© 2018 IJSRSET | Volume 4 | Issue 11 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET1184112

A Comparative Simulation Study of ARIMA and Fuzzy Time Series Model for Forecasting Time Series Data

Haji A. Haji, Kusman Sadik, Agus Mohamad Soleh
Department of Statistics, Bogor Agricultural University, IPB Bogor, 16680, Indonesia

ABSTRACT

Simulation study is used when real world data is hard to find or time consuming to gather and it involves generating data set by specific statistical model or using random sampling. A simulation of the process is useful to test theories and understand behavior of the statistical methods. This study aimed to compare ARIMA and Fuzzy Time Series (FTS) model in order to identify the best model for forecasting time series data based on 100 replicates on 100 generated data of the ARIMA (1,0,1) model.There are 16 scenarios used in this study as a combination between 4 data generation variance error values (0.5, 1, 3,5) with 4 ARMA(1,1) parameter values. Furthermore, The performances were evaluated based on three metric mean absolute percentage error (MAPE),Root mean squared error (RMSE) and Bias statistics criterion to determine the more appropriate method and performance of model. The results of the study show a lowest bias for the chen fuzzy time series model and the performance of all measurements is small then other models. The results also proved that chen method is compatible with the advanced forecasting techniques in all of the consided situation in providing better forecasting accuracy.

Keywords: ARIMA, Bias, Chen, Fuzzy Time Series, MAPE, RMSE, Simulation,Yu FTS

I. INTRODUCTION

The development of a forecasting method is quite rapidly resulting, there are many choices of methods that can be used to forecast time series data according to needs of the users. So it is necessary to compare one method with another method to get the best forecast results with high accuracy. In 1970, the idea of forecasting future events of a time series as a combination of its past values received a strong impulse after Box & Jenkins (1970). In that work, Box & Jenkins proposed a modeling cycle for the autoregressive (AR) model, which assumes that future values of a time series can be expressed as a linear combination of its past value. Of course this linearity assumption implies certain limitations, and in the last years much research has been devoted to nonlinear models. Nonlinear and non-stationary models are more flexible in capturing the characteristics of data and, in some cases, are better in terms of estimation and forecasting. These advances do not rule out linear models at all, because these models are a first approach which can be of great help to further estimate some of the parameters. Furthermore, modeling of any real-world problem by using nonlinear models must start by evaluating if the behavior of the series follows a linear or nonlinear pattern.

Box-Jenkins method or often also called ARIMA is a method that is intensively developed and studied by Statisticians Box and Jenkins; therefore their names are frequent associated with the ARIMA process applied for data analysis and data forecasting time

series. ARIMA is actually an attempt to search for the most data patterns matches from a bunch of data, so the ARIMA method entails completely historical data and current data to produce short-term forecasts (Sugiarto and Harijono, 2000).

The time series forecast has been a widely used forecasting method. Although time series forecast can deal with many forecasting problems, it cannot solve forecasting problems in which the historical data are vague, imprecise, or are in linguistic terms. To address this problem, Song and Chissom (1993a, b, 1994) presented the definitions of fuzzy time series by using fuzzy relational equations and approximate reasoning. Since then, a number of researchers have built on their research and developed different fuzzy forecasting methods (Chen (1996, 2002); Hwang, Chen & Lee (1998); Chen & Hwang (2000); Huarng (2001a,b); Lee & Chou (2004)).

The proposed method by Chen (1996) uses simplified arithmetic operations rather than those complicated maximum composition operations in Song and Chissom (1993a). In Chen's method, the variation of enrollment of this year is related to the trend of past years'. To define the degrees of variations, he performs systematic calculations to obtain the relation between the variation of last year and other previous years. Then, he can get the forecasting enrollments from the derived relation. Huarng (2001a, 2001b) proposes Heuristic models by integrating problem-specific heuristic knowledge with Chen's model to improve forecasting, since Chen's model is easy to calculate, straightforward to integrate heuristic knowledge, and could forecast better than others. Lee et al. (2006), based on the two-factor high-order fuzzy time series and historical data, proposed two-factor high-order fuzzy logical relationships to increase the forecasting accuracy rate.

