ssion_time_series_model_in_enrollment_university_forecasting.pdf

by

5th International Symposium on Computational and Business Intelligence

Fuzzy Random Auto-Regression Time Series Model in Enrollment University Forecasting

Riswan Efendi[1,2], Noor Azah Samsudin[1], Nureize Arbaiy[1], Mustafa Mat Deris[1]
[1]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia
Batu Pahat, Malaysia
[2]Mathematics Department
UIN Sultan Syarif Kasim
Pekanbaru, Indonesia
e-mail: riswan@uthm.edu.my; azah@uthm.edu.my, nureize@uthm.edu.my; mmustafa@uthm.edu.my

***Abstract*— the statistical models required the large data in the time series forecasting. While, to forecast the limited data or small data cannot be suggested by using these models. In this paper, we are interested to apply fuzzy random auto-regression model to handle the university enrollment data. The accuracy of the forecasting model can be improved through the left-right procedure. The yearly enrollment data of Alabama University are examined as benchmark data to evaluate the performance of proposed model. The results indicate that the smaller left-right spread of triangular fuzzy number produced the higher forecasting accuracy if compared with the existing models.**

***Keywords*-component; fuzzy random variable, enrollment, left-right procedure, auto-regression model**

I. INTRODUCTION

Forecasting of university enrollment is very essential activity in development of the higher education, such as providing information for decision making and budget planning. Many models have been presented to forecast the enrollment data. However, the forecasting accuracy is still issue and not easy task to obtain, as many factors have impacts to university enrollment. In fuzzy time series, many models have been presented to handle the enrollment trend using the classical approaches [1], [2], [3], [4] and the hybrid techniques [5], [6], [7]. On the other hand, the randomness, the vagueness and the possibility of input data are not yet considered in the models.

In this paper, we introduce a left-right procedure to model the randomness, the vagueness and the possibility of university enrollment of Alabama using fuzzy random auto-regression time series [8], [9], [10]. This procedure is very appropriate to handle the single point input of the time series data, such as, enrollment, production, visitor, and others. Because these data sets are measured using single value. The detail of theories, proposed procedure, and implementation are discussed in Sections II, III, and IV, respectively.

II. BASIC THERORIES OF FUZZY RANDOM MODEL

A. Fuzzy Random Variable (FRV)

Based on [5],[6], the fuzzy random variable Y, with possibility distribution μ_Y, the possibility, necessity, and credibility of event $\{Y \le r\}$ can be written mathematically as follows:

$$Pos\ \{Y \le r\} = sup\ \mu_Y(t),\ t \le r, \tag{1}$$

$$Nec\ \{Y \le r\} = 1 - sup\ (t),\ t \ge r, \tag{2}$$

$$Cr\ \{Y \le r\} = ½\ (1 + sup_{t\le r}\mu_Y(t) - sup_{t\ge r}\mu_Y(t)). \tag{3}$$

From Eq. (3), we note that the credibility measure is an average of the possibility and the necessity measures, i.e., $Cr\{.\} = \frac{Pos\{.\}+Nec\{.\}}{2}$. Credibility measure is to develop a certain measure, which is a sound aggregate of the two extreme cases. That is the possibility (which expresses a level of overlap and highly optimistic in this sense) and necessity (that articulates a degree of inclusion and is pessimistic in its nature). Based on credibility measure, the expected value of fuzzy variable is presented as follows [8].

Definition 1. Expected value of fuzzy variable [9]

Let Y be a fuzzy variable. The expected value of Y is defined as:

$$E(Y) = \int Cr\{Y \ge r\}\,dr - \int Cr\{Y \le r\}\,dr, \tag{4}$$

under the condition that the two integral are finite. Assume that $Y = [a^l, c, a^r]_T$ is triangular fuzzy variable or number (TFV = TFN). Making use of Eq. (4), we determine the expected value of Y to be

$$E(Y) = \frac{(a^l + 2c + a^r)}{4}. \tag{5}$$

Definition 2. Fuzzy random variable [9]

