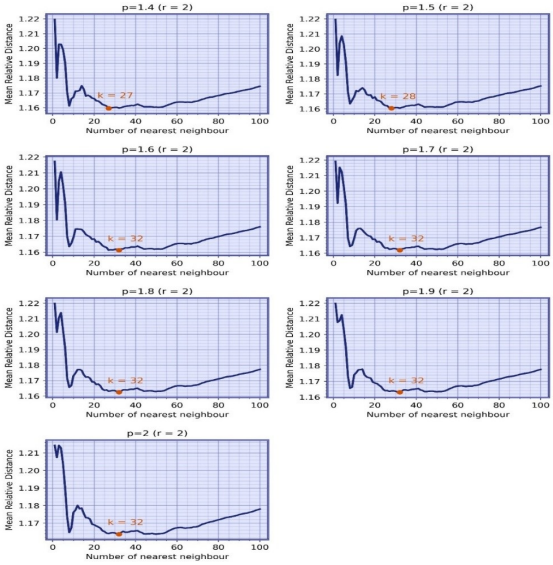

Figure. 3. *Relationship Between k and MRD for Various p (r = 2)*

Figure 3 shows that the best performance is achieved when the modified Cheng and Lee k-nearest neighbor is applied to data with 27 nearest neighbors for $p = 1.4$, 28 nearest neighbors for $p = 1.5$, and 32 nearest neighbors for $p = 1.6$ to $p = 2$. Almost similar to the previous result, the number of nearest neighbors in this case ranges from 26 to 32. The MRD for this optimal combination of (p,k) selected in this step is summarized in Table 6.

Table 6. *MRD Across Values of Distance Exponent Parameter (r = 2)*

Minkowski exponent (p)	**Number of nearest neighbors (k)**	**MRD**
1.4	27	1.15962
1.5	28	1.16046
1.6	32	1.16125
1.7	32	1.16187
1.8	32	1.16244
1.9	32	1.16306
2.0	32	1.16368

The MRD values in Table 6 range from 1.15962 to 1.16368, indicating stability across values of p. However, compared to $r = 1$, the MRD yielded is consistently larger. It implies that increasing the weighting parameter from $r = 1$ to $r = 2$ degrades the performance of the modified Cheng and Lee KNN fuzzy regression in predicting the house price TFN. This table also shows that the best performance is achieved at $p = 1.4$ and $k = 27$, with the lowest MRD of 1.15962.

4.4. *Comparative analysis*

Table 7 summarizes the previous result on the optimal combination of r, p, and k that yields the best performance of the M-CL-KNNFR when implemented on the house price dataset. It is found that the best predictive performance is achieved at $r = 1$, $p = 1.6$, and $k = 29$.

Table 7. *Summary of Optimal Combination of r, p, and k*

Weighting Parameter (r)	**Minkowski exponent (p)**	**Number of nearest neighbors (k)**	**MRD**
0	1.8	5	1.26808
1	1.6	29	1.13819
2	1.4	27	1.15962

Moreover, compared to possibilistic fuzzy regression, the modified Cheng and Lee approach shows a better ability to predict a TFN for given house prices. The possibilistic fuzzy regression yields an MRD 2.2 times higher than the modified Cheng and Lee approach.

5. Conclusion

This study aimed to generate rules for the modified Cheng and Lee KNN fuzzy regression to predict the TFN of a given house's price. The rules include the best distance measure used, the best weighting scheme, and the optimal number of nearest neighbors used as the basis of prediction. Subsequently, the best-performing variant of modified Cheng and Lee KNN fuzzy regression was then compared to possibilistic fuzzy regression based on MRD. The result from the possibilistic fuzzy regression shows that house age and the distance to the nearest MRT station negatively impact the house prices, while the number of convenience stores and latitude positively impact the house prices. Latitude is the only variable that contributes to the heterogeneity and uncertainty in house pricing. A possible explanation is that latitude may reflect broader spatial differences that are not directly observed in the available variables.

Results from M-CL-KNNFR show that the best-performing approach of the modified Cheng and Lee KNN fuzzy regression is achieved when this approach employs the Minkowski distance with the exponent parameter $p = 1.6$, an unequal weighting scheme with $r = 1$, and 29 nearest neighbors. Compared to possibilistic fuzzy regression, this approach gives better performance in predicting the TFN of house prices. In addition, the possibilistic fuzzy regression yields an MRD 2.2 times

higher than the modified Cheng and Lee approach. Most existing studies treat housing prices as single-point values, implicitly assuming that market prices can be determined with precision despite negotiation processes and market uncertainty. This study addresses that gap by modeling housing prices as fuzzy values, allowing the estimation framework to reflect not only the relationship between variables but also the inherent uncertainty in real-world price formation.

6. Acknowledgment

This article originated from the first author's doctoral dissertation submitted to Universiti Pendidikan Sultan Idris. The author gratefully acknowledges the financial support provided by Universitas Andalas throughout this study.

