# Estimating Relationships in Multi-Dimensional Data Sets by Means of Asymmetric Fuzzy Regression

Raphael A. Krauthann, Tobias Kruse, Hinnerk Jannis Müller, Michael Stumpf, and Peter Rausch
*Department of Computer Science*
*Nuremberg Institute of Technology Georg Simon Ohm*
Keßlerplatz 12, 90489 Nuremberg, Germany
{krauthannra64754|kruseto64083|muellerhi65413|michael.stumpf|peter.rausch}@th-nuernberg.de

**KEYWORDS**

Fuzzy Regression Analysis, Modeling, Machine Learning, Data Mining, Predictive Analytics, Outlier Detection, Asymmetric Fuzzy Regression

**ABSTRACT**

In spite of all progress of AI and Machine Learning, making predictions based on real-world data is still a challenging task. For this purpose, Tanaka's approach of symmetric fuzzy linear regression is explained, and open issues are outlined. These issues occur if the instances of data sets are not symmetrically distributed. For this purpose, new solutions based on enhancements of Tanaka's approach are discussed. A real-world scenario for predicting house prices is used to illustrate the ideas. It is shown that the new asymmetric approach works at least as well as the symmetric version but is superior in certain situations.[1]

## I. INTRODUCTION

Although a lot of research has been done in the field of AI and Machine Learning, making predictions based on real-world data is still a challenging task. Thus, a new approach is presented to estimate relationships in multi-dimensional data sets by applying an asymmetrical fuzzy regression technique. Regression approaches are prevalent and applied in numerous areas, for instance, in the fields of financial risk measurement (Valaskova et al. 2018), sales revenue prediction in the telecommunications industry (Welc and Esquerdo 2018), or stock price prediction (Patel, Patel, and Darji 2018). Nevertheless, in many cases, a point forecast is not always appropriate (Rausch and Jehle 2013). For instance, if the impact of weather parameters on sales of ice cream is analyzed, no clear causal relationship between independent and dependent variables is apparent, but a fuzzy linear correlation can be observed (Rausch and Jehle 2013). To solve this type of real-world issues, fuzzy versions of regression approaches were developed, see for instance (Tanaka, Uejima, and Asai 1982). Approaches providing lower and upper boundaries for fuzzy linear relationships are already available. However, this paper shows that these approaches are not appropriate for all types of data sets, for instance, if symmetric lower and upper boundaries are an inadequate representation of a fuzzy relationship. Thus, in the following sections, a new solution to overcome this issue is presented. Section II describes a real estate data set which will be used for illustration purposes. Section III provides a brief survey of available solutions and their limitations. In Section IV, details of Tanakas' popular approaches are provided, and enhancements to handle issues in asymmetrically distributed data sets are introduced as a new approach. To illustrate its features in Section V, it is applied to the data set, and the impact of predictors on sales prices of houses is analyzed. In Section VI, the findings are evaluated. It is worked out in which situations the associated features are beneficial. Note that a comparison to other types of AI techniques is not within the scope of this paper and should be handled in a follow-up study. Finally, plans for future research into further enhancements and other fields of application are presented.

## II. DATA SET

The data set used in the following sections to illustrate the new approach represents sales of individual residential property in Ames, Iowa, from 2006 to 2010. It contains 2930 transactions and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) that play a role in the valuation of property values (Cock 2011). This data set is the basis of the Kaggle Competition "House Prices: Advanced Regression Techniques". The contest's goal is to predict the sales price for each house. In order to discuss the features of the new fuzzy method presented in this paper, only a few attributes of the data set are used. The temporal attributes considered are the year of construction, the year of renovation, and the month and year of sale. Furthermore, the data set contains information on lot size, the quality of the material used, and its condition, which are also considered.

## III. RELATED WORK

The data set described has been used in several publications, such as (Fan, Cui, and Zhong 2018; Vik-

[1] The algorithms in this paper were implemented in Python 3 and are freely accessible at (`https://github.com/TheGarkine/fuzz_regression_py`).

Communications of the ECMS, Volume 34, Issue 1,
Proceedings, ©ECMS Mike Steglich, Christian Mueller,
Gaby Neumann, Mathias Walther (Editors)
ISBN: 978-3-937436-68-5/978-3-937436-69-2(CD) ISSN 2522-2414

torovich et al. 2018; Yan 2017). The data set analyses followed a similar process:

First, the data set was preprocessed, removing outliers and handling missing values, as well as removing unwanted features and possibly adding custom features. Subsequently, several base methods and one or more ensemble methods were selected. Fan et al. used Lasso and Ridge linear regression models, which introduce penalty terms. Besides, random forest, support vector regression with linear and Gaussian kernel, and extreme gradient boosting trees as base methods are applied (Fan, Cui, and Zhong 2018). As ensemble strategy, they used a simple weighted linear combination of the best performing models. Viktorovich et al. used Lasso linear regression, ElasticNet, extreme gradient boosting trees and neural networks as base methods. Different combinations were assessed as integrating ensemble methods (Viktorovich et al. 2018). Whereas Yan used gradient boosting trees, random forest, and regularized regression with both Lasso and Ridge penalization as base methods (Yan 2017). On top of that, different ensemble methods were applied, whereby a multi-class ROC random forest was proposed.

