Ecotoxicology and Environmental Safety 227 (2021) 112875

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

journal homepage: www.elsevier.com/locate/ecoenv

A new hybrid fuzzy time series model with an application to predict PM_{10} concentration

Yousif Alyousifi [a,b,*], Mahmod Othman [c], Abdullah Husin [d], Upaka Rathnayake [e]

[a] *Department of Mathematics, Faculty of Applied Science, Thamar University, Dhamar 87246, Yemen*
[b] *Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia*
[c] *Department of Foundation and Applied Science, Universiti Teknologi PETRONAS, Seri Iskandar 32160, Malaysia*
[d] *Department of Information System, Universitas Islam Indragiri, Riau, Indonesia*
[e] *Department of Civil Engineering, Faculty of Engineering, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka*

ARTICLE INFO

Edited by Dr. Hao Zhu

Keywords:
C-Means clustering technique
Forecasting
Fuzzy relationship group
Fuzzy time series (FTS)
Markov transition probability matrix

ABSTRACT

Fuzzy time series (FTS) forecasting models show a great performance in predicting time series, such as air pollution time series. However, they have caused major issues by utilizing random partitioning of the universe of discourse and ignoring repeated fuzzy sets. In this study, a novel hybrid forecasting model by integrating fuzzy time series to Markov chain and C-Means clustering techniques with an optimal number of clusters is presented. This hybridization contributes to generating effective lengths of intervals and thus, improving the model accuracy. The proposed model was verified and validated with real time series data sets, which are the benchmark data of actual trading of Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) and PM_{10} concentration data from Melaka, Malaysia. In addition, a comparison was made with some existing fuzzy time series models. Furthermore, the mean absolute percentage error, mean squared error and Theil's U statistic were calculated as evaluation criteria to illustrate the performance of the proposed model. The empirical analysis shows that the proposed model handles the time series data sets more efficiently and provides better overall forecasting results than existing FTS models. The results prove that the proposed model has greatly improved the prediction accuracy, for which it outperforms several fuzzy time series models. Therefore, it can be concluded that the proposed model is a better option for forecasting air pollution parameters and any kind of random parameters.

1. Introduction

In 1965, Zadeh (1965) proposed the fuzzy set theory and Fuzzy Logic as an extension to the already available classical crisp logics to multivariate form. Song and Chissom (1993, 1994) introduced the first-order fuzzy time series (FTS) model by replacing the values of the time series with fuzzy sets in order to deal with the fuzzy, incomplete sequences containing noise. Various FTS models have been developed by improving the three main stages, which are fuzzification, fuzzy inference, and defuzzification, to reach high accuracy of the forecasting models (Abdullah and Ling, 2012). For instance, Chen (1996) has improved Song and Chissom's model by employing fuzzy logical relation tables to reduce the computational complexity of the model. Huarng (2001) and Huarng and Yu (2006) improved the forecasting accuracy of Chen's model (Chen, 1996) by extending the model by determining the intervals using average-based and distribution-based lengths. Therefore, many forecasting methods based on this framework were proposed over the past decades. Most of these researches have used an interval-based FTS model to handle the fuzzification of the time series and have applied fuzzy logic relationships, which can be executed on the FTS dataset, for making a forecast.

Prediction of air pollution is a very important task on multiple levels - community, national and global as it is beneficial to air pollution assessment where the result found can be used for managing the air quality. Accurate forecasting enables people to plan ahead, decreasing the effects on health and the costs associated. Particularly, predicting the concentration of air pollutants is vital for assessing the effects of air pollutants on human health (Yan et al., 2019; Alyousifi et al., 2018, 2019; Wang et al., 2016). Many researchers have applied the FTS models for forecasting air pollution levels (Alyousifi et al., 2020a, 2021a,

* Corresponding author at: Department of Mathematics, Faculty of Applied Sciences, Thamar University, Dhamar 87246, Yemen.
E-mail address: yalyousifi@tu.edu.ye (Y. Alyousifi).

https://doi.org/10.1016/j.ecoenv.2021.112875
Received 9 May 2021; Received in revised form 14 September 2021; Accepted 4 October 2021

2021b). For instance, Cagcag et al. (2013) have predicted air pollution in Ankara based on a new seasonal FTS model. In addition, Koo et al. (2020) have utilized some statistical models in order to forecast the air pollution events and made a comprising to determine the adequate model. Furthermore, Cheng et al. (2011) have introduced an FTS model for forecasting daily O_3 concentrations in Taiwan. Apart from that, the FTS model based on Fuzzy K-Medoids clustering was suggested for minimizing the sensitivity of outliers in order to produce adequate forecasts of air pollution (Dincer and Akkuş, 2018). Furthermore, an integrated model of the fuzzy theory and advanced optimization algorithm was introduced in air pollution forecasting by Yang et al. (2019). Therefore, the hybrid models in prediction are quite often used in the related literature. A hybrid FTS model proposed by Wang et al. (2018) has taken the attention of the research world in predicting air pollution in China.

The hybrid models in time series forecasting usually perform better than their single models. For example, the hybrid models in conventional time series, such as the Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) models, which are combined of the autoregressive (AR) and moving average (MA) models, have shown greater improvement in the model performance than their single models (Box et al., 2015; Zhang, 2003). In addition, a hybrid model based on combining of linear model ARIMA and a nonlinear model ANN has proven improved forecasting accuracy (Zhang, 2003; Fraiha Lopes et al., 2020). Furthermore, Mohamadi et al. (2017) proposed a hybrid ARIMA-GARCH model for predicting epileptic seizures and the method was successful. Singh et al. (2020) have also applied the hybrid wavelet-ARIMA model for predicting death cases due to COVID-19 recently and their findings were impressive. They have found that the hybrid model has outperformed the single models of its components not only in predicting air pollution levels but also in some other random parameters.

Many researchers have proposed a variety of hybrid FTS models in order to address some of these issues in fuzzy time series. For instance, a hybrid FTS based Markov chain model (FTSMC) was proposed by Tsaur (2012), which was used for determining the proper weights of the fuzzy relationships among the data points of the time series pattern. FTSMC is superior to many methods available in the literature in terms of model accuracy. Nevertheless, it has a drawback in utilizing an arbitrary partitioning of intervals. Therefore, it could not determine the appropriate length of the intervals, which is considered a shortcoming in determining the effective length of the interval in this hybrid model. Apart from this, Alyousifi et al. (2020b) have proposed the use of the FTSMC model based on the optimal grid partition method for modeling air pollution events in Malaysia. Moreover, Chen and Chen (2015) constructed a hybrid FTS model based on granular computing. Zhang et al. (2020) proposed a hybrid FTS model based on multiple linear regression and clustering techniques. They found that the model demonstrates its ability in dealing with uncertainties and enhances the rate of forecasting accuracy. Furthermore, Singh (2018) proposed a new fuzzy time series model based on artificial neural networks for forecasting rainfall. He found that the model has the robustness to deal with one-factor time series data sets more efficiently than existing FTS models. Recently, Singh (2021) proposed a new fuzzy-quantum time series model which is a combination between a developed quantum optimization algorithm (QOA) and fuzzy time series. The model proposed showed its ability in converging very fast compared to the existing models and can evolve one step ahead of forecasted results.