This method used to compute forecasting and applied in time series data. The main aim of the fuzzy time series is to predict the value of time series data which can be widely used in any real time data (Hansun 2012). The fuzzy time series process uses a linguistic variable whose value linguistic is a fuzzy set. The concept of the fuzzy time series method has been many developed in some forecasting problems, one of them by Saxena et al. (2012) which presents the fuzzy time series method in modeling the data by using the percentage change of data each year based on the number of frequencies. In predicting a data we are often confronted with elections which method best suits the data we are going to forecast

This Simulation study is used to obtain empirical results about the performance of statistical methods incertain scenarios,as opposed to more general analytic results,which may cover many scenarios.It is not always possible,or may be difficult, to obtain analytic results. Simulation study come into their own when methods make wrong assumptions because they can assess the resilience of methods in such situations.This is not always possible with analytic results, where results may apply only when data arise from aspecific model.

\

This paper was conducted a simulation study of ARIMA generated data by ARMA (1,1) model and fuzzy time series model in order to know the characteristics and performancebetween two models and between chen and Yu FTS. Furthermore, this paper intents to identify which model is able to provide better information in decision making, especially when the data is insufficient. The results could improve the understandings of how much time period is included in data to affect forecasting performance, by analyzing and comparing Chen (1996) and Yu (2005) model with ARIMA (1,0,1) model.

This simulation study is a part of comparative study of ARIMA and fuzzy time series model for forcasting rainfall data a case of Bogor city-Indonesia.

II. LITERATURE REVIEWS

A. Simulation of Data Generation

In this simulation study time series is based on the non-seasonal ARIMA model, ARIMA (p, d, and q) where (p, d, and q) represent the non-seasonal part of the model. This model incorporates characteristics of interest: increasing trend i.e. non-stationary in the mean, and the presence of seasonal variation. We have set the length of the data (n=100) for the simulated time series.

We use different Scenarios with combination of 4 ARMA (1,1) parameter values , namely a) $\phi = 0.5$ and $\theta = 0.6$, b) $\phi = 0.9$ and $\theta = 0.6$, c) $\phi = 0.1$ and $\theta = 0.9$, d) $\phi = 0.2$ and $\theta = 0.9$. While variance to generate errors also uses 4 values, namely 0.5,1,3 and 5. as many as 100 data series will be raised each with 100 repititions.

B. ARIMA model

ARIMA, also known as Box–Jenkins models (Box and Jenkins 1976), has been a very popular type of time series forecast models in different fields. It has the function of transforming a non-stationary time series into a stationary time series, The synthetic data is the ARMA (1,1) model with a non-seasonal AR term and a non-seasonal MA term, , used in this study.

$$Z_t = \phi Z_{t-1} + e_t - \theta e_{t-1} \qquad (1)$$

where Z_t is a time series data and ϕ is AR parameter, θ is MA parameter and e_t are random variables with $\mu = 0$ and $\sigma 2 = 1$. Box and Jenkins proposed four primary stages in building ARIMA model, known as Box-Jenkins procedure,

a. Identification of the model:
   By starting with an examine the stationary assumption with time series plot, Autocorrelation Funcion (ACF), Partial Autocorrelation Funcion (PACF), and augmented Dickey-Fuller (ADF) test or Box-Cox transformation. Next to determine tentative models based on sample data to identify p, d, and q values after the time series data to be stationary.
b. Estimation of parameters
   After getting appropraite value of p,q,d,P,Q and D, the next stage is to find the values of c, θ,φ,Θ,andΦ. There are several different methods that are used to estimate the parameters. All of them should produce very similar estimates, but may be more or less efficient for any given model. Most of them are Moment, Least square and Maximum likelihood method (Melard , 1984).
c. Diagnosticc model
   Diagnostics test is applied to understand whether the estimated parameters and residuals of the fitted ARIMA model are significant, checks the residual assumption using Ljung-Box test. The hypotheses being tested is:
   H_0: Uncorrelated residuals
   H_1: correlated residuals.

The statistical test is done by calculating the value of Q^*as follows:

$$\boldsymbol{Q^* = T(T+2)\sum_{t=1}^{T}\left(\frac{\hat{r}_{k)}^2}{T-k}\right)} \qquad (2)$$

With
$\hat{r}_k^2$: residual correlation at k-lag
k : number of lags being tested
T : length training data set

The decision is to reject H_0 if $Q^* > \mathcal{X}_\alpha^2$ $(df = T - p - q)$ or by checking p-value < α. If the decision is accepted, it can be said that the ARIMA model used is feasible for forecasting.

d. Fitting and Prediction of ARIMA model

Once a model has been identified and all the parameters have been estimated, we can predict future values of a time series with this model.Evaluation of forecasting results is determined by the value of MSE and MAPE.