Suppose that (Ω, Σ, Pr) is a probability space and F_v is a collection of fuzzy variables defined on possibility space $(\Gamma, \mathcal{P}(\Gamma), Pos)$. A fuzzy random variable is a mapping $X : \Omega \to F_v$ such that for any Borel subset B of R, $Pos\{X(\omega) \in B\}$ is a measurable function of ω. Let X be a fuzzy random variable on Ω. From the previous definition, we know, for each $\omega \in \Omega$, that $X(\omega)$ is a fuzzy variable. Furthermore, a fuzzy random variable X is said to be positive if, for almost every ω, fuzzy variable $X(\omega)$ is positive almost surely. The expected value of the fuzzy variable $X(\omega)$ is denoted by $E(X(\omega))$, For any fuzzy random variable X on Ω, for each $\omega \in \Omega$. This has been proved to be a measurable function of ω [8], i.e., it is random variable. Hence, the expected value of the fuzzy random variable X is defined as the mathematical expectation of the random variable $E(X(\omega))$.

Definition 3. Expected value of fuzzy random variable [9] Let X be a fuzzy random variable defined on probability space (Ω, Σ, Pr). Then, the expected value of X and variance of X are defined as

$$E(X) = \int \Omega \left[\int Cr\{\xi(\omega) \geq r\}\, dr - \int Cr\{\xi(\omega) \leq r\}\, dr \right] Pr(\omega), \quad (6)$$

$$Var(X) = E(X - e)^2, \quad (7)$$

where $e = E(X)$ is given by Eq. (6).

B. *Fuzzy Random Auto-Regression Model (FR-AR)*

For time series data points, autoregressive or AR(p) model can be written as:

$$Y_t = \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + e_t, \quad (8)$$

where $\emptyset_1, \ldots, \emptyset_p$ are coefficients of $Y_{t-1}, \ldots, Y_{t-p}$, respectively, e_t is an error models at time-t. While, based on [9], the input and output data $\tilde{Y}_{t-p}$ for all $p = 0, 1, 2, \ldots, n$ are fuzzy random variables, which are written as:

$$\tilde{Y}_t = \cup_{i=1}^{n} \left[\left(Y_{it}^{l}, Y_{it}^{c}, Y_{it}^{r} \right)_{T}, P_{it} \right], \quad (9)$$

where $\tilde{Y}_t$ is a time series data at time-t and its formatted as a triangular fuzzy number [left, *l*; center, *c*; ; right, *r*]. Based on Eq. (9), all values given as fuzzy numbers with probabilities, P_{it}. These data, $\tilde{Y}_t$ also can be presented in Table I.

TABLE I. FUZZY RANDOM INPUT-OUTPUT DATA

Sample	Output	Input		
1	$\tilde{Y}_t$	$\tilde{Y}_{t-1}$	$\tilde{Y}_{t-2}$	$\cdots$ $\tilde{Y}_{t-k}$
2	$\tilde{Y}_{t-1}$	$\tilde{Y}_{t-2}$	$\tilde{Y}_{t-3}$	$\cdots$ $\tilde{Y}_{t-(k+1)}$
…	…	…	…	…
n	$\tilde{Y}_{t-n}$	$\tilde{Y}_{t-(n+1)}$	$\tilde{Y}_{t-(n+2)}$	$\cdots$ $\tilde{Y}_{t-(k+n)}$

III. DATA PREPARATION

The single point data has been frequently implemented by previous fuzzy and non-fuzzy models to forecast the time series data. Apparently, the existing fuzzy forecasting models have not considered three major issues in data representation: randomness, vagueness, and possibility. For example, different financial analyst may arrive to different conclusion when observing trend of stock prices within certain period of time. Possibly, some analysts will evaluate the stock prices performance according their experience and expertise. As a result, there are various possibilities in the evaluation, e.g., "low, medium, high, and very high", or other prices. The variation in the performance evaluation is considerably stochastic in nature. However, representing the performance evaluation using single value is inappropriate. Therefore, the price evaluation is more reasonable to be presented in fuzzy form [5]. This preparation procedure is presented in Section III*A*.

A. *Left-Right Procedure*

The explanation of left-right (LRS) procedure in [9], [10] is not clearly enough. Although, the LRS procedure is commonly used in dealing with single data input, yet the rational of choosing inconsistent left-right spread values are not well discussed. On the other hand, we are concerned with consistent left-right spread (CLRS) procedure in adjusting single input time series data to build desired FR-AR model.