Clustering techniques are commonly adopted by the fuzzy time series model, and they are mostly applied to determine the fuzzy sets to generate appropriate partitioning of the universe of discourse (Zhang et al., 2020; Van Tinh et al., 2016). The hybrid model of the FTS model and the C-Means clustering technique have outperformed the FTS models and can be found in the literature. For example, Van Tinh et al. (2016) have proposed a hybrid FTS and K-Means clustering for predicting enrollment at the University of Alabama, United States of America. Likewise, Kai et al. (2010) have proposed a forecasting model for FTS based on K-Means clustering. Cheng et al. (2016) have also defined a similarity with a fuzzy logic relationship and employed C-means to improve the forecasting accuracy. In addition, Chen and Chang (2010) have applied a fuzzy C-means clustering algorithm to construct fuzzy rules to make a prediction. Furthermore, Askari et al. (2015) have applied fuzzy C-means clustering in combination with FTS, which showed an improvement in the prediction accuracy of the model. However, all of these had a limitation in dealing with repetitions of observations and the number of clusters chosen randomly, which may lead to insufficient length of intervals and inadequate forecasting accuracy.

The accuracy of the FTS modeling approach depends on two main factors, namely the length of intervals and the handling of repeated fuzzy sets. Therefore, to overcome the above-stated drawbacks, this study proposes a novel hybrid FTS model (FTSMC-CMeans) by integrating the fuzzy time series with Markov chain and C-Means clustering algorithm. In particular, the C-Means clustering algorithm was applied with the optimal number of clusters for determining the appropriate length of the intervals, and the Markov chain was implemented for handling the repeated fuzzy sets and determining the proper weights. Although the idea of using the C-Means clustering algorithm for partitioning historical datasets into intervals of different lengths has been adopted by several researchers, this study is different from them. The C-Means clustering algorithm was applied after determining the optimal number of clusters in the proposed method. Moreover, it was integrated with FTS and Markov chain, which can help to improve the prediction result significantly. Therefore, it is expected that the hybrid model overcomes the limitations showcased and produces better prediction accuracy of the model. The validation of the model performance was determined by comparing it with some existing FTS models. The next sections of this paper are organized as follows. Section 2 presents the methodology, involving basic definitions, and explains the algorithms of the proposed forecasting model, and describes the key steps in detail. Section 3 demonstrates the implementation of the proposed model and algorithm using the data of TAIEX for each step. Next, Section 4 presents the experimental results of the model proposed for the TAIEX and PM_{10} data and displays the comparison of forecasting results of the model and existing models while the conclusions are given in Section 5.

2. Methodology

As stated earlier, FTS is a forecasting model based on fuzzy set theory and fuzzy logic. FTS models use fuzzy sets for their advantages in solving non-linear time series for prediction. For example, FTS can model nonlinear and uncertain systems, incorporate expert opinions and experiences in the modeling process, handle linguistic variables, do not require statistical assumptions, and finally provide adequate performance for data sets with a small number of observations, such as a sample size of 15 or 20. In general, the basic steps for designing FTS models to produce a forecast are:

(1) First, define the universe of discourse (U), divide U into an equal number of intervals
(2) Then, determine the fuzzy sets on the universe of discourse and fuzzify the time series
(3) Next, establish the model of the existing fuzzy logic relationships in the fuzzified time series and
(4) Finally, calculate the forecasts and defuzzify the forecast values.

2.1. *Basic fuzzy time-series definitions*

In this subsection, the main FTS definitions are listed below.

A fuzzy set is a class with varying degrees of membership in the set. Let U be the universe of discourse, $U = \{u_1, u_2, \ldots, u_n\}$, which is discrete and finite, then fuzzy set A of U can be defined as given in Eq. (1).

$$A = \frac{f_A(u_1)}{u_1} + \frac{f_A(u_2)}{u_2} + \ldots = \sum_i \frac{f_A(u_i)}{u_i} \tag{1}$$

where

f_{A_i} is the membership function of fuzzy set A; $f_{A_i}: U \rightarrow [0,1]$, $f_{A_i}(u_r) \in [0,1]$ and $1 \leq r \leq n$. The $f_A(u_i)$ is the degree of membership of the element u_i in the fuzzy set A. Based on these the following time series definitions can be listed.

Definition 1. Let $X(t)(t = 0,1,2,\ldots), X(t)(t = 0,1,2,\ldots)$, a subset of real numbers, be the universe of discourse on which fuzzy sets $f_j(t)$ $(j = 1,2,\ldots)$ are defined, and let $F(t)$ be a collection of $f_1(t), f_2(t)$.
Then $F(t)$ is called an FTS on $X(t)$ (Song and Chissom, 1993, 1994).

Definition 2. If $F(t)$ is caused by $F(t-1)$, i.e., $F(t-1) \rightarrow F(t)$, then this relationship can be represented as shown in the following Eq. (2).

$$F(t) = F(t-1) \circ R(t, t-1) \tag{2}$$

where $R(t, t-1)$ is a fuzzy relationship between $F(t)$ and $F(t-1)$. Here, R is the union of fuzzy relations and "∘" is the max-min composition operator. It is also called the first-order model of $F(t)$ (Song and Chissom, 1993, 1994).

Definition 3. Let $F(t)$ be an FTS, and $R(t, t-1)$ be the first−order model of $F(t)$. If $R(t, t-1) = R(t-1, t-2)$ for any time t, i.e., $R(t, t-1)$ is independent of t, and $F(t)$ only has finite elements. Then $F(t)$ is called a time-invariant FTS. Otherwise, it is called a time-variant FTS (Song and Chissom, 1993, 1994).

Definition 4. Suppose that $F(t-1) = A_i$ and the $F(t) = A_j$. The relationship between two consecutive observations $F(t-1)$ and $F(t)$ is referred to a fuzzy logical relationship (FLR), which can be defined as $A_i \rightarrow A_j$, where A_i and A_j are the left-hand and right-hand sides (or the previous state and current state) of FLR, respectively (Song and Chissom, 1993, 1994; Alyousifi et al., 2019).

Definition 5. Suppose the FLRs, $A_i \rightarrow A_{j1}$, $A_i \rightarrow A_{j2}$, …, $A_i \rightarrow A_{jm}$. If the FLRs having the same previous state, then they can be grouped into the same fuzzy logical relationship group (FLRG) (Chen, 1996). Thus, these FLRs can be grouped into the same FLRG as: $A_i \rightarrow A_{j1}$, A_{j2}, …, A_{jm} (Song and Chissom, 1993, 1994).