C. Chen (1996) fuzzy time series

In general Chen (1996) improved the approach proposed by Song and Chissom (1993a; 1993b). Chen's method uses a simple operation, instead of complex matrix operations, in the establishment step of fuzzy relationships. The algorithm of Chen's method can be given as follows:

1) Define the universe of discourse and intervals for the rules of abstraction. Based on the issue domain, the universe of discourse can be defined as: *U*= [stating, ending]. As the lenght of interval is determined U can be partitioned into several equally length intervals.
2) Define fuzzy sets based on the universe of discourse and fuzzified the historia data
3) Fuzzify observed rules.
4) Establish Fuzzy Logical Relationships (FLRs) and group (FLRG) them based on the current states of the data of the fuzzy logic relationships.
6) Forecast. Let F(t -1) = Ai .
7) Defuzzify. If the forecast of F(t) is Aj1 ,Aj2 ,…,Ajk, the defuzzified result is equal to the arithmetic average of the midpoints of Aj1,Aj2,...Ajk

E. Yu (2005) fuzzy time series (Yu FTS)

The steps of the algorithm of the weighted method proposed by Yu (2005) can be given below:

1) Define the discourse of universe and subintervals. Based on min and max values in the data set, Dmin and Dmax variables are defined., then choose two arbitrary positive numbers which are D1 and D2 in order to divide the interval evenly,

$$U = [Dmin - D1 , Dmax - D2] .$$

2) Define fuzzy sets based on the universe of discourse and fuzzify the historical data.
3) Fuzzify observed rules.
4) Establish fuzzy logical relationships (revised Chen's method).
5) Forecast: Use the same rule as Chen's.
6) Defuzzify: Suppose the forecast of F(t) is Aj1 ,Aj2 ,…,Ajk. The defuzzified matrix is equal to a matrix of the midpoints of Aj1 ,Aj2 ,…,Ajk :

$$M(t) = [mj1,mj2,\ldots,mjk]$$

where, M(t) represents the defuzzified forecast of F(t).

7) Assigning weights. Suppose the forecast of F(t) is Aj1 ,Aj2 ,…,Ajk. The corresponding weights for Aj1 ,Aj2 …,Ajk , say w¢1,w¢2 ,…,w¢k are specified as:

$$w' = \frac{w_i}{\sum_{h=1}^{k} w_h} \quad (3)$$

where, $w_1 = 1, w_i = w_{i-1} + 1$ for $2 \le i \le k$. We then obtain the weight matrix as:

$$W(t) = [w ,w ,\ldots,w] = \left[\frac{1}{\sum_{h=1}^{k} w_h}, \frac{2}{\sum_{h=1}^{k} w_h}, \ldots, \frac{k}{\sum_{h=1}^{k} w_h}\right] \quad (4)$$

where, wh is the corresponding weight for Ajk

8) Calculating the final forecast values. In the weighted model, the final forecast is equal to the product of the defuzzified matrix and the transpose of the weight matrix:

$$\hat{F}(t) = M(t) * w(t)^T = [m_{j1}, m_{j2}, \ldots, m_{jk}] * \left[\frac{1}{\sum_{h=1}^{k} w_h}, \frac{2}{\sum_{h=1}^{k} w_j}, \ldots \frac{k}{\sum_{h=1}^{k} w_h}\right]^T \quad (5)$$

Where:

* : Matrix product operator
M(t) : 1xk matrix
W(t)T : k x1 matrix, respectively

III. METHODS AND MATERIAL

A. Material

One sample size, long term with size N=100 are used and data are generated from parametric non-stationary ARMA (1,1) model with 8 different

combination of parameters and 4 different of variance errors from generated data set.

B. Method of Analysis

We generated a time series for each model of length 100 for different orders of p and q with all possibilities $p + q \leq 5$ with stationarity data ($d = 0$), We generate the four non-stationary time series models and select the most appropriate models with the smallest mean absolute percentage error and the smallest average root mean square error. The following steps are used to generate time series data in this simulation study.