Bibliography

[1] Chen, C., Ma, X., Zhang, X., 2024, Empirical Study on Real Estate Mass Appraisal Based on Dynamic Neural Networks, *Buildings*, Vol. **14**: 2199

[2] Erciyes, A.H., 2025, Learning The Value Of Place: Machine Learning Models for Mass Valuation, *Buildings*, Vol. **15**: 2773

[3] Jafary, P., Shojaei, D., Rajabifard, A., Ngo, T.D., 2024, Automating Property Valuation at The Macro Scale of Suburban Level: A Multi-Step Method Based on Spatial Imputation Techniques, *Habitat International*, Vol. **148**: 103075

[4] Karanikolas, N., 2025, Artificial Intelligence and Real Estate Valuation, *Information*, Vol. **16**: 1049

[5] Ota, A., 2024, Variation in Property Valuations Conducted by Artificial Intelligence Services, *Equilibrium Economics*, Vol. **1**: 13

[6] Jiang, E.X., Zhang, A.L., 2025, Collateral Value Uncertainty and Mortgage Credit Provision, *Journal of Financial Economics*, Vol. **169**: 104054

[7] El Jaouhari, A., Samadhiya, A., Kumar, A., Šešplaukis, A., Raslanas, S., 2024, Mapping The Landscape: A Systematic Literature Review on Automated Valuation Models and Strategic Applications in Real Estate, *International Journal of Strategic Property Management*, Vol. **28**: 286–301

[8] Jaroszewicz, J., Horynek, H., 2024, Aggregated Housing Price Predictions with No Information About Structural Attributes—Hedonic Models: Linear Regression and a Machine Learning Approach, *Land*, Vol. **13**: 1881

[9] Nor, M.I., Hussein, S.N., 2025, Modeling Residential Property Prices in Emerging Climate-Responsive Urban Markets: A Hybrid Modeling Framework for Baidoa City-Somalia, *Frontiers in Built Environment*, Vol. **11**: 1615229

[10] Rey-Blanco, D., Zofío, J.L., González-Arias, J., 2024, Improving Hedonic Housing Price Models by Integrating Optimal Accessibility Indices Into Regression and Random Forest Analyses, *Expert Systems with Applications*, Vol. **235**: 121059

[11] Wan, H., Roy Chowdhury, P.K., Yoon, J., Bhaduri, P., Srikrishnan, V., Judi, D., Daniel, B., 2025, Explaining Drivers of Housing Prices with Nonlinear Hedonic Regressions, *Machine Learning with Applications*, Vol. **21**: 100707

[12] Wang, Z., Wang, Y., Xia, X., Chen, S., Jiang, W., 2025, How Does Built Environment Influence Housing Prices in Large-Scale Areas? An Interpretable Machine Learning Method by Considering Multi-Dimensional Accessibility, *ISPRS International Journal of Geo-Information*, Vol. **14**: 436

[13] Dubois, D., Prade, H., 1980, Fuzzy Sets and Systems: Theory and Applications, Academic Press, New York
[14] Zadeh, L.A., 1965, Fuzzy Sets, *Information and Control*, Vol. **8**: 338–353
[15] Chaira, T., 2019, Fuzzy Set and Its Extension: The Intuitionistic Fuzzy Set, Wiley
[16] Zimmermann, H.J., 2001, Fuzzy Set Theory and Its Application, 4th, Springer, New York
[17] Bector, C.R., Chandra, S., 2005, Fuzzy Mathematical Programming and Fuzzy Matrix Games, Springer
[18] Tanaka, H., Uejima, S., Asai, K., 1982, Linear Regression Analysis with Fuzzy Model, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. **12**: 903–907
[19] Tanaka, H., 1987, Fuzzy Data Analysis by Possibilistic Linear Models, *Fuzzy Sets and Systems*, Vol. **24**: 363–375
[20] Tanaka, H., Watada, J., 1988, Possibilistic Linear Systems and Their Application to the Linear Regression Model, *Fuzzy Sets and Systems*, Vol. **27**: 275–289
[21] Sakawa, M., Yano, H., 1992, Multiobjective Fuzzy Linear Regression Analysis for Fuzzy Input–Output Data, *Fuzzy Sets and Systems*, Vol. **47**: 173–181
[22] Hojati, M., Bector, C.R., Smimou, K., 2005, A Simple Method for Computation of Fuzzy Linear Regression, *European Journal of Operational Research*, Vol. **166**: 172–184
[23] Kohli, S., Godwin, G.T., Urolagin, S., 2021, Sales Prediction Using Linear and KNN Regression, in Advances in Machine Learning and Computational Intelligence, Springer, Singapore
[24] Song, Y., Liang, J., Lu, J., Zhao, X., 2017, An Efficient Instance Selection Algorithm for k Nearest Neighbor Regression, *Neurocomputing*, Vol. **251**: 26–34
[25] Kumbure, M.M., Luukka, P., Collan, M., 2020, A New Fuzzy k-Nearest Neighbor Classifier Based on the Bonferroni Mean, *Pattern Recognition Letters*, Vol. **140**: 172–178
[26] Fathabadi, A., Seyedian, S.M., Malekian, A., 2022, Comparison Bayesian, k-Nearest Neighbor, and Gaussian Process Regression Methods for Quantifying Uncertainty of Suspended Sediment Concentration Prediction, *Science of The Total Environment*, Vol. **818**: 151760
[27] Cheng, C., Lee, E., 1999, Nonparametric Fuzzy Regression—K-NN and Kernel Smoothing Techniques, *Computers & Mathematics with Applications*, Vol. **38**: 239–251
[28] Yozza, H., 2026, *The Integrated KNN Regression in Improvement of Fuzzy Regression Performance*, Doctoral dissertation at Universiti Pendidikan Sultan Idris, unpublished
[29] Kim, B., Bishu, R.R., 1998, Evaluation of Fuzzy Linear Regression Models by Comparing Membership Functions, *Fuzzy Sets and Systems*, Vol. **100**: 343–352
[30] Yeh, I., 2018, Real Estate Valuation [Dataset], UCI Machine Learning Repository
[31] Rahman, H.M., Arbaiy, N., Wen, C.C., Efendi, R., 2019, Autoregressive Modeling with Error Percentage Spread based Triangular Fuzzy Number, *International Journal of Recent Technology and Engineering*, Vol. **8**: 252