2.2. C-Means clustering technique

The C-Means clustering technique (Hartigan, 1979; Zhang and Zhu, 2012) is one of the well-known unsupervised learning algorithms. It is a partitioning clustering algorithm, which partitions a given data set into a set of C clusters. The result of the C-Means clustering technique depends on the number of clusters. Apart from that, the main matter in partitioning clustering is determining the optimal number of clusters in a data set. For instance, the C-Means clustering requires the user to specify the number of clusters k in order to be generated. Accordingly, in this paper, an optimal number of clusters was determined based on three different methods, which are elbow, average silhouette, and gap statistic methods (Zhang and Zhu, 2012), through using two functions in R programming, which are *fviz_nbclust() and NbClust()* (Hartigan, 1979). By using these functions, the optimal number of clusters determines the best partition selected. For further details about these methods, the reader can see Hartigan (1979) and Kaufman and Rousseeuw (2009). The C-Means algorithm which was used in this study can be summarized as Pseudo-code in Table 1.

2.3. Hybrid fuzzy time series model

The algorithm of this study was adopted from the arithmetic processes proposed by Tsaur (Alyousifi et al., 2018). The flow chart of the calculations in the proposed model is shown in Fig. 1.

The steps of the model's algorithm are as follows based on Fig. 1.

Table 1
Pseudo-code of C-Means clustering algorithm with an optimal number of clusters.

```
Input: $C$ i.e, the number of clusters and the set of points $(x_1, x_2, \ldots, x_k)$.
1. Choose an optimal number of clusters $(C)$ using fviz_nbclust () and NbClust () R functions and obtain the data points.
2. Place the centroids $c_1, c_2, \ldots, c_k$.
3. For each data point $x_i$:
     - Find the nearest centroid $(c_1, c_2, \ldots, c_k)$
     - Assign the point to that cluster.
4. Update $\{x_i\}$ to minimize $SS(C_k) = \sum_{i=1}^{k} \sum_{x_i \in C_k} (x_i - \overline{x}_k)^2$, $C$is the cluster, $x_i$ design a data point belonging tothe cluster $Ck$and$\overline{x}_k$ is the mean value of the points assigned to the cluster $C_k$.
5. For each cluster $j = 1..k$ - new centroid = mean of all points assigned to that cluster.
6. Repeat steps 3,4 and 5 until convergence or until the end of a fixed number of iterations.
End
```

Step 1. Define the universe of discourse (U).

Step 2. Divided U into subintervals based on the C-Means Clustering technique.

Step 3. Define the fuzzy sets A_i for each time series observation on U. A fuzzy set A_i $(i = 1, 2, \ldots, n)$ can be defined as shown in Eq. 3.

$$A_i = \frac{f_{A_i}(u_1)}{u_1} + \frac{f_{A_i}(u_2)}{u_2} + \ldots + \frac{f_{A_i}(u_n)}{u_n} \tag{3}$$

where f_{A_i} is the membership function of A_i; $f_{A_i}: U \rightarrow [0, 1]$.

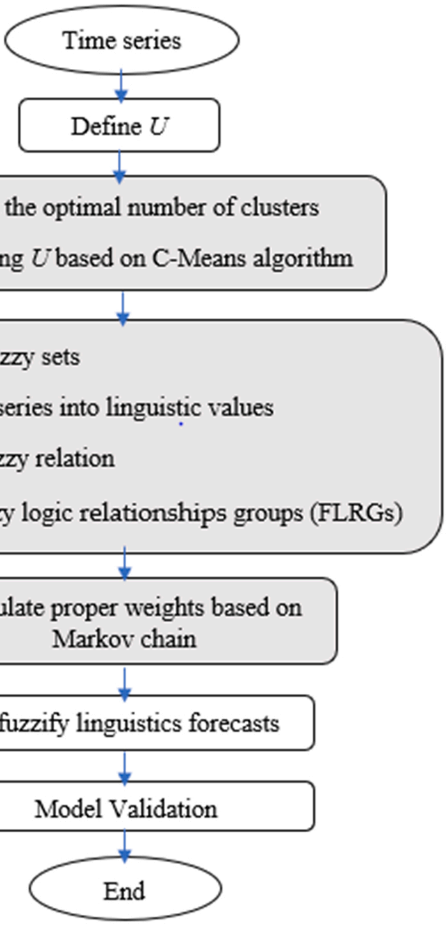

Time series
Define U
* Select the optimal number of clusters
* Partitioning U based on C-Means algorithm
* Define the fuzzy sets
* Fuzzify time series into linguistic values
* Determine fuzzy relation
* Establish fuzzy logic relationships groups (FLRGs)
Calculate proper weights based on Markov chain
Defuzzify linguistics forecasts
Model Validation
End

Fig. 1. Flowchart for the hybrid model.

$f_{A_i}(u_r) \in [0,1]$ and $1 \le r \le n$.

Step 4. Fuzzify the actual values of the time series into fuzzy numbers based on the maximum membership value in accordance with the intervals in Step 2.

Step 5. Establish the fuzzy logical relationships (FLRs) and determine fuzzy logical relation groups (FLRGs).

Step 6. Create the Markov transition probability matrix based on FLRGs. The matrix P is $P_{n\times n}$, and P_{ij} is transition probability from state A_i to state A_j. P_{ij} can be calculated as shown in Eq. (4).

$$P_{ij} = \frac{N_{ij}}{Ni.}, i,j = 1,2,\ldots,n \tag{4}$$

where N_{ij} is the frequencies and $N_{i.} = \sum_{j=1}^{n} N_{ij}$ is the total frequencies.

Step 7. Calculate the forecasted values by considering the following two cases.

Case 1. If the FLRG of A_2 is one-to-one, i.e., $A_i \rightarrow A_k$, with $P_{ik} = 1$ and $P_{ik} = 0, j \neq k)$ then the forecasting of $F(t)$ is m_k , the midpoint of $u_k, k = 1, 2.. n$, which is determined by using Eq. (5).

$$F \quad (t+1) = \quad m_k \quad P_{ik} \quad = \quad m_k \tag{5}$$

Case 2. If the FLRG of A_i is one-to-many, i.e., $A_i \rightarrow A_1, A_2, \ldots A_n$, $i = 1,2, \ldots, n$). Thus, if thestate is A_i for the actual value $Y(t)$ at time t, the forecasted value $F(t+1)$ is calculated according to Eq. (6).

$$F \ (t+1) = m_1 p_{i1} + m_1 p_{12} + \ldots + m_{i-1} p_{i(i-1)} + Y(t) p_{ii} + m_{i+1} p_{i(i+1)} + \ldots + m_n p_{in} \tag{6}$$

where $m_1, \quad m_2, \quad \ldots, \quad m_n$are the midpoint of $u_1, \quad u_2, \quad \ldots, u_n$ and m_i replaced by $Y(t)$.