After selecting appropriate base and ensemble methods, a hyperparameter optimization took place, where the best performing combination of the selected methods was determined. Since the output of the base models is the input for the ensemble methods, this is a multi-layered process or stacked generalization as described in (Wolpert 1992). Decent results are achievable with this practice. However, in many cases, the results of ensemble methods are difficult to understand and are often treated as opaque "black-boxes" (Cortez and Embrechts 2011). There are approaches and ongoing research summarized under the term Explainable Artificial Intelligence (Arrieta et al. 2020), which try to make "black-box"-models more understandable, for instance, with sensitivity analyses (Cortez and Embrechts 2011), by extracting rules (Tickle et al. 1998) or by visualizing different layers (Yosinski et al. 2015). To avoid the issue mentioned above, the focus in the following sections is on fuzzy linear regression.

While conducting the literature review, it became clear that developments on the research topic of fuzzy regression have not received much attention in the last two decades. Thus, there is not much preliminary work on this research subject. However, a few sources could be identified: D'Urso and Gastaldi presented an asymmetric fuzzy regression approach for crisp data sets with a fuzzy output component using a new metric (D'Urso and Gastaldi 2001). In (Neubauer 2010), much knowledge about fuzzy regression and further developments can be found. The methods discussed can handle fuzzy input, which is not the focus of this paper.

One of the first publications in this field is from Tanaka et al., who also tried to model house prices (Tanaka, Uejima, and Asai 1982). They used available crisp data in five dimensions to estimate prices for Japanese prefabricated houses using an algorithm with a linear optimization problem.

Diamond also investigated fuzzy regression in 1988 (Diamond 1988). He solved the problem of fuzzy-to-fuzzy regression by implementing a new metric and proving its properties. Additionally, his ideas of crisp-to-fuzzy regressions were presented. His regression resulted in symmetric triangular fuzzy numbers.

In 1997, Tanaka and Lee published a fuzzy regression approach to process crisp data sets (Tanaka and Lee 1997). Their work is the basis for the approaches presented in this paper. Therefore, their ideas are outlined in Section IV. Additionally, Lee and Tanaka presented a non-symmetric fuzzy regression idea with crisp input and output, in 1999 (Lee and Tanaka 1999). However, the approach presented has the significant drawback that it weights the fuzzy properties to the linear correlation unequally. This feature is explained in Section VI.

## IV. APPROACH

This section presents the general approaches, starting with a definition of fuzzy linear functions, their properties, and other fundamentals. Afterward, Tanaka's approaches are reviewed. Note that many regression algorithms are often denoted using matrices and vectors (see (Neubauer 2010) or (Tanaka, Uejima, and Asai 1982)), while this paper relies on sums over real values, iterating through input dimensions. This is done to make the general idea of the algorithms and the underlying quadratic optimization problem more understandable. Based on this, a new asymmetric approach is introduced. Tanaka's so-called optimization without expert knowledge (Tanaka and Lee 1997) is also integrated into this approach.

### A. *Fundamentals*

#### A.1 *Data Set Notation and Definition*

The data set shall be defined as $D$ containing $p$ instances. Each instance of $D$ has $n+1$ dimensions or attributes. For the sake of this paper, $x_j$ means the $j$-th data point with all its attributes. Therefore, the notation $x_{ji}$ represents the value of the $i$-th dimension of the $j$-th instance within $D$. The algorithms expect numerical, non-negative values ($\mathbb{R}_0^+$). Concluding, the data set can is defined as

$$D \subset {\mathbb{R}_0^+}^{n+1}, |D| = p \tag{1}$$

#### A.2 *Fuzzy Linear Regression*

The goal is to find a fuzzy linear function $f : \mathbb{R}^n \to \widetilde{\mathbb{R}}$. These functions can be generally described as:

$$Y = \sum_{i=0}^{n} \gamma_i X_i \tag{2}$$

where

$$X_0 = 1 \tag{3}$$

$$X_i \in \mathbb{R}, i = 1, \ldots, n+1 \tag{4}$$

$$\gamma_i \in \widetilde{\mathbb{R}}, i = 0, \ldots, n \tag{5}$$

In Equation (2), $n$ denotes the number of input dimensions in the data analyzed. The $X_i$ represents the crisp $i$-th dimension of the data set, while $\gamma_i$ values

are fuzzy coefficients calculated by the linear regression algorithm for the $i$-th dimension. For algorithmic purposes, $X_0$ is initialized with the value of *one* so that the sum includes the $\gamma_0$ case. In accordance with the data set notation, $X_{n+1}$ denotes the dimension analyzed that is to be approximated.

