Modification frequency density with euclid distance classify for weather forecasting

R Wulandari 1, B Surarso2, B Irawanto3, Farikhin4,

1Postgraduate Program of Mathematics, Department of Mathematics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia
2,3,4 Department of Mathematics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

Corresponding author: dariwulan711@gmail.com

Abstract. Method of weather forecasting fuzzy time series has some drawbacks, weather forecasting needs many weather components, big data size and forecaster experience, one of which is very difficult if it is worn on large data whereas weather Forecasting using time series forecast has large data characteristics, high dimensions, and continuous also require many large components and amounts of data, this affects speed and accuracy in weather forecasts, but the science develops in tandem with the development of the fuzzy Forecasting method, the fuzzy forecasting method that can be worn on large data will facilitate the weather forecasting, in this paper will be explained methods of forecasting for large data that is using the Euclid distance to classify data, then using the frequency density partitioning modification method applied to the database that has been grouped. The data used was KW Hipel Al McLeod 1994 using year variables and temperature, data from 1782 to 1988, data covers 206 global average temperature data annually. The final solution of the weather forecasting fuzzy time series problem is to analyze the value of AFER and MSE, the value of this paper is AFER 0.0021 and MSE 0.0025, the smaller the AFER and MSE values indicate that the method is good enough to forecast the weather.

1. Introduction

Weather is the state or behavior of the atmosphere at a certain time whose nature changes over time, air has a very dynamic nature [1]. The weather has many data components and a large amount of data. This leads to poor accuracy and forecast speed. The amount of data in a forecast greatly affects the results of the forecast. A forecast that is composed of large data usually has low accuracy, and if too little data is collected, the results are also far from expected. Attempts to implement a return period or cycle system commonly used by practitioners from a particular field need to be well considered.

Forecasting is a topic that is closely related to time series data. Time series is data that consists of one object but includes several time periods such as daily, monthly, weekly, annual, and others. In general, quantitative forecasting can be applied when there are 3 of the following conditions [2]: Information about the past (historical data) is available, this information can be quantified in numerical form, it can be assumed that some aspects of the past pattern will continue. in the future. While time series data usually has large data characteristics. To deal with that problem researcher try to find pattern in time series data. Subsequence clustering is performed on a single time series [3]. To measure the distance between two subsequences, we use Euclidean distance that has been widely used in time series domain [4]. Short Time Series Distance. Short time series (STS) distance is the squared of the gradient
distance between two time series data [5]. The idea of this paper is to discretize time series data into several sub-sequences, by processing plotting the data into time series graph, the next step divides the graph into several windows called subsequence time series.

2. Preliminaries

2.1. Fuzzy time series

**Definition 1**
Suppose \( X \) is a universe of discussion. Fuzzy set \( A \) is defined as \( \mu_A: X \rightarrow [0,1] \). Where \( \mu_A \) is the membership function of the \( A \) fuzzy set, and \( \mu_A(x) \) is the degree of \( x \) membership of the \( A \) fuzzy set members.

The fuzzy triangle membership functions is denoted by \( A = [a, b, c] \) the degree of membership as follows:

\[
\mu_A(x) = \begin{cases} 
\frac{x-a}{b-a}, & a \leq x \leq b \\
\frac{c-x}{c-b}, & b \leq x \leq c \\
0, & \text{otherwise}
\end{cases}
\]

**Definition 2**
Suppose \( Y(t) (t = \ldots, 0, 1, 2, 3, \ldots) \) and \( Y(t) \subseteq R \), is the universe of discussion of the fuzzy set \( f_i(t), i = 1, 2, \ldots \) and the collections of \( f_i(t), i = 1, 2, \ldots \) is \( F(t) \) are linguistic variables too.

**Definition 3**
Suppose \( F(t) (t = \ldots, 0, 1, 2, 3, \ldots) \) is a fuzzy time series, If \( F(t) \) this happens because \( (t-1), F(t-2), \ldots, F(t-h) \) then the FLR will be represented as \( F(t-h), \ldots, F(t-2), F(t-1) \rightarrow F(t) \) and is called \( h \)-order FLR, and represent \( h = 1 \) called one-order.

**Definition 4**
Jilani, Burney, and Ardil (2007) using the AFER (Average Forecast Error Rate) and MSE (Mean Squared Error) methods to find the magnitude of the irregularities which occurred in the data results of the forecasting actual data. As for the calculation of AFER and MSE:

\[
AFER = \frac{|X_i - F_i|}{X_i} \times 100%
\]

\[
MSE = \frac{\sum_{i=1}^{n}(X_i - F_i)^2}{n}
\]

Where \( X_i \) is the actual value over \( X = \{x_1, x_2, \ldots, x_n\} \) at \( i \) time, \( y_i (i = 1, 2, \ldots, n) \) is a prediction of \( X_i \) and the smaller the AFER and MSE values the better the forecasting performance.