Step 8. Adjust the predicted values by adding the differences of actual values $Y(t)$. The adjusted forecasted values can be written by Eq. 7.

$$\widehat{F} \quad (t+1) = F \quad (t+1) + diff(Y(t)) \tag{7}$$

3. Proposed model and algorithm

The proposed model involved two key aspects which were applied to approach the lengths of intervals and proper weights on time series data to increase the forecasting accuracy. First, the original historical data were used instead of the variations of historical data in the forecasting model. Second, the C-Means algorithm was developed to adjust the interval lengths to obtain the optimal partition. The proposed model algorithm was implemented using two datasets to verify the proposed model's effectiveness. First, the algorithm of the proposed model was implemented for TAIEX data, and its results were provided for each step. A detailed explanation of the proposed model is given in subsection 3.1. Second, the algorithm is implemented for the PM_{10} concentration (the results of each step are not shown here due to length limitations). The best model is selected based on the smallest value found of the statistical criteria presented in Section 3.2.

3.1. Implementation of the hybrid forecasting model

The algorithm of the proposed model was applied to the time series of the TAIEX data from the 5th of January 2015 to the 30th of May 2015 to forecast the TAIEX. The model was implemented, and its results were provided for each step. A time series plot of the TAIEX data is given in Supplementary Figure S1. To validate the proposed model, the TAIEX data was used to evaluate the performance and compare it with existing models. The steps of the model algorithm are given below.

Step 1. Define the universe of discourse U from TAIEX data. Since $U = [D_{min} - D_1, \quad D_{max} + D_2]$,

then, $U = [9048.34 - 48.34, \quad 9973.12 + 26.88]$, thus, $U = [9000, \quad 10000]$

Table 2
TAIEX values expressed as fuzzy numbers.

N	Date	TAIEX Value	linguistic values	Fuzzy set relationships
1	**2015/1/5**	9274.11	A2	–
2	**2015/1/6**	9048.34	A1	A2→A1
3	**2015/1/7**	9080.09	A2	A1→A2
4	**2015/1/8**	**9238.03**	A2	A2→A2
5	**2015/1/9**	9215.58	A2	A2→A2
6	**2015/1/12**	9178.3	A2	A2→A2
7	**2015/1/13**	9231.8	A2	A2→A2
:	:	:	:	:
:	:	:	:	:
94	**2015/5/27**	9693.54	A17	A15→A17
95	**2015/5/28**	9712.84	A17	A17→A17
96	**2015/5/29**	9701.07	A17	A17→A17

.

Step 2. Partitioning the universe of discourse. The C-Means clustering method was applied for partitioning the universe of discourse U as shown in Supplementary Figure S2, where each partition has a different length. Particularly, the U has been partitioned into 19 intervals with unequal lengths as follows.

u_1= [9000, 9088.907]	u_2= [9088.907, 9212.099]
u_3= [9212.099, 9366.364]	u_4= [9366.364, 9432.377]
u_5= [9432.377, 9466.752]	u_6= [9466.752, 9500.015]
u_7= [9500.015, 9512.45]	u_8= [9512.45, 9529.959]
u_9= [9529.959, 9566.25]	u_{10}= [9566.25, 9579.918]
u_{11}= [9579.92, 9605.19]	u_{12}= [9605.19, 9611.19]
u_{13}= [9611.19, 9615.54]	u_{14}= [9615.54, 9618.41]
u_{15}= [9618.41, 9623.735]	u_{16}= [9623.735, 9644.033]
u_{17}= [9644.033, 9688.38]	u_{18}= [9688.38, 9752.852]
u_{19}= [9752.852, 10000]	

Step 3. Define fuzzy sets. The C-Means method of partitioning the universe of discourse U is demonstrated here. Fuzzy sets A_i, $(i = 1, 2, \ldots, n)$ were determined based on the interval u_i that has already formed using the C-Means method in the previous step with the membership function based on Eq. (3) as follows (refer to Eq. 8).

$$A_i = \left\{ \begin{array}{c} \frac{1}{u_1} + \frac{0.5}{u_2} i = 1 \\ \frac{0.5}{u_1} + \frac{1}{u_2} + \frac{0.5}{u_3} 2 \le i \le n-1 \\ \frac{0.5}{u_{n-1}} + \frac{1}{u_n} i = n \end{array} \right\} \tag{8}$$

Supplementary Table S1 reveals the fuzzy sets A_i. A greater value of i indicates that the fuzzy set of TAIEX values will move from the lowest to the highest fuzzy set of TAIEX values.

Step 4. Fuzzify the dataset into linguistic values. Transform the TAIEX data into fuzzy numbers and determine the fuzzy logic relationships (FLRs) as can be observed in Table 2, which reveals the alterations of the observed TAIEX to be the linguistic values. Since u_1 has the maximum membership degree in fuzzy set A_1, observation 9274.11 is transferred to fuzzy set A_2, and 9048.34 is transferred to fuzzy set A_1, meaning that all data of TAIEX are fuzzified similarly. The TAIEX values and the corresponding fuzzified values found from the fuzzification process are reported in Table 2.

Step 5. Establish fuzzy logical relationship groups (FLRGs) and the frequency (account) matrix of the fuzzy relation between observations. This step shows that the FLRs may be grouped into fuzzy logic relationship groups (FLRGs).

The groups given in Table 3 represent eighteen groups of the fuzzy time series found with multiple FLRs. Moreover, it can be seen from Table 3 that a transition frequency matrix or frequency matrix (count) of

Table 3
Fuzzy logical relationship groups and Markov weighted FTS based on the C-Means method for TAIEX data.