Let Y_t be time series at $t = 1, 2, \ldots, n$ and let k be possible spread values, $k = 1, 2, \ldots, n$. Y_t can be transformed into TFN format using k values. Note that, although various k values are allowed to adjust the CLRS of TFN, but at any Y_t a consistent k value must be used. As a result, the transformation can be written in mathematical form such as in Eq. (10).

$$Y_t \rightarrow [Y_t - k, Y_t, Y_t + k]_T \rightarrow \left[Y_t^{l}, Y_t^{c}, Y_t^{r} \right]_T \rightarrow \langle Y_t^{c}, k \rangle = \tilde{Y}_t, \quad (10)$$

Which $Y_t^{l}, Y_t^{c}, Y_t^{r}$ are results of transformation of Y_t using LRS procedure. Figure 1 shows an example of Y_t being transformed into TFN at $t = 3$ using $k = 1, 2, \ldots, n$.

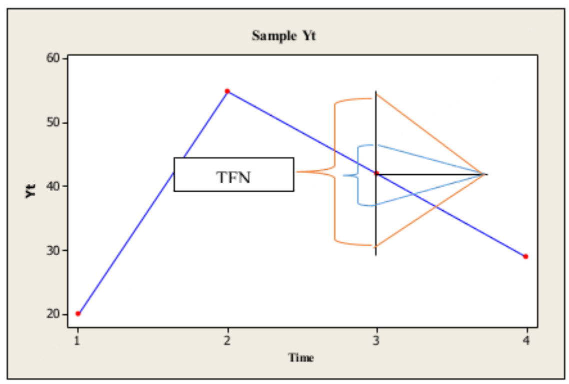

Figure 1. An example of transformation data into TFN

It is important to choose a k value that has potential to reduce the forecasting error of FR-AR model. Different k values will lead to different TFN shapes. As a result, the TFN shapes will influence the expected value and standard deviation of fuzzy random data (FRD). Essentially, the expected value and standard deviation will influence the parameter estimation FR-AR model. Figure 2 shows the simulation results with four different time series data using $k = 1, 2, \ldots, 10$. Based on data size, there are various initial values of k being used. In data set 1, initial value of k is 5, in data set 2, initial value of k is 3, in data set 3, initial value of k is 2 and in data set 4, initial value of k is 1. Note that, for each data set in Figure 2 three k values are plotted to highlight differences in MSE. Apparently, the smaller k is able to produce smaller mean square error (MSE). In other words, there is a significant relationship between spread value of k and MSE.

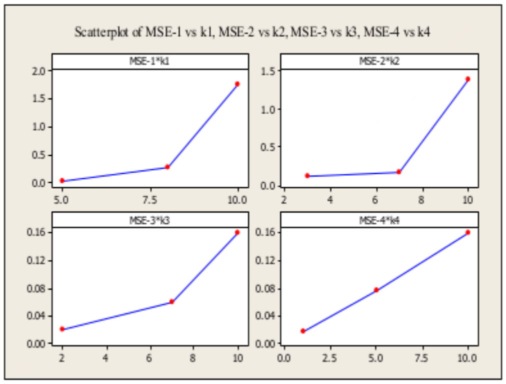

Figure 2. The simulation results to investigate the influence of k values to MSE

B. Building FR-AR Model Using Algorithm

In this section, we will present the algorithm details in building FR-AR model [9], [10]. Indeed, the transformation step for the type of data presented in Sections A will be included. Additionally, the parameter estimate using linear programming (LP) approach will be discussed too. The details of the algorithm are presented in Steps 1-7 as follows:

Step 1: Transform the actual time series into TFN based on the procedures in Section A.

Step 2: Determine fuzzy data (FD) using the TFN results in Step 1.

Step 3: Divide FD into two groups of fuzzy random data, FRD_1 and FRD_2.

The length (l) for each FD can be defined as the difference between upper (b_i) and lower data (a_i), such that $i = 1, 2, 3, \ldots, n$. By using value of l, the fuzzy random time series data can be divided into FRD_1 and FRD_2.

Step 4: Calculate the expected value $E(Y)$ and $Var\ (Y)$ of FRD.

Before building FR-AR with parameter estimates, a confidence-interval that is induced by $E(Y)$ and $Std.Dev(Y)$ of a fuzzy random variable (FRD), then we consider the one-sigma confidence $(1 \times \sigma)$ interval (CI) of each FRD [8]. The CI of FRD can be expressed as:

$$CI \cong [E(Y) - Std.Dev\ (Y), E(Y) + Std.Dev\ (Y)].$$

Note that, the CI being adopted in this step is different from the one that we used to estimate sample mean. The corresponding CI (output) results for each FRD_1 (input) and FRD_2 (input).