1. General steps of generating data from an ARMA (1,1) are as follows:

a. Generate e~$N(0, \delta_e^2)$ with n=100, and variance of error specific to each scenario. The parameter μ=15 is used to generate an ARMA (1,1) with non-zero means.

b. Adding the generated error value to the ARMA (1,1) model formula in order to create time series data with an ARMA (1,1) each scenario of $\phi 1$ model θ1 parameters. This process is repeated for 100 replication.

c. Splitting 100 simulated data into two data sets as training data set (88 sample data) and testing data set (last 12 sample data).

Comparing and estimating the results using ARIMA model and combination of Chen and Yu fuzzy time series model.

IV. RESULTS AND DISCUSSION

A. Exploration and visualisation of result

We first explore the results. Figure 1 plots the first replicate of generated data set with δ_e^2 = 0.5 and 4 different ARMA(1,1) parameters , with means displayed with different colors, the generated data with parameter ϕ = 0.9 and θ=0.6 shows a much longer range of fluctuation compared with other parameters.

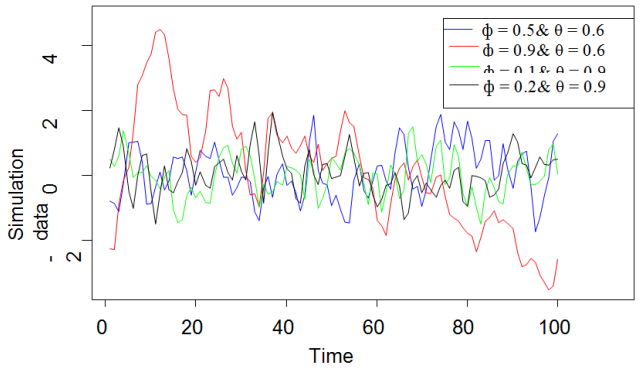

Figure 1. First replicate plot of generated data with a) ϕ=0.5 and θ=0.6 b) ϕ =0.9 and θ=0.6 c) ϕ =0.1 and θ=0.9 d) ϕ =0.2 and θ=0.9 and δ_e^2= 0.5.

The first illustration of generated data with ARMA (1.1) model parameter values of ϕ = 0.5 and θ = 0.6 and 4 different variance error rates as shown in Figure 2 above the data generated by δ_e^2=5 having a data deviation that tends to be larger than others. whereas for data generated from $\delta_e^2 = 0.5$ *and* $\delta_e^2 = 1$, although some data show large data deviations but overall still smaller than the deviation of $\delta_e^2 = 5$

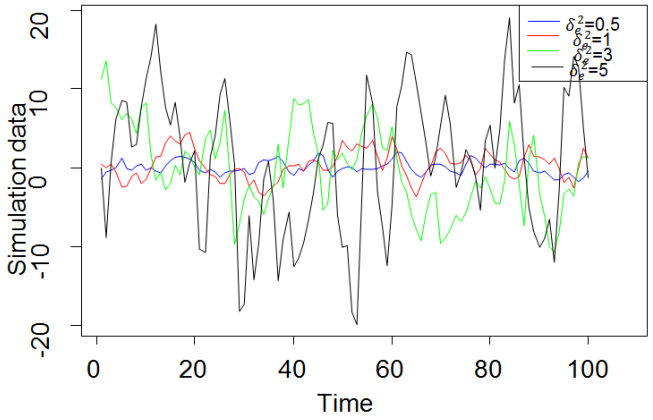

Figure 2. First re-illustration of generation results with ϕ = 0.5 and θ = 0.6 with 4 various variance errors values.

B. Comparison of Accuracy measures of Simulation testing data set with ARIMA and fuzzy time series model

The results shows the comparison between the proposed models versus classical models for long terms based on selected criterion of forecasting accuracy for simulated models. The distribution of different forecasting meausres Bias,RMSE and MAPE are estimated. The results show that the Chen model is preferable in selecting the most appropriate forecasting model over all the other models for long terms beacause both forecasting meausres has smallest values then the other models. In addition, the ARIMA model performs better than the other model in term of Bias. Furthermore, the Bias measures for Yu model for smallest variance error equal 0.05. This result indicates that Yu model is more efficient than the other models for long terms and for all parameters.