Fuzzy logical relationships Group (FLRGs)	Markov weight elements for each group
A1 → (2)A1, A2, A18	A1→ A1(0.5), A18(0.25), A2(0.25)
A2→ A1, (8)A2, A3	A2 → A1(0.1), A2(0.8), A3(0.1)
A3 → (2)A3, A4, A5	A3 → A3(0.5), A4(0.25), A5(0.25)
A4 → A3, A4, A5, A7	A4 → A3(0.25), A4(0.25), A5(0.25), A7(0.25)
A5 → A4, A5, A6, A8	A5 → A4(0.25), A5(0.25), A6(0.25), A8(0.25)
A6 → A8, A9	A6 → A8(0.5), A9(0.5)
A7 → A4, A5, A7, A8	A7 → A4(0.25), A5(0.25), A7(0.25), A8(0.25)
A8 → A7,A8,A12, A13, A15, A16	A8 → A12(0.143), A13(0.143), A15(0.143), A16 (0.286), A7(0.143), A8(0.143)
A9 → (2)A8, A10, A12, A14	A9 → A10(0.2), A12(0.2), A14(0.2), A8(0.4)
A10 → A7, A9, A12, A15, A17	A10 → A12(0.2), A15(0.2), A17(0.2), A7(0.2), A9(0.2)
A11 → A12	A11 → A12(1.0)
A12 → A10, A11, A12, (2)A14, A15, A18	A12 → A10(0.143), A11(0.143), A12(0.143), A14 (0.28), A15(0.143), A18(0.143)
A13 → A10	A13 → A10(1.0)
A14 → A6, A9, A10, A16	A14 → A10(0.25), A16(0.25), A6(0.25), A9(0.25)
A15 → A8, A15, (2)A16, A17	A15 → A15(0.2), A16(0.4), A17(0.2), A8(0.2)
A16 → A9, A15, (2)A17	A16 →A 15(0.2), A17(0.4), A9(0.4)
A17 → A10, (2)A12, A14, (10)A17, A18	A17 → A10(0.067), A12(0.133), A14(0.067), A17 (0.66), A18(0.067)
A18 → A1, (2)A17, (5) A18	A18→ A1(0.125), A17(0.25), A18(0.625)

the fuzzy relationship between observations can be determined, which can be represented as a count matrix $\boldsymbol{N}_{19\times 19}$, as presented in Supplementary Figure S3.

Step 6. Assign the Markov weights based on the matrix of frequencies from step 5 by using Eq. (4). By defining 19 states for each of the fuzzy sets, matrix $\boldsymbol{P}_{19\times 19}$ is produced. Given the number of transitions of the fuzzy values and the elements of Markov weights per group, the obtained Markov weights based on FLRGs as shown in Table 3. The Markov weights found can be used for establishing the transition probability matrix $\boldsymbol{P}_{19\times 19}$, which can be used for calculating the forecasting values in the next step. Then, the transition process diagram can be established using the weights to visualize the Markov weighted matrix (given in Eq. 9), as shown in Supplementary Figure S4.

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1\quad 19} \\ p_{21} & p_{22} & & p_{2\quad 19} \\ \vdots & & \ddots & \vdots \\ p_{19\quad 1} & p_{19\quad 2} & \cdots & p_{19\quad 19} \end{bmatrix} \tag{9}$$

where $p_{ij} = \frac{N_{ij}}{N_{i.}}$ is the transition probability from state A_i to A_j.

For example, in the case of FLRG, $A_1 \rightarrow A_1, A_2, A_{18}$. Then, $N_{11} = 4$, $N_{12} = 2$, $N_{1\quad 18} = 2$ and $N_{1.} = 8$.Thus, $p_{11} = 0.5$, $p_{12} = 0.25$ and $p_{1\quad 18} = 0.25$; otherwise, $p_{1j} = 0$. Similarly, defining fuzzy sets, the fuzzy logical relationships, the fuzzy logical relationship groups, and Markov weights can be found using the other partition methods considered in this study.

Step 7. Forecast values were calculated by using Eqs. (4) or (5) based on Markov weights. For example, the forecast value for the next day (6/1/2015) was calculated by using Eq. (5) as presented in Eq. 10.

$$F_{t+1} \quad (2015/1/6) = \quad m_1 \quad p_{11} + Y(t) \quad p_{12} + m_3 \quad p_{13} + 4 \quad p_{14} + m_5 \quad p_{15}$$

$$F_{t+1} \quad (2015/1/6) = (9044.453)(1/10) + (9274.11)(8/10) + (9289.231)(1/10) \tag{10}$$

$$F_{t+1} \quad (2015/1/6) = 9252.657$$

In the same way, the forecast values were calculated based on the obtained results of each partition method in order to fit the optimum partition method that provides the best results.

Step 8. The forecasted values were adjusted based on Eq. (7). For example, the adjusted values $\widehat{F}$ can be calculated as follows.

$\widehat{F} \quad (2015/1/6) = \quad F_{t+1} \quad (2015/1/6) \mp \quad |\text{diff} \quad (Y(t), \quad m_2)| = 9252.657 - 123.607 \quad = \quad 9129.05$

Similarly, the other forecasted values were calculated.

3.2. Model evaluation

It is important to check the forecasting performance to identify the best model with the smallest error. Several statistical tests can be considered to measure model validation. The three statistical criteria used in this study in order to validate the forecasting accuracy of the proposed model are MAPE, RMSE, and Thiels' U-statistics, which are given in Eqs. (11–13).

$$\text{MAPE} \quad = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - F_i}{Y_i}\right| \times 100 \tag{11}$$

$$\text{RMSE} \quad = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - F_i)^2}{N}} \tag{12}$$

$$\text{Theil}'\text{sU} = \frac{\sqrt{\sum_{i=1}^{N}(Y_i - F_i)^2}}{\sqrt{\sum_{i=1}^{N}Y_i^2} \quad + \quad \sqrt{\sum_{i=1}^{N}F_i^2}} \tag{13}$$

where Y_i is the real data, F_i are the forecasted values and N is the total number of observations. These statistical criteria are widely used in the literature, which can be calculated directly from real data and predicted values, rather than dealing with unknown parameters that must be estimated. The statistical calculations measure the residual errors with the smallest value and are selected as the best model for prediction.

4. Experimental results

The proposed model was implemented to forecast the daily TAIEX data from 1st of May 2015–31st of December 2015, and also addressed other forecasting issues, such as the empirical data for weekly PM_{10} concentrations from 1st of January 2012–31st of December 2014 collected from Melaka, Malaysia. These two datasets were also used to conduct a comparative study. To validate the proposed model, both datasets were used to evaluate the model performance and compare it with existing models.

4.1. TAIEX forecasting

Table 4 and Fig. 2 show a comparison between the proposed model and some existing fuzzy time series models proposed by Chen's model (Chen, 1996), Sliva2 et al.'s model (Silva et al., 2017), Yu's model (Yu, 2005), Cheng's model (Cheng et al., 2006), Severiano et al.'s model (Severiano et al., 2017), Sliva et al.'s model (Silva et al., 2019), Sadaei et al.'s model (Sadaei et al., 2014), and Tsuar's model (Tsaur, 2012) to

Table 4
Statistical criteria of the Hybrid model and some fuzzy time series models using the TAIEX.

	Model	RMSE	MAPE	U-Statistic
1	Chen's model	150.24	1.37	2.32
2	Yu's model	145.79	1.25	2.26
3	Cheng's model	146.92	1.33	2.27
4	Sliva et al.'s model	133.49	1.02	2.06
5	Tsuar's model	136.09	1.19	2.11
6	Sadaei et al.'s	158.79	1.34	2.46
7	**Hybrid Model**	**66.84**	**0.51**	**1.03**

the TAIEX data.