In the following sections, the performance of three algorithms inspired by Tanaka's approach (Tanaka and Lee 1997) is compared. Additionally, these solutions are transferred to an asymmetrical algorithm to solve the issues mentioned in Section I. The goal is to compare the approaches in both the performance and relevance of their results. Each algorithm gives a fuzzy linear approximation, including a centerline and lower and upper boundaries for expected values. These boundaries may also be referenced as a tunnel. This concept is typically not implemented by other algorithms, such as neural networks.

Asymmetric approaches could be superior in representing real-world scenarios (described in Subsection IV-D). This work tests this hypothesis. All of the procedures presented provide a linear function with triangular fuzzy numbers, which are either symmetric or asymmetric. Symmetrical triangular fuzzy numbers, which can be used to represent the $\gamma_i$ parameters are, denoted as $(a; c)_S$, where $a$ is the center of the number and $c$ the symmetric spread. Asymmetrical triangular fuzzy numbers are denoted as $(a; l; u)_{LR}$, where $a$ is the center of the number, $l$, and $u$ describes the lower and upper boundary distances.

### A.3 *h-Level*

Some fuzzy algorithms have an $h$-level parameter (Tanaka and Lee 1997). The $h$-level defines the minimum membership for each data point of the data set with the resulting linear regression. The closer the $h$-level comes to 1, the wider the tunnel becomes. Thus, the fuzziness of the resulting regression increases.

## B. Linear Regression with Symmetric Triangular Fuzzy Numbers, Ignoring Linear Correlations

Using Equation (2) the linear goal function for symmetric triangular fuzzy numbers can generally be described as follows:

$$Y = \sum_{i=0}^{n} (a_i; c_i)_S X_i \tag{6}$$

For the first version of the algorithm, the linear correlation within the data set is ignored. Instead, the spread of the resulting symmetric triangular fuzzy coefficients is reduced, so that all values are within the $h$-level of the fuzzy function. Therefore, the following linear programming problem has to be solved.

$$\min_{(a_0,c_0),\ldots,(a_n,c_n)} \sum_{i=0}^{n} \sum_{j=1}^{p} c_i x_{ji} \tag{7}$$

subject to:

$$\sum_{i=0}^{n} \Big( (a_i + (1-h)c_i) x_{ji} \Big) \geq x_{j(n+1)},\ j=1,\ldots,p \tag{8}$$

$$\sum_{i=0}^{n} \Big( (a_i - (1-h)c_i) x_{ji} \Big) \leq x_{j(n+1)},\ j=1,\ldots,p \tag{9}$$

$$c_i \geq 0, \qquad i=0,\ldots,n \tag{10}$$

The constraints defined in (8) and (9) cause the regression to have all values within the spread with a minimum membership of $h$. Additionally, it is desirable to prevent spreads to become less fuzzy for higher numbers, which is achieved by (10) (Tanaka and Lee 1997).

## C. Linear Regression with Symmetric Triangular Fuzzy Numbers, Using Linear Correlations

Ignoring the data set's linear correlation between dimensions can result in errors in the interpretation of the resulting linear regression. Hence, this property should not be ignored. Tanaka provides a solution combining the properties of least-squares regression with the fuzzy effect of Section IV-B (Tanaka and Lee 1997). Again, the goal is to create a function similar to Equation (6).

The resulting quadratic optimization problem can now be written as:

$$\begin{aligned} \min_{(a_0,c_0),\ldots,(a_n,c_n)} & k_1 \sum_{j=1}^{p} \left( x_{j(n+1)} - \sum_{i=0}^{n} a_i x_{ji} \right)^2 \\ & + k_2 \sum_{j=1}^{p} \left( \sum_{i=0}^{n} c_i x_{ji} \right)^2 \end{aligned} \tag{11}$$

subject to (8), (9) and (10).

Since the same constraints are still valid, those from the previous approach are reused. The factors $k_1$ and $k_2$ are user-specified positive real values and enable adjustments of the function. The ratio between $k_1$ and $k_2$ influences the weight of the linear property in contrast to the slimness of the resulting tunnel. Therefore, $k_1$ and $k_2$ can be seen as the importance of the linear property and the crispness, respectively.