Figures 3 show the numerical results of the RMSE for the simulation testing data set for the variance error and 4 parameters by the ARIMA and fuzzy time series.the numerical results under the three interest time series models mentioned above are highly consistent (the values almost fall on the same line but the chen model is equivalent to the deterministic model in this case). For the Chen model showed that best model to be used for forecast because it has lowest RMSE value compared to other model in each variance error term being selected.From figure 4 above it can be seen that the best method to forecast the time series data is the Chen fuzzy time series model, because both the RMSE and MAPE errors values for this method has lower value compared to other methods and also contained Bias statistics with higher values than Bias statistics of other methods.

The bias of the eight parameters with four difference variance errors is presented in Fig.5. The results shows that the conditions for $\sigma_e^2 = 0.5$ are very small actually are similar for all models. As shown in the figure, the bias becomes negative as `variance error increases for ARIMA model especially when ϕ =0.1 and θ=0.9 for both $\sigma_e^2 = 3$ and $\sigma_e^2 = 5$, which is to be expected. But for Yu model has high bias for that condition.The relationship between the bias and other forecasting accuracy measures is roughly linear for all methods. Furthermore, The largest bias for $\sigma_e^2 = 5$ is associated with the Yu fuzzy time series model for positive values followed by Chen model and for ARIMA get the negative value of bias. With regard to the ordering of the forecasting methods, differences are found between simulation values of and between the time series data set.

For small values of variance error /, the smallest bias is shown by both methods.

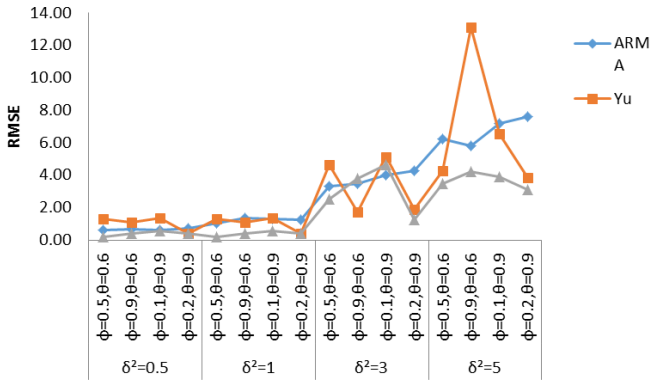

Figure 3.RMSE for simulation testing data for varinace error and 4 parameters by forecasting method

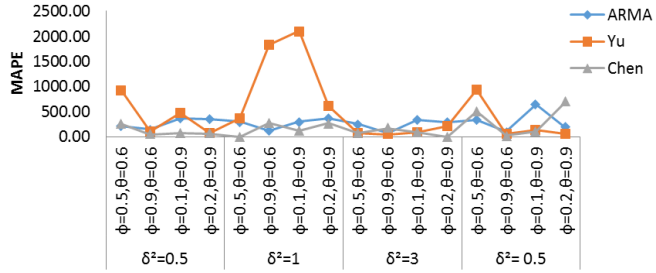

Figure 4.MAPE for simulation testing data for variance error and 4 parameters by forecasting method

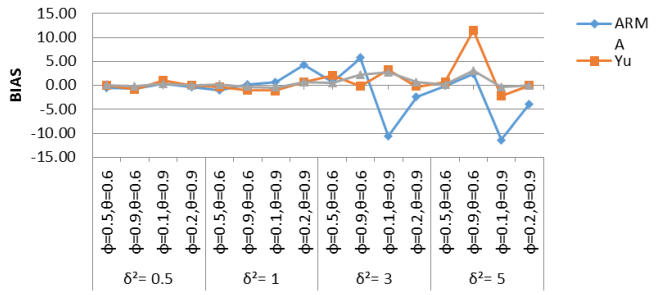

Figure 5. BIAS for simulation testing data for variance error and 4 parameters by forecasting method

V. CONCLUSION

In this paper, ARMA and fuzzy time series approaches are applied to the simulated data generated from ARIMA (1,0,1) models and compared with fuzzy time series model of Chen and Yu. Then performance of these fitted models are assessed by means of measurements.The results show that chen model has an acceptable performance for modeling and forecasting of time series data in all of the considered situations. It is an interesting result because of data sets are generated from the specified ARIMA model, also Chen model has a good fitness for this simulattion study. So that an application of Chen FTS is suggested because of precision of the method.