It can be seen from Table 4 that the hybrid model outperforms the compared models, producing better results with the smallest error values of RMSE, MAPE and U-statistics than the error values of the existing models. The hybrid model provides the most accurate prediction and was consistent in all statistical criteria, as shown in Table 4. The RMSE, MAPE and U-statistics are 66.84, 0.51, and 1.03, respectively. They indicate that the proposed model is adequate and capable of providing reasonable prediction accuracy. Moreover, the results are supported by Fig. 2(a) and (b) as the original TAIEX compared to the predicted values based on the hybrid model are quite similar compared to the other models.

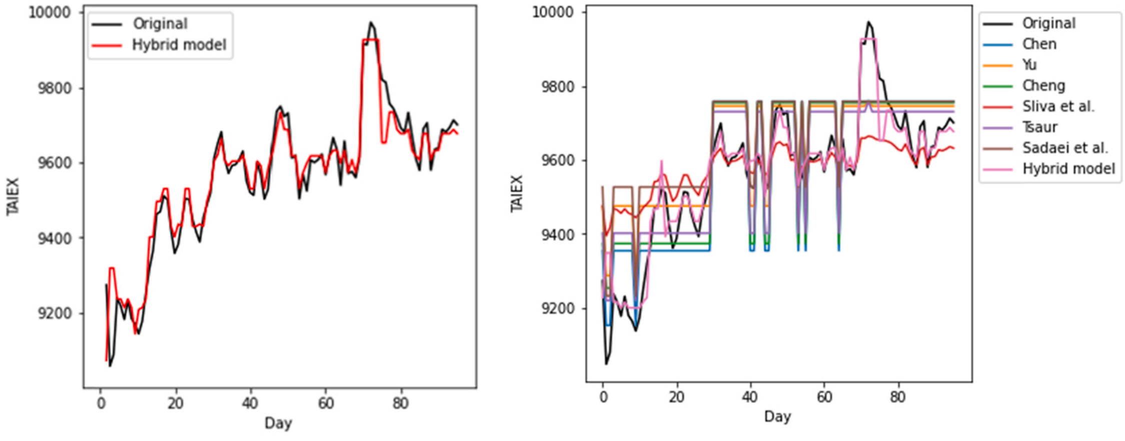
4.2. *Air pollution forecasting*

The algorithm of the model was also applied to training and test data sets of PM_{10} concentrations to evaluate the performance of the model and compare it with some of the existing fuzzy time series models to further validate the proposed hybrid model. To implement the algorithm of the proposed model, similar to Section 3.1, it was started by defining the universe of discourse U from PM_{10} data and then applying the C-Means clustering method. Based on the C-Means clustering method, the universe of discourse was divided into 25 intervals of unequal length. The found intervals were used to form the fuzzy logic relation group and assigned Markov weights. Due to the length limitation of the paper, the results of each step are not presented here. Therefore, only the final results and the comparative study between the proposed model and the existing models are presented. As can be seen in Table 5, the proposed model produces the smallest values of the applied statistical criteria compared to the existing models. This shows that the model outperforms the existing models, which means that the proposed hybrid model is a better option for predicting PM_{10} concentration.

It can be seen from Fig. 3(a) that the proposed model performs significantly well with negligible errors (compared to other models), indicating that the predicted values based on the hybrid model are quite similar to the original data of PM_{10}. Moreover, Fig. 3(b) demonstrates the results of the comparison between the proposed model and some existing models using the training dataset of PM_{10} concentrations. It was found that the model predicted the PM_{10} data are well within the acceptable levels and produced the smallest error. The hybrid model produces the most accurate prediction with respect to all statistical criteria as shown in Table 5. For the training data, RMSE, MAPE and U-statistics are 7.55, 6.83 and 0.55, respectively. Similarly, the results for the test data are 5.01, 7.25 and 0.45, respectively. Moreover, the proposed model outperformed all the existing models. Moreover, the model showed its superiority compared to the other models. This implies that the proposed model is an improved option for predicting air pollution events. Similarly, Fig. 3(c) shows a comparison between the proposed model and some existing models using the test data set of PM_{10} concentrations, indicating that the predicted values based on the hybrid model are quite similar to the original data of PM_{10}, indicating that the proposed model is superior to the existing models. The purpose of PM_{10} prediction is to act as an early warning system for air quality control and management to keep air quality within the specified guidelines.

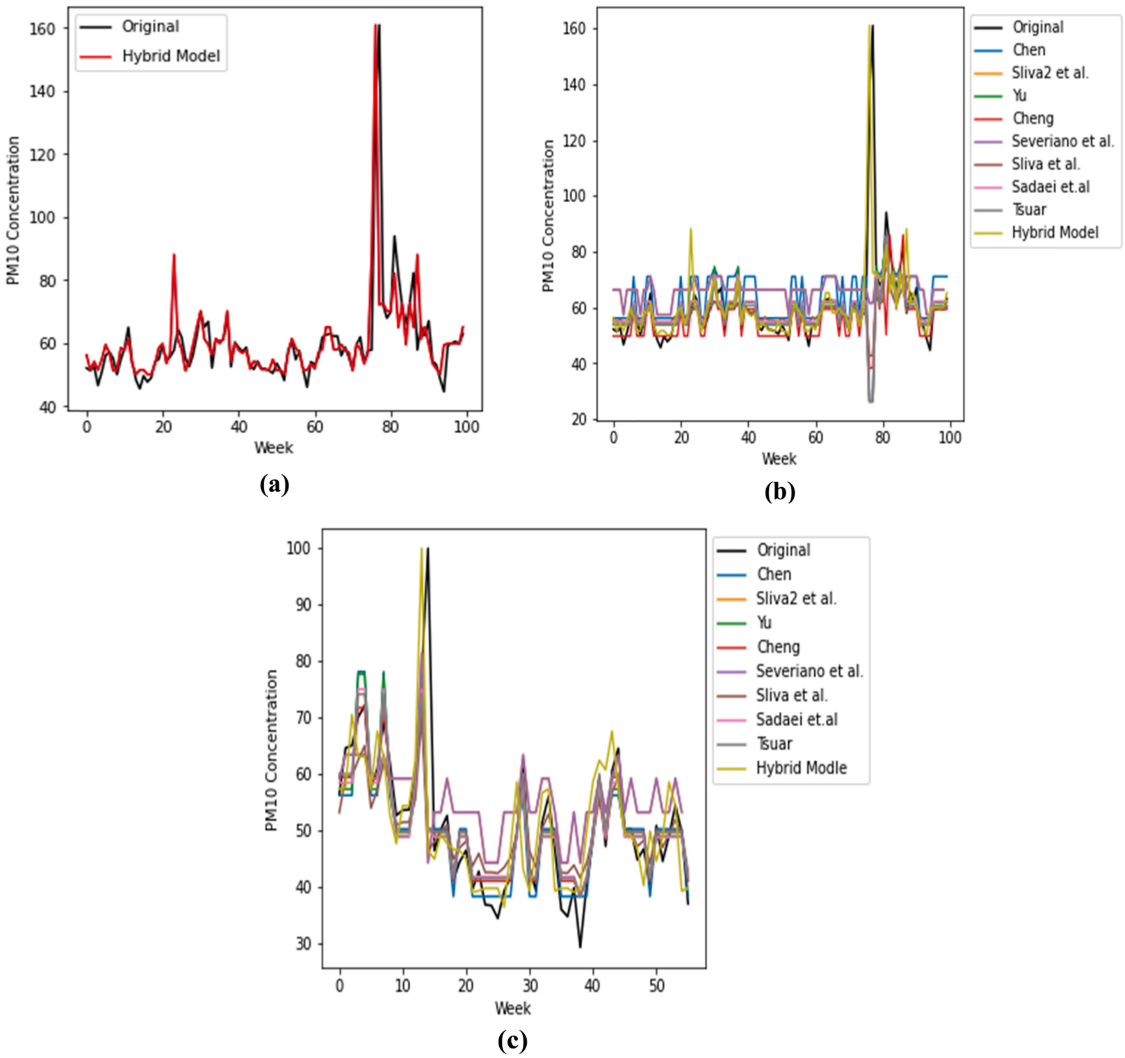
Table 5
Statistical criteria of the hybrid model and eight fuzzy time series models using the training and testing PM10 data.