## D. Linear Regression with Asymmetric Triangular Fuzzy Numbers, Using Linear Correlations

In the asymmetric case, the definition of the linear asymmetric fuzzy numbers is taken from Subsection IV-A and transformed into the following general function:

$$Y = \sum_{i=0}^{n} (a_i; l_i; u_i)_{LR} X_i \tag{12}$$

Thus, a modified version of Tanaka's symmetric fuzzy regression algorithm (Tanaka and Lee 1997) is used. Asymmetric approaches have the advantage that they are not so much affected by outliers lying only above or beneath the centerline. In this case, an extreme point only affects the side it occurs on (above or below the actual linear relationship). For the optimization, the

squares of both boundaries are combined. Since the fuzziness has now potentially twice the importance in the optimization, the combination of the two parts is halved. Finally, this result is combined with the least square regression component, and both are weighted with the two crisp factors, $k_1$ and $k_2$.

$$\min_{(a_0,l_0,u_0),\ldots,(a_n,l_n,u_n)} k_1 \sum_{j=1}^{p} \left( x_{j(n+1)} - \sum_{i=0}^{n} a_i x_{ji} \right)^2 + \frac{k_2}{2} \sum_{j=1}^{p} \left( \left( \sum_{i=0}^{n} l_i x_{ji} \right)^2 + \left( \sum_{i=0}^{n} u_i x_{ji} \right)^2 \right) \quad (13)$$

$$\sum_{i=0}^{n} \left( (a_i + (1-h) u_i) x_{ji} \right) \geq x_{j(n+1)}, \; j = 1, \ldots, p \quad (14)$$

$$\sum_{i=0}^{n} \left( (a_i - (1-h) l_i) x_{ji} \right) \leq x_{j(n+1)}, \; j = 1, \ldots, p \quad (15)$$

$$u_i \geq 0, l_i \geq 0, \qquad i = 0, \ldots, n \quad (16)$$

As in the symmetric case, the lower boundary needs to be less then or equal to, and the upper boundary greater than or equal to all recorded values. Therefore, Equations (8), (9), and (10) are modified to use the new lower ($l$) and upper ($u$) boundaries.

Figure 1 shows the asymmetric approach applied to the test values from Tanaka's work (Tanaka and Lee 1997). The boundaries of the asymmetric fuzzy regression are closer to the values of the data set compared to Tanaka's symmetrical fuzzy regression approach, since each side is only limited to their respective extrema. Note that this is limited to the space of the training set and is not valid for predictions outside of this area.

In addition, Figure 1 shows the impact on regression. $k_1 = 1$ and $k_2 = 10$ are set for the symmetrical and asymmetrical approaches, meaning that the slimness of the tunnel is more important than the least-squares regression of the centerline. Note that the upper and lower boundaries are not drawn for the symmetrical case, since they are very similar to the asymmetrical case with these parameters. The centerline of the asymmetric case (plotted solid) better matches the actual linear relationship. This effect can be measured using the root mean squared error, which is measured between the actual data points and the centerline. The symmetrical case scores 1.88, and the asymmetrical reduces this by approximately 8% to 1.73 for the given data set.

### E. Optimization Using Tanaka's Reliable and Suspicious Sets

In the symmetric cases of Subsection IV-C and Section IV-D, filtering untypical values is also possible. For this purpose, Tanaka's definitions of *suspicious* and *reliable* values are used (Tanaka and Lee 1997). First, a crisp least-squares regression is computed, and the resulting coefficients are defined as $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_n$. Subsequently, the standard deviation $\sigma$ based on the resulting regression is calculated. A point of the data set is defined *reliable* when it is within the interval:

$$[\alpha x_j - t\sigma; \alpha x_j + t\sigma], j == 1, \ldots, p \quad (17)$$

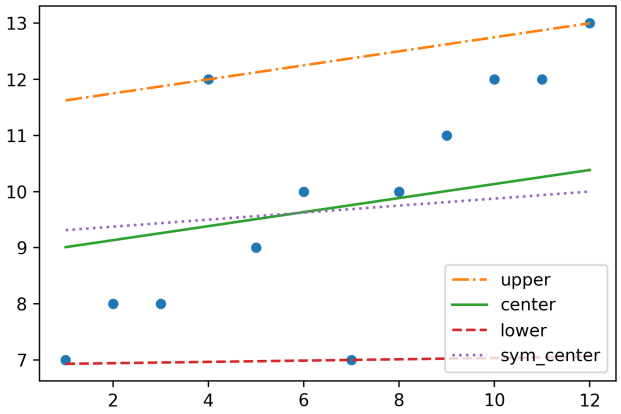


Fig. 1: Visualization of an Asymmetric Triangular Fuzzy Regression and Comparison to the Symmetrical Center Line

$t$ is a factor used to widen or tighten the set of *reliable* values ($Re$). Values outside of this interval are members of the set of *suspicious* points ($Su$).