Furthermore the results showed that RMSE values for both models showed a much stable variance for $\sigma_e^2 = 0.5$ and $\sigma_e^2 = 5$.The variance of RMSE from Yu FTS is in a range of 0.41 – 13.14 and variance of RMSE from Chen FTS is in a range of 0.18 – 4.65. also the variance of RMSE with ARIMA increased from 0.634 to 7.633.

Despite ARMA and Yu FTS method seemed look better than Chen FTS in term of Bias, the more increased in ARMA(1,1) parameter showed Bias on Yu method is better than ARMA even with a closer result with Chen FTS.

VI. AACKNOWLEDGEMENT

The author would like to express his deepest gratitude to the Bogor Agricultural University Indonesia,for providing the opportunity to study especially Department of Statistics and the Ministry of Education Directorate General of Higher Education *Indonesian* "*DIKTI*" for funding this research.

VII. REFERENCES

[1]. Bisht, D.C.S., M.M. Raju, M.C. Joshi. 2009. Simulation of Water Table Elevation Fluctuation using FuzzyLogic and ANFIS. Computer Modelling and New Technologies, Vol.13 (2): 1623

[2]. Box,G. & G.Jenkins, (1976).Time series analysis: forecasting and control. (Revised ed) Holden day, San Francisco.

[3]. Chen, S.M. 1996. Forecasting enrolments based on fuzzy time series. Fuzzy Sets and Systems 81(3): 311-319.

[4]. Cheng,C.H., Chen, T.L., Teoh, H.J. & Chiang, C.H. 2008. Fuzzy time-series based on adaptive expectation model for TAIEX forecasting. Expert Systems with Applications 34(2): 1126-1132.

[5]. Jilani T.A. and Ardil. C., "Fuzzy metric approach for fuzzy time series forecasting based on frequency density based partitioning", in proceedings of world academy of science, engineering

[6]. L.A. Zadeh, Fuzzy set, Fuzzy Set Information and control 8 (1965)338-353 Management, University of Diponegoro, Semarang, Indonesia, 2008. MS Lutkepohl, Helmut, New Introduction to Multiple Time Series Analysis, Berlin:Springer Science + Business Media, Inc., 2005.

[7]. Lee, L.W., Wang, L.H. and Chen, S.M., 2007, Temperature prediction and TAIFEX forecasting based on fuzzy logical relationships and genetic algorithms, Expert Systems with Applications, 33, p. 539-550.

[8]. Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998). Forecasting Methods and Applications, 3rd Edition, John Wiley, New York

[9]. Salehfar, H, N. Bengiamin, and J. Huang. 2000. A Systematic approach to linguistic fuzzy modeling based on input-output data. Proceedings of the 2000 Winter Simulation Conference. J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, Eds., University of North Dakota, U.S.A.

[10]. S.M. Chen, Forecasting enrolments based on higher-order fuzzy time series, Cybernetics and systems. An international Journal 33 (2002) 1-16

[11]. S.R. Singh, A Simple method of forecasting based on fuzzy time series, applied mathematics and computation 186 (2007) 330-339

[12]. Song, Q. & Chissom, B.S.1993a. Forecasting enrolments with fuzzy time series - Part I. Fuzzy Sets and Systems 54(1): 1-9.

[13]. Song, Q. & Chissom, B.S. 1993b. Fuzzy time series and its models. Fuzzy Sets and Systems 54(3): 269-277. .

[14]. Wang, C.H. and Hsu, L.C., 2008, Constructing and applying an improved fuzzy time series model: Taking the tourism industry for example, Expert Systems with Applications, 34, p. 2732-2738.

[15]. Yu, H.K.2005.Weighted fuzzy time series models for TAIEX forecasting. Physical A: Statistical Mechanics and Its Applications 349(3-4): 609-624.

[16]. Zadeh, L. A. (1996). "Fuzzy Logic: Computing with Words", IEEE Transactions on Fuzzy Systems, 4, 103-111

[17]. Zhang, W.X. & Li, T. & Ma J.F. & Li, A.J. (1990). Set-valued measure and fuzz valued measure, Fuzzy Sets and Systems 36,181-188