	Model	Using training dataset			Using testing dataset		
		RMSE	MAPE	U-Statistic	RMSE	MAPE	U-Statistic
1	Chen's model	18.04	14.67	1.32	9.12	14.26	0.80
2	Sliva2 et al.'s model	16.10	17.79	1.18	10.23	18.68	0.89
3	Yu's model	17.06	9.74	1.25	8.87	14.20	0.78
4	Cheng's model	16.70	10.88	1.22	8.67	13.69	0.76
5	Severiano et al.'s	16.10	17.79	1.18	10.23	18.68	0.89
6	Sliva et al.'s model	15.67	9.28	1.14	8.90	13.92	0.78
7	Sadaei et al.'s	17.06	9.82	1.25	8.73	13.99	0.77
8	Tsuar's model	17.07	9.40	1.25	8.66	13.86	0.76
9	**Hybrid Model**	**7.55**	**6.83**	**0.55**	**5.01**	**7.25**	**0.45**



(a) **(b)**

Fig. 2. (a) comparison of original TAIEX vs the forecasted values based on the hybrid model; and (b) comparison of the model with some existing FTS models.



(a)

(b)

(c)

Fig. 3. (a) comparison of original PM_{10} vs the forecasted values based on the hybrid model; (b) comparison of the model with eight existing FTS models using training data; and (c) comparison of the model with eight existing FTS models using testing data.

5. Conclusions

A novel hybrid fuzzy time series model was proposed, which was implemented for TAIEX and PM_{10} concentration prediction. The model can also be implemented for many types of time series data, especially for non-seasonal data. Hybridization of the model has contributed in producing adequate partitioning and improved the model accuracy accordingly. The proposed model was investigated by comparing it with several FTS models known in the literature. The comparison has clearly shown the ability of the model to avoid the arbitrary selection of intervals and to deal with recurrent observations, which significantly improves the model accuracy. The proposed prediction method is a promising method to improve prediction accuracy. The experimental results showed that the hybrid model obtains higher prediction accuracy compared to the existing FTS model. Moreover, the empirical results demonstrated that the high-order FTS model outperformed the first-order FTS model with a lower prediction error. For TAIEX and air pollution prediction, the proposed model has achieved superior prediction accuracy compared to the conventional and advanced time series methods proposed in the literature. Moreover, the prediction values found by the model show its flexibility in FTS for air pollution prediction. In general, the proposed model has the flexibility to be applied to many types of time series data. In future studies, a comparative study using other partitioning methods such as automatic clustering and fuzzy K-medoid clustering (FKM) in combination with the FTSMC model will be conducted to investigate the method which is more powerful to improve the model performance and then to achieve high prediction accuracy.

Funding

This research was funded by International Collaborative Research Fund, Universiti Teknologi PETRONAS, Malaysia, grant number (015ME0-222).

CRediT authorship contribution statement

Y.A. conceived of the presented idea and he developed the model and performed the computations. **M.O.** and **A.H.** verified the analytical methods. **Y.A.** encouraged **U.R.** to investigate the paper structure and revised the findings of this work. All authors discussed the results and contributed to the final revised manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors are grateful to the Department of Environment Malaysia for providing air pollution data (PM_{10}). In addition, the authors would like to thank Universiti Teknologi PETRONAS for providing financial support and good facilities. The authors also wish to thank the anonymous reviewers for their critical comments and views that led to the improvement of this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ecoenv.2021.112875.

References

Abdullah, L., Ling, C.Y., 2012. Intervals in fuzzy time series model preliminary investigation for composite index forecasting. ARPN J. Syst. Softw. 2 (1), 7–11.

Alyousifi, Y., Ibrahim, K., Kang, W., Zin, W.Z.W., 2019. Markov chain modeling for air pollution index based on maximum a posteriori method. Air Q. Atmos. Health 1–11.

Alyousifi, Y., Kıral, E., Uzun, B., Ibrahim, K., 2021a. New application of fuzzy Markov chain modeling for air pollution index estimation. Water Air Soil Pollut. 232 (7), 1–13.

Alyousifi, Y., Masseran, N., Ibrahim, K., 2018. Modeling the stochastic dependence of air pollution index data. Stoch. Environ. Res. Risk Assess. 32, 1603–1611.

Alyousifi, Y., Othman, M., Almohammedi, A.A., 2021b. A novel stochastic fuzzy time series forecasting model based on a new partition method. IEEE Access 9 (5), 80236–80252.

Alyousifi, Y., Othman, M., Faye, I., Sokkalingam, R., Silva, P.C., 2020a. Markov weighted fuzzy time-series model based on an optimum partition method for forecasting air pollution. Int. J. Fuzzy Syst. 22, 1468–1486.

Alyousifi, Y., Othman, M., Sokkalingam, R., Faye, I., Silva, P.C., 2020b. Predicting daily air pollution index based on fuzzy time series Markov chain model. Symmetry 12 (2), 293.

Askari, S., Montazerin, N., Zarandi, M.F., 2015. A clustering-based forecasting algorithm for multivariable fuzzy time series using linear combinations of independent variables. Appl. Soft Comput. 35, 151–160.