2.2. Time series metric euclid distance

The model using time series data mining with cluster analysis in sequence time series, as material for data simulation, the data was obtained from the time series data library (KW Hipel Al McLeon 1994) using the variable year and temperature, data used from 1782 to 1988. The data consisted of 206 records which were global average temperature data every year. The data is processed by plotting the data into a time series graph, and divides into several subsequences. The data is divided into 10 subsequences and 20 width. Using the euclid distance equation, each subsequence produces points obtained by calculating the similarity distance. Each subsequence produces 3 center points of similarity.

2.3. Metric approach
The following figure is the flowchart of the Metric Approach forecasting method.

![Flowchart of Metric Approach forecasting method](image)

**Figure 1.** Flowchart of Metric Approach forecasting method

3. **Simulation**

**Step 1 Determine the universe of discourse**

Determine the universe of discourse. From the temperature, data used from 1782 to 1988, we knew that the value of \( X_{\min} = 14.31 \) dan \( X_{\max} = 15.87 \)

\[ U = [X_{\min} - D_1, X_{\max} + D_2] = [14; 16] \]

**Step 2 Processing data mining**

The data is processed by plotting the data into a time series graph, and divides into several subsequences, the data divided into 10 subsequences and 20 width, by using euclid distance equation each subsequence produces point obtained by calculating the similarity distance. Each subsequence produces 3 center points of similarity.

**Step 3 Divide into the interval**

Divide the universe of discourse into 7 intervals.

| Table 1. Fuzzy interval |
|-------------------------|
| **7 Intervals**          |
| U1                      | [14.00 ; 14.29]     |
| U2                      | [14.29 ; 14.57]     |
| ...                    | ...                 |
| U7                      | [15.71 ; 16.00]     |

**Step 4 Define a fuzzy triangle based on table**
Let $A_1, A_2, \ldots, A_k$ be the fuzzy sets which are linguistic values of the linguistic values of the linguistic variable temperature data. In this paper, the 7 intervals are redivided into 12 intervals named linguistic values. From $u_1$ triangular membership functions can be made from the interval $[14.29; 14.57]$.

![Membership function $A_1$]

The membership function $\mu_{A_1}(x)$ is given by:

$$\mu_{A_1}(x) = \begin{cases} 
\frac{x - 14.29}{14.43 - 14.29}, & 14.29 \leq x \leq 14.43 \\
\frac{14.57 - x}{14.57 - 14.43}, & 14.43 \leq x \leq 14.57 \\
0, & \text{otherwise}
\end{cases}$$

And further by replacing the as linguistic. Based on the fuzzify historical enrollments obtained in step 4, we can get the fuzzy logical relationship group (FLRG).

**Step 5 Setting and building FLR and FLRG**

FLR (Fuzzy Logic Relation Group) is a relation that will be used to change the value of the fuzzy rule becomes the value of the crips (defuzzyfikasi). Establish fuzzy logical relationships based on the fuzzified enrollments where the fuzzy logical relationship "$A_p A_q A_r \rightarrow A_s$" denotes that "if the fuzzified enrollments of year $p, q$ and $r$ are $A_p, A_q$ and $A_r$ respectively, then the fuzzified enrollments of year $r$ is $A_r$". Then $A_j \rightarrow A_{j-1}, A_j, A_{j+1}$.

For $A_1$ is $A_1 \rightarrow A_1, A_2$.

The same rules for FLRG $A_{12} \rightarrow A_{11}, A_{12}$.

**Step 6 Defuzzyfikasi**

For the determination of the value of forecasting by determining the value of trend prediction by defuzzification of linguistic forms through the centroid method.

Trend prediction is the defuzzification form of linguistic with the centroid method is represented by the following:

$$t_j = \begin{cases} 
\frac{1 + 0.5}{a_j + 0.5}, & \text{if } j = 1 \\
\frac{0.5 + 1 + 0.5}{a_{j-1} + a_j} + \frac{0.5}{a_{j+1}}, & \text{if } 2 \leq j \leq n - 2 \\
\frac{0.5 + 1}{a_{n-1} + a_n}, & \text{if } j = n
\end{cases}$$

Where $(a_{j-1}, a_j, a_{j+1})$ is the middle value of interval fuzzy $A_{j-1}, A_j, A_{j+1}$ respectively.