The fuzzy objective function Equation (2) can now be extended by the fuzzy error parameter $E$, which is of type $(0; e)_S$.

$$Y = \sum_{i=0}^{n} \gamma_i X_i + E \quad (18)$$

This error $E$ should be minimal. Hence, it is considered within the optimization process as shown in the new goal (19):

$$\min_{(a_0,c_0),\ldots,(a_n,c_n),e} k_1 \sum_{j=1}^{p} \left( x_{j(n+1)} - \sum_{i=0}^{n} a_i x_{ji} \right)^2 + k_2 \sum_{j=1}^{p} \left( \sum_{i=0}^{n} c_i x_{ji} \right)^2 + k_3 e^2 \quad (19)$$

subject to Equation (10) and

$$\sum_{i=0}^{n} \left( (a_i + (1-h) c_i) x_{ji} \right) \geq x_{j(n+1)} \quad \forall x_j \in Re \quad (20)$$

$$\sum_{i=0}^{n} \left( (a_i - (1-h) c_i) x_{ji} \right) \leq x_{j(n+1)} \quad \forall x_j \in Re \quad (21)$$

$$\sum_{i=0}^{n} \left( (a_i + (1-h) c_i) x_{ji} \right) + e \geq x_{j(n+1)} \quad \forall x_j \in Su \quad (22)$$

$$\sum_{i=0}^{n} \left( (a_i - (1-h) c_i) x_{ji} \right) - e \leq x_{j(n+1)} \quad \forall x_j \in Su \quad (23)$$

Note that (22) and (23) are relaxed versions of the previously known constraints since they consider $e$. These constraints create two *tunnels*. The inner one requires all *reliable* values to be within the boundaries, while all *suspicious* values can be further away within the outer boundaries.

### F. Integrating the Optimization Technique into the Asymmetric Approach

Finally, Tanaka's method of reducing the impact of outliers can be applied to the asymmetric adaptation. The general regression function is equal to Equation (18). All coefficients $\gamma_i$ are LR fuzzy numbers.

Note that the fuzzy error parameter $E$ is now asymmetric as well and can be represented by $(0; e_l; e_u)_{LR}$. This results in the following quadratic optimization problem:

$$\min_{(a_0,l_0,u_0),\ldots,(a_n,l_n,u_n),e_l,e_u} k_1 \sum_{j=1}^{p} \left( x_{j(n+1)} - \sum_{i=0}^{n} a_i x_{ji} \right)^2 + \frac{k_2}{2} \sum_{j=1}^{p} \left( \left( \sum_{i=0}^{n} l_i x_{ji} \right)^2 + \left( \sum_{i=0}^{n} u_i x_{ji} \right)^2 \right) + k_3 (e_l^2 + e_u^2) \quad (24)$$

subject to Equation (16) and

$$\sum_{i=0}^{n} \Big( (a_i + (1-h) u_i) x_{ji} \Big) \geq x_{j(n+1)} \quad \forall x_j \in Re \quad (25)$$

$$\sum_{i=0}^{n} \Big( (a_i - (1-h) l_i) x_{ji} \Big) \leq x_{j(n+1)} \quad \forall x_j \in Re \quad (26)$$

$$\sum_{i=0}^{n} \Big( (a_i + (1-h) u_i) x_{ji} \Big) + e_u \geq x_{j(n+1)} \quad \forall x_j \in Su \quad (27)$$

$$\sum_{i=0}^{n} \Big( (a_i - (1-h) l_i) x_{ji} \Big) - e_l \leq x_{j(n+1)} \quad \forall x_j \in Su \quad (28)$$

The calculation of the reliable and suspicious sets $Re$ and $Su$, according to Tanaka's approach, is described in Section IV-E.

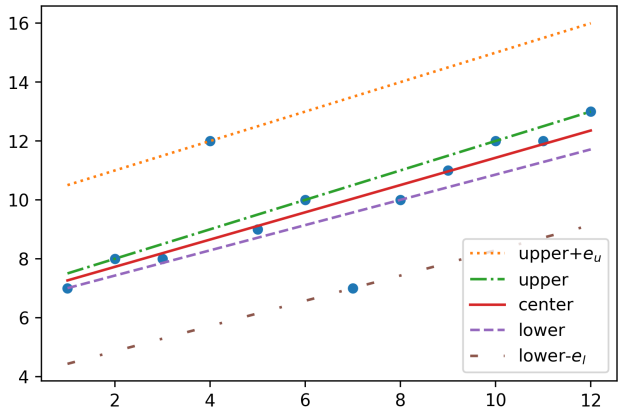


Fig. 2: Asymmetric Triangular Fuzzy Regression with Outlier Detection

In Figure 2, an asymmetric fuzzy regression with outlier elimination is visualized. The algorithm was executed with the parameters $k_1 = 1$, $k_2 = 10$, $k_3 = 1$ and $t = 2$. Obviously, the tunnel becomes significantly thinner to match the data set more closely than before since it is allowed to ignore the outer two data points. The calculated coefficients for the approaches with outlier detection (OD) and without ($\overline{\text{OD}}$) and – if applicable – error values are described in Table I.