Step 5: The general FR-AR model can be expressed as:

$$Y_t = [\widehat{\emptyset}_1^l, \widehat{\emptyset}_1^r]\tilde{Y}_{t-1} + [\widehat{\emptyset}_2^l, \widehat{\emptyset}_2^r]\tilde{Y}_{t-2}, \quad (11)$$

Note that, a linear equation system is to be developed using the input (FRD_1, FRD_2) and output (FRD). There are two linear equation system to be developed for left-right inputs. Then, both linear equation systems will be solved using linear programing (LP), namely, simplex approach to estimate the parameters of FR-AR model.

Step 6: Determine the predicted FR-AR model from Step 5.

$$\hat{Y}_t = [\widehat{\emptyset}_1^l, \widehat{\emptyset}_1^r]\tilde{Y}_{t-1} + [\widehat{\emptyset}_2^l, \widehat{\emptyset}_2^r]\tilde{Y}_{t-2}, \quad (12)$$

which $\hat{Y}_t$ is a predicted fuzzy random time series at time-t, $\widehat{\emptyset}_1^l, \widehat{\emptyset}_1^r$ and $\widehat{\emptyset}_2^l, \widehat{\emptyset}_2^r$ are pair predicted parameters of fuzzy random time series $(\tilde{Y}_{t-1}, \tilde{Y}_{t-2})$, respectively. The uniqueness of the estimated model parameters is that left-right values are always the same. Therefore, interestingly Eq. (12) is capable to forecast three different values, low, medium and high as follows:

$$\hat{Y}_t = \begin{cases} [\widehat{\emptyset}_1^l]\tilde{Y}_{t-1} + [\widehat{\emptyset}_2^l]\tilde{Y}_{t-2}, & \text{forecast of low data} \\ [\widehat{\emptyset}_1^r]\tilde{Y}_{t-1} + [\widehat{\emptyset}_2^r]\tilde{Y}_{t-2}, & \text{forecast of high data} \\ \frac{\hat{Y}_t^l + \hat{Y}_t^r}{2}, & \text{forecast of medium data} \end{cases}, \quad (13)$$

Unlike existing models which are limited to forecast single value only.

IV. Implementation

In this section, we present the implementation of proposed algorithm to forecast the yearly enrollment of Alabama University from 1972 to 1992. This data set is a benchmark and frequently used in fuzzy time series forecasting. By following steps in Section III B, FR-AR model of enrolment can be written as:

$$\hat{Y}_t = [0.94, 1.076](Y_{t-1})_T. \quad (14)$$

Based on Eq. (14), we eventually obtained same parameter estimates for left and right values ($\widehat{\emptyset}_1^l = 0.94$ and $\widehat{\emptyset}_1^r = 1.076$) for each model. Then, the root of mean square error (RMSE) of the models are compared with the existing fuzzy time series (FTS) models in Table II.

TABLE II. RMSE Comparison

Model	RMSE
Song &Chissom [1]	650.4
Chen [3]	638.3
Huarng [7]	476.9
Lee & Chou [7]	501.2
Cheng et al. [7]	478.4
Yolcu et al. [7]	805.1
Qiu et al. [7]	511.3
Joshi & Kumar [7]	433.7
Ismail et al. [6]	400.2
Kumar & Gangwar [7]	493.5
Bisht & Kumar [7]	428.5
Proposed FR-AR	149.2**

Due to the limitation FTS models, we use average values only for comparison. Recall that, the data are divided into training and testing data. The RMSE for the FR-AR models are smaller than FTS models. The RMSE values of proposed model imply that the forecasting error can be reduced significantly. In other words, the FR-AR model benefits from the use of left-right values in proposed procedure becuase the estimated parameters can reduce the randomness, vagueness and possibility of data. Thus, the variation will be minimise too and can be controled. As a result, the predicted values tend to be near to the medium data. Thus the accuracy is improved. Moreover, the actual enrollment and the forecasted values which derived by proposed FR-AR and some FTS models are also illustrated in Figure 3.