**Table 2. Fuzzified historical data**
Step 7 fault finding forecasting

Jilani, Burney, and Ardil (2007) using the AFER (Average Forecast Error Rate) and MSE (Mean Squared Error) methods to find the magnitude of the irregularities which occurred in the data results of the forecasting actual data. As for the calculation of AFER:

\[
AFER = \frac{\left| X_i - F_i \right|}{X_i} \times 100% = 0.002108698
\]

The Mean Squared Error (MSE) is another method to evaluate the methods of forecasting.

\[
MSE = \frac{\sum_{i=1}^{n}(X_i - F_i)^2}{n} = 0.00248
\]

We use MSE and AFER to compare the forecasting result of different forecasting methods, where \(A_i\) denotes the actual enrollment and \(F_i\) denotes the forecasting enrollment of year \(i\).

| Window sequence | Year actual | average temperature | Increase/decrease | Fuzzy rule | FLRG |
|-----------------|-------------|---------------------|-------------------|------------|------|
| I               | 1784        | 15,14               | Start             | A7         | A6,A7,A8 |
|                 | 1790        | 15,80               | Increase          | A11        | A10,A11,A12 |
|                 | 1796        | 15,86               | increase          | A12        | A11,A12 |
|                 | 1964        | 15.04               | Decrease          | A6         | A5,A6,A7 |
|                 | 1974        | 14.96               | Decrease          | A5         | A4,A5,A6 |

| Table 3. Criteria of AFER |
|---------------------------|
| AFER | Criteria |
|<10%  | Very good |
|10%-20%| Good  |
|20%-50%| Good enough |
|>50%  | Bad     |

4. Results and discussion

4.1. Result

In this section, we have proposed a method to forecast the known historical weather data, using the Euclid distance to classify data, then using the frequency density partitioning modification method applied to the database that has been grouped.

| Window sequence | Year actual | Actual temperature | Increase/decrease | Fuzzy rule | FLRG | Forecast temperature |
|-----------------|-------------|---------------------|-------------------|------------|------|----------------------|
| I               | 1784        | 15,14               | Start             | A7         | A6,A7,A8 | 15.10855045 |
|                 | 1790        | 15,80               | Increase          | A11        | A10,A11,A12 | 15.59191325 |
|                 | 1796        | 15,86               | Increase          | A12        | A11,A12 | 15.75884775 |
According to table 4 we can make a chart of comparison between forecast result and actual value. For future work, we will compare other methods for forecasting data based on different intervals to understand the higher forecasting accuracy.

![Chart of comparison forecast result](chart.jpg)

**Figure 3.** Chart of comparison forecast result

4.2. Discussion
Research about the optimization of fuzzy rules can improve the accuracy of predictions. For further work we will include other variables that are more relevant or logical, and Fuzzy rules are more relevant according to the actual conditions.

5. Conclusion
In this paper, we have presented a modification frequency density with euclid distance classify for weather forecasting. From table 2 and figure 3, we can see that the AFER of the forecasting results 0.002108698 < 10%, values indicates that the method is good enough to forecast the weather.

References
[1] Kartasapoetra A G 2004 *Klimatologi Pengaruh Iklim terhadap Tanah dan Tanaman* Jakarta PT Bumi Aksara
[2] Makridakis S, Wheelwright S C and McGee V E 1992 *Metode dan Aplikasi Peramalan* Jakarta Erlangga
[3] Keogh E and Lin J 2005 *Knowledge and Information Systems* 8 2 154–177
[4] Rodpongpun S, Niennattrakul V, Ratanamahatana C A 2012 *Knowledge-Based Systems* pp. 361–368
[5] Zolhavarieh S, Aghabozorgi S and Teh Y W 2014 *A Review of Subsequence Time Series Clustering* *The Scientific World Journal*
[6] Jilani T A, Burney SMA and Ardil C 2007 *International Journal of Computer and Information Engineering* 4 1194 – 1199
[7] Cheng Shou Hsiung 2016 *Fuzzy Time Series Forecasting Based on Logical Relationships and Similarity Measure* 272-287

[8] Elfajar A, Setiawan and Dewi C 2017 *Peramalan Jumlah Kunjungan Wisatawan Kota Batu Menggunakan Metode Time Invariant Fuzzy Time Series* 85-94

[9] Saxena, Preetika 2012 *International Journal Computer Technology and Application* 3 957-961

[10] Fu T 2011 *Eng. Appl. Artif. Intell.* 24 1 164-181

[11] Lin J, Keogh E, Leonardi S and Patel P 2002 Proc. 2nd Work. Temporal Data Min. pp. 53-68

[12] Zarlis M, Buaton R, Efendi S 2019 *Optimization Time Series Model with RBT (Rule Based Time Series) Base On Industrial Revolution 4.0*

[13] Irawanto B, Ningrum R W, Wulandari R, Surarso B, Farikhin 2019 *Journal of Physics* 1321

[14] Winarso, Paulus Agus 2000 *Sistem Prakiraan Cuaca dan Iklim di Indonesia* Prosiding Temu Ilmiah Prediksi Cuaca dan Iklim Nasional Lembaga Penerbangan dan Antariksa Nasional pp 193-199

[15] Zakir, Achmad 2000 *Operasional Prakiraan Cuaca Jangka Pendek di Badan Meteorologi dan Geofisika* Prosiding Temu Ilmiah Prediksi Cuaca dan Iklim Nasional pp 15-17