TABLE I: Coefficients and Error Values

| Alg. | $e_l$ | $l_0$ | $l_1$ | $a_0$ | $a_1$ | $u_0$ | $u_1$ | $e_u$ |
|---|---|---|---|---|---|---|---|---|
| OD | 2.57 | 0.23 | 0.03 | 6.8 | 0.46 | 0.2 | 0.03 | 3 |
| $\overline{\text{OD}}$ | - | 1.97 | 0.11 | 8.88 | 0.12 | 2.6 | 3e-6 | - |

## V. EXPERIMENTS

After the implementation of all approaches, their performances on the data set were tested. At the beginning, only seven of the original eighty attributes have been used. The selection was based on experimenting and calculating their Pearson correlation to the target value.

The best combination of features was found by trying various permutations of features in combination with different values of $k_1$, $k_2$, $k_3$ and $t$. The values of the weights $k_{\{1,2,3\}}$ were in $\{1, 10\}$ while $t$ was selected from the set $\{0.5, 1, 2\}$. Since the model calculations run rather fast, the complete permutations of the seven input features were tested. This approach resulted in 28,585 initial tests.

After the first batch was evaluated, it was assumed that the fuzzy linear model is more suitable when the number of features increases. Therefore, a second batch has been prepared using additional five feature dimensions, increasing the input dimension to $n = 12$. Besides, the set of the $k_{\{1,2,3\}}$ has been increased to $\{1, 10, 100, 1000\}$, and $t$ could be any of $\{0.2, 0.5, 1, 2, 3, 5\}$.

Since this confirmed a hypothesis about the influence of parameterization on the metrics observed, a third batch of experiments was executed, using all of the numerical input data except those features which can be derived from other features. This resulted in an input dimension $n = 34$.

## VI. EVALUATION

After testing many configurations of the algorithms, enough data was gathered to evaluate the results. The evaluations were conducted after each batch. Two metrics were used for comparison: the root mean squared error (RMSE) and the root mean squared logarithmic error (RMSLE). The latter is also used as a comparison metric in the Kaggle Challenge for this data set.

In the first batch, the asymmetric approach was significantly superior when compared to the symmetric approaches. In the following batches, when the number of input dimensions was increased, the symmetric approach presented by Tanaka (Tanaka and Lee 1997) performed as well as the extended asymmetric approach. Table II shows the best results that were gathered.

Rows 1 to 4 represent the best results of the first batch, which were achieved by using all seven features. Note that the first experiments also tested all permutations between the seven dimensions selected. Row 1 and 2 include the best (lowest) RMSE, and the next two rows the best RMSLE values.

The best performing symmetrical (Tanaka's) approach was the one with outlier detection (SOD). This has proven to be the best solution for the new asymmetrical version (abbreviated as AOD), too. It should be mentioned that the symmetric and asymmetric models without outlier detection were tested as well, but they never achieved the highest score of any batch.

Rows 5 to 8 show the impact of more extreme values of $k_{\{1,2,3\}}$ and $t$ on the results. Five additional input features extended these tests. The next rows (9 to 12) show how the algorithms perform on all available, nonlinear dependent input features.

TABLE II: Experimental Results

| | Alg. | Feat. | k1 | k2 | k3 | t | RMSE | RMSLE |
|---|---|---|---|---|---|---|---|---|
| 1 | SOD | 7 | 1 | 1 | 10 | 1 | 43807 | 0.730 |
| 2 | AOD | 7 | 10 | 1 | 1 | 2 | 43630 | 0.779 |
| 3 | SOD | 7 | 1 | 10 | 1 | 1 | 61499 | 0.310 |
| 4 | AOD | 7 | 1 | 10 | 10 | 2 | 49326 | 0.269 |
| 5 | SOD | 12 | 1000 | 10 | 1 | 5 | 35048 | 0.733 |
| 6 | AOD | 12 | 10 | 1 | 100 | 3 | 37786 | 0.185 |
| 7 | SOD | 12 | 1 | 1 | 10 | 2 | 37296 | 0.187 |
| 8 | AOD | 12 | 1 | 10 | 1000 | 2 | 39618 | 0.185 |
| 9 | SOD | 34 | 100 | 1000 | 1 | 1 | 30683 | 0.199 |
| 10 | AOD | 34 | 10 | 1000 | 1 | 1 | 30689 | 0.199 |
| 11 | SOD | 34 | 1 | 1 | 100 | 1 | 32901 | 0.166 |
| 12 | AOD | 34 | 1 | 1 | 100 | 1 | 33186 | 0.167 |
| 13 | L99 | 34 | 1 | 1 | - | - | 31071 | 0.216 |
| 14 | LSR | 34 | - | - | - | - | 31071 | 0.218 |