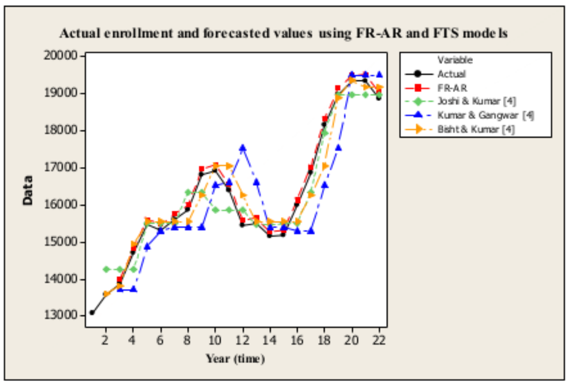

Figure 3. Actual enrollment and forecasted values using some models

V. CONCLUSION

In this paper, we build an improved FR-AR model by employing left-right procedure in data preparation. While this procedure is more appropriate for single point data. Additionally, most existing statistical models require differencing technique to ignore seasonal and trend components. In contradiction, such differencing technique is not required in the proposed procedure. Moreover, the smaller left-right spread of TFN is suggested to achieve the better forecasted values. In the comparison of RMSE, we considered fuzzy models, namely, FTS. Based on application, our proposed FR-AR model with new input type outperforms other both existing models. Based on our perspective, this proposed model can be implemented in handling the out-sample forecast as compared with most existing fuzzy time series models. Besides that, our proposed model is also able to produce the different forecasted values by time generation.

ACKNOWLEDGMENT

This study is supported by Ministry of Higher Education, Malaysia. Fundamental Research Grant Scheme, Vot. 1609 and in part by a great from Research Gates IT Solution Sdn. Bhd.

REFERENCES

[1] Q. Song and B. S. Chissom, “Forecasting enrollment with fuzzy time series Part-I,” Fuzzy Sets and Systems, vol. 54, pp. 1-9, 1993.

[2] Q. Song and B. S. Chissom, “Forecasting enrollment with fuzzy time series Part-II,” Fuzzy Sets and Systems, vol. 64, pp. 1-8, 1994.

[3] S. M. Chen, “Forecasting enrollment based on fuzzy time series,” Fuzzy Sets and Systems, vol. 81, pp. 311-319, 1996.

[4] S. M. Chen and N. Y. Chung “Forecasting enrollment using high-order fuzzy time series and genetic algorithm,” International Journal of Intelligent Systems, vol. 21, pp. 485-501, 2006.

[5] U. Yolcu, E. Egrioglu, V. R. Uslu, M. A. Basaran and C. H. Baladag, “A new approach for determining the length of intervals for fuzzy time series,” Applied Soft Computing, vol. 9, pp. 647-651, 2009.

[6] Z. Ismail, R. Efendi and M. M. Deris,”Inter-quartile range approach to length-interval adjustment of enrollment data in fuzzy time series forecasting,” International Joournal of Computational Intelligence and Applications, vol. 12, pp. 48-58, 2013.

[7] K. Bisht and S. Kumar, “Fuzzy time series forecasting method based on hesitant fuzzy sets,” Expert Systems with Applications, vol. 64 (c) pp. 577 – 568, 2016

[8] J. Watada, S. Wang and W. Pedrycz, “Building confidence-interval based fuzzy random regression model,” IEEE Trans. Fuzzy Syst, vol. 17, pp. 1273-1283, 2009.

[9] L. Shao, Y-H. Tsai, J. Watada and S. Wang,”Building fuzzy random auto-regression model and its application, Intel. Decision Tech. (SIST), vol. 14, pp. 24-30, 2012.

[10] R. Efendi, N. Arbaiy and M. M. Deris,”Estimation of confidence-interval for yearly electricity load consumption based on fuzzy random autor-regression,” Advances Intel. Syst. Compt. (CIIS), vol. 1, pp. 61-68, 2017.

ssion_time_series_model_in_enrollment_university_forecasti...

ORIGINALITY REPORT

17%	%	17%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Computational Intelligence in Information Systems", Springer Nature, 2017 Publication	12%
2	Lu Shao, You-Hsi Tsai, Junzo Watada, Shuming Wang. "Chapter 16 Building Fuzzy Random Autoregression Model and Its Application", Springer Science and Business Media LLC, 2012 Publication	4%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		