Box, George E.P., Gwilym, M.Jenkins, Reinsel, Gregory C., Ljung, Greta M., 2015. Time Series Analysis: Forecasting and Control, 5th ed. John Wiley & Sons, Hoboken, New Jersey.

Cagcag, O., Yolcu, U., Egrioglu, E., Aladag, C.A., 2013. Novel seasonal fuzzy time series method to the forecasting of air pollution data in Ankara. Am. J. Intell. Syst. 3 (1), 13–19.

Chen, S.M., 1996. Forecasting enrolments based on fuzzy time series. Fuzzy Sets Syst. 81 (3), 311–319.

Cheng, C.H., Chen, T.L., Chiang, C.H., 2006. Trend-weighted fuzzy time-series model for TAIEX forecasting neural information processing. Springer Berlin / Heidelberg. Lect. Notes Comput. Sci. 42 (34), 469–477.

Cheng, S.-H., Chen, S.-M., Jian, W.-S., 2016. Fuzzy time series forecasting based on fuzzy logical relationships and similarity measures. Inf. Sci. vol. 327, 272–287.

Cheng, C.H., Huang, S.F., Teoh, H.J., 2011. Predicting daily ozone concentration maxima using fuzzy time series based on a two-stage linguistic partition method. Comput. Math. Appl. 62 (4), 2016–2028.

Chen, S.-M., Chang, Y.-C., 2010. Multi-variable fuzzy forecasting based on fuzzy clustering and fuzzy rule interpolation techniques. Inf. Sci. 180 (24), 4772–4783.

Chen, M.Y., Chen, B.T., 2015. A hybrid fuzzy time series model based on granular computing for stock price forecasting. Inf. Sci. 294, 227–241.

Dincer, N.G., Akkuş, Ö., 2018. A new fuzzy time series model based on robust clustering for forecasting of air pollution. Ecol. Inform. 43, 157–164.

Fraiha Lopes, R.L., Fraiha, S.G., Gomes, H.S., Lima, V.D., Cavalcante, G.P., 2020. Application of hybrid ARIMA and artificial neural network modelling for electromagnetic propagation: an alternative to the least squares method and ITU recommendation P. 1546-5 for Amazon urbanized cities. Int. J. Antennas Propag. 18 (3), 1–12.

Hartigan, J.A., 1979. A K-means clustering algorithm: algorithm AS 136. Appl. Stat. 28 (11), 126–130.

Huarng, K., 2001. Effective lengths of intervals to improve forecasting in fuzzy time series. Fuzzy Sets Syst. 123 (3), 387–394.

Huarng, K., Yu, T.H., 2006. Ratio-based lengths of intervals to improve fuzzy time series forecasting. IEEE Trans. Syst. Man Cybern. – Part B: Cyber 36, 328–340.

Kai, C., Fang-Ping, F., Wen-Gang, C., 2010. A novel forecasting model of fuzzy time series based on k-means clustering. IWETCS. IEEE 223–225.

Kaufman, L., Rousseeuw, P.J., 2009. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons,.

Koo, J.W., Wong, S.W., Selvachandran, G., Long, H.V., 2020. Prediction of air pollution Index in Kuala Lumpur using fuzzy time series and statistical models. Air Q. Atmosphere Health 1–12.

Mohamadi, S., Amindavar, H., & Hosseini, S.A.T., Arima-garch modeling for epileptic seizure prediction. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 994–998). IEEE. (2017).

Sadaei, H.J., Enayatifar, R., Abdullah, A.H., Gani, A., 2014. Short-term load forecasting using a hybrid model with a refined exponentially weighted fuzzy time series and an improved harmony search. Int. J. Electr. Power Energy Syst. 62, 118–129.

Severiano, C.A., Silva, P.C., Sadaei, H.J., Guimarães, F.G., Very short-term solar forecasting using fuzzy time series. In 2017 IEEE international conference on fuzzy systems (FUZZ-IEEE). 1–6 (2017).

Silva, P.C., Sadaei, H.J., Guimarães, F.G., 2017. Interval forecasting with fuzzy time series. Conf.: IEEE Symp. Ser. Comput. Intell.

Silva, P.C., Sadaei, H.J., Guimarães, F.G., 2019. Probabilistic forecasting with fuzzy time series. IEEE Trans. Fuzzy Syst. 99.

Singh, S., Parmar, K.S., Kumar, J., Makkhan, S.J.S., 2020. Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19. Chaos Solitons Fracta. 135, 109866.

Singh, P., 2018. Rainfall and financial forecasting using fuzzy time series and neural networks based model. Int. J. Mach. Learn. Cybern. 9 (3), 491–506.

Singh, P., 2021. FQTSFM: A fuzzy-quantum time series forecasting model. Inf. Sci. 566, 57–79.

Song, Q., Chissom, B.S., 1993. Forecasting enrollments with fuzzy time series-part I. Fuzzy Sets Syst. 54, 1–10.

Song, Q., Chissom, B.S., 1994. Forecasting enrollments with fuzzy time series-part II. Fuzzy Sets Syst. 62 (1), 1–8.

Van Tinh, N., Vu, V.V., Linh, T.T.N., 2016. A new method for forecasting enrolments combining time-variant fuzzy logical relationship groups and K-means clustering. Int. Res. J. Eng. Technol. 3 (3), 1–32.

Tsaur, R.C., 2012. A fuzzy time series-Markov chain model with an application to forecast the exchange rate between the Taiwan and US dollar. Int. J. Innov. Comput., Inf. Control 8 (7), 4931–4942.

Wang, H., Jiao, M., Tan, Y., 2016. Air quality index forecast based on fuzzy time series models. J. Resid. Sci. Technol. 13 (5), 12.

Wang, J., Li, H., Lu, H., 2018. Application of a novel early warning system based on fuzzy time series in urban air quality forecasting in China. Appl. Soft Comput. 71, 783–799.

Yang, H., Zhu, Z., Li, C., Li, R., 2019. A novel combined forecasting system for air pollutants concentration based on fuzzy theory and optimization of aggregation weight. Appl. Soft Comput. 105972.

Yan, Y., Li, Y., Sun, M., Wu, Z., 2019. Primary pollutants and air quality analysis for urban air in China: evidence from Shanghai. Sustainability 11 (8), 2319.

Yu, H.K., 2005. Weighted fuzzy time series models for TAIEX forecasting. Phys. A: Stat. Mech. Appl. 349 (34), 609–624.

Zadeh, L.A., 1965. Fuzzy sets. Inf. Control 8 (7), 338–353.

Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50, 159–175.

Zhang, Y., Qu, H., Wang, W., Zhao, J., 2020. A novel fuzzy time series forecasting model based on multiple linear regression and time series clustering. Math. Probl. Eng.

Zhang, Z., Zhu, Q., 2012. Fuzzy time series forecasting based on k-means clustering. Open J. Appl. Sci. 2, 100–103.