In row 13 (L99), the results of Lee's algorithm from 1999 (Lee and Tanaka 1999) are used with the parameters optimized for RMSE. The final row 14 shows the RMSE and RMSLE scores of the least-squares regression (LSR). Lee's approach and the LSR yield almost identical results. This can also be seen when comparing the $a$-component of the fuzzy coefficients with the crisp coefficients of the LSR. Hence, the difference between them is never higher than 0.02%. Since the boundaries are not squared, the weighting between the least-squares and the tunnel optimization is skewed. Thus, $k_1$ and $k_2$ can not be seen as equal weights, and the influence of $k_2$ depends on the data set. For comparison to Equation (24), Equation (29) describes the fuzzy component of the optimization problem of Lee and Tanaka's algorithm, which is similar to the approach.

$$L99_{\text{fuzzy}} = k_2 \sum_{j=1}^{p} \left( \sum_{i=0}^{n} l_i x_{ji} + \sum_{i=0}^{n} u_i x_{ji} \right) \tag{29}$$

As already mentioned, the following experiments with a higher number of input dimensions yielded the same results for symmetric and asymmetric approaches. This insight is clearly outlined in Table II (see rows 9 to 12). The final experiments resulted in an almost identical configuration for the SOD and AOD parameters.

*Interpretation of the Results*

From the experimental results, it can be concluded that the new asymmetrical approach has the advantage of boundaries, which can be derived from a richer solution space due to their independence. If a data set has only outliers above or below the linear tendency, the algorithm is not forced to give both boundaries more room in an equal, symmetric fashion. As the number of input features increased, no noticeable linear correlation between many of them and the target dimension could be found. Since the test set is rather large, this may result in a symmetrical distribution of the data points beneath and above the linear tendency of the sales price. Eventually, this explains the algorithms' pattern of returning the same results ($c_i = l_i = u_i$) for many dimensions. This results in $a_i$, which is calculated the same way. The results were measured and are shown in Figure 3.

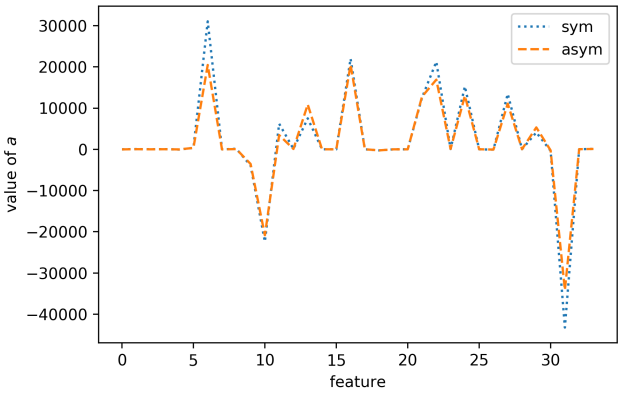


Fig. 3: Visualization of the Asymmetric and Symmetric Algorithm $a_i$ Results

## VII. CONCLUSION AND FUTURE WORK

In this work, a new algorithm has been presented using Tanaka's (Tanaka and Lee 1997) approach to generate fuzzy linear regressions based on crisp input data and outlier detection. It returns asymmetrical triangular fuzzy numbers instead of symmetrical ones. Furthermore, different fuzzy regression approaches have been tested in a real-world scenario, predicting house prices for a given data set. Based on these experiments, it can be said that the asymmetrical approach should be preferred when calculating fuzzy linear regression with crisp input, since it performs at least as good as the symmetrical idea and outperforms other fuzzy regression approaches which are dependent on the input data. Apart from these promising results, the transparent functionality of the new approach is a remarkable advantage, due to the explainability of its results. At any point during the analysis process, the results of the algorithm is comprehensible.

In the future, it has to be investigated, which parameters affect the performance of the algorithm. The asymmetric fuzzy coefficients have approximately 50% more parameters. This impacts the runtime of the QP solving algorithms. The computational effort depends on the approach chosen. A comparison with other regression algorithms with respect to transparency and accuracy, as well as algorithmic runtime would be conceivable. Furthermore, it would be interesting to apply the approach to different scenarios. Another idea could be to expand fuzzy linear regression with crisp input data to other fuzzy distributions. Currently, triangular fuzzy numbers are used to create the fuzzy linear functions, although in many cases, the data are not distributed linearly. Using Gaussian bell curves could improve the performance of the approaches, which needs to be implemented and further investigated. Additionally, a comparison of the results from this study to the output of other types of AI techniques would be very interesting.

## VIII. REFERENCES


Arrieta, Alejandro Barredo et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115.

Cock, Dean De (2011). "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project". In: *Journal of Statistics Education* 19.3.

Cortez, Paulo and Mark J. Embrechts (2011). "Opening black box Data Mining models using Sensitivity Analysis". In: *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE.

Diamond, Phil (1988). "Fuzzy Least Squares". In: *Information Sciences*, pp. 141–157.

D'Urso, Pierpaolo and Tommaso Gastaldi (2001). "Linear Fuzzy Regression Analysis with Asymmetric Spreads". In: *Advances in Classification and Data Analysis*. Springer Berlin Heidelberg, pp. 257–264.

Fan, Chenchen, Zechen Cui, and Xiaofeng Zhong (2018). "House Prices Prediction with Machine Learning Algorithms". In: *Proceedings of the 2018 10th International Conference on Machine Learning and Computing - ICMLC 2018*. ACM Press.

Lee, Haekwan and Hideo Tanaka (1999). "Fuzzy Approximations with Non-Symmetric Fuzzy Parameters in Fuzzy Regression Analysis". In: *Journal of the Operations Research Society of Japan* 42.1, pp. 98–112.

Neubauer, Dagmar (2010). "Fuzzy-Regression bei Fehlern in den Daten : Modellierung und Analysepotentiale". PhD thesis. Johann Wolfgang von Goethe University.

Patel, Janki, Miral Patel, and Mittal Darji (2018). "Stock Price Prediction Using Clustering and Regression: A Review". In: *Sustainability* 3.1, pp. 1967–1961.

Rausch, Peter and Birgit Jehle (2013). "Data Supply for Planning and Budgeting Processes under Uncertainty by Means of Regression Analyses". In: *Business Intelligence and Performance Management: Theory, Systems, and Industrial Application*. Ed. by Peter Rausch, Alaa F. Sheta, and Aladdin Ayesh. Springer U.K., pp. 163–178.

Tanaka, Hideo and Haekwan Lee (1997). "Fuzzy Linear Regression Combining Central Tendencyand Possibilistic Properties". In: *Proceedings of 6th International Fuzzy Systems Conference*. IEEE.

Tanaka, Hideo, Satoru Uejima, and Kiyoji Asai (1982). "Linear Regression Analysis with Fuzzy Model". In: *IEEE Transactions on Systems, Man and Cybernetics* 12, pp. 903–907.

Tickle, A.B. et al. (1998). "The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks". In: *IEEE Transactions on Neural Networks* 9.6, pp. 1057–1068.

Valaskova, Katarina et al. (2018). "Financial Risk Measurement and Prediction Modelling for Sustainable Development of Business Entities Using Regression Analysis". In: *Sustainability* 10, p. 2144.

Viktorovich, Parasich Andrey et al. (2018). "Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning". In: *2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC)*. IEEE.

Welc, Jacek and Pedro J. Rodriguez Esquerdo (2018). *Applied Regression Analysis for Business*. Springer International Publishing.

Wolpert, David H. (1992). "Stacked Generalization". In: *Neural Networks* 5.2, pp. 241–259.

Yan, Jiaju (2017). "Multi-Class ROC Random Forest for Imbalanced Classification". PhD thesis.

Yosinski, Jason et al. (2015). "Understanding Neural Networks Through Deep Visualization".



**RAPHAEL A. KRAUTHANN**

Raphael Krauthann studies computer science at Nuremberg Tech since 2015. In his studies, he focuses on big data and modeling.
In his professional life, he assists companies in improving their process by digitizing and automating them. Additionally, he researches in the area of embedded systems and control.



**TOBIAS KRUSE**

Tobias Kruse is a master's student in computer science at Nuremberg Tech since 2019. He earned his bachelor's degree with his thesis about time-traveling debuggers. During his bachelor studies, he worked on the conception and implementation of programming languages, compilers, and interpreters. Currently, he focuses his studies on scalable software architectures, big data systems, and data visualization.



**HINNERK JANNIS MÜLLER**

Hinnerk Müller studies Information Systems & Management at Nuremberg Tech since 2015. He designed various accounting interfaces for tour operator booking systems during his employment as a working student. His bachelor's degree focused on visualization and creation of dashboards for tour operators' data. Upcoming research for his master's thesis deals with Explainable AI.



**MICHAEL STUMPF**

Michael Stumpf is a research assistant at Nuremberg Tech. He holds a master's degree in Information Systems, and his current work and research are focused on business intelligence, business analytics, and fuzzy technologies. Additionally, he is interested in the digital transformation of processes and process automation.



**PETER RAUSCH**

Peter Rausch is a professor of Information Systems at Nuremberg Tech and has published many papers, articles, and book chapters. He holds a Ph.D. in business administration and has spent several years working in the fields of software development, business process optimization, and consulting. His current research activities are focused on fuzzy technologies, business planning, and process automation.