Hybrid ARIMA-ANFIS for Rainfall Prediction in Indonesia

Ria Faulina[1], Suhartono[2]

[1]Department of Statistics Institut Teknologi Sepuluh Nopember, Surabaya-Indonesia
rifamoslem@gmail.com

[2]Department of Statistics Institut Teknologi Sepuluh Nopember, Surabaya-Indonesia
gmsuhartono@gmail.com m

Abstract: ***In Indonesia, rainfall prediction is very important especially for food production. The M3-competition shows that more complicated model not always yield better forecast than simpler one. Conversely, from this competition there is a statement shows that when the various methods are being combined, the accuracy is better than the individual method. This paper proposed hybrid and ensemble model of forecasting method for ten-daily rainfall prediction based on ARIMA (Autoregressive Integrated Moving Average) and ANFIS (Adaptive Neuro Fuzzy Inference System) at six certain area in Indonesia. To find an ensemble forecast from ARIMA and ANFIS models, the averaging and stacking method was implemented. In this study, Triangular, Gaussian, and Gbell function are used as membership function in ANFIS. The best model is measured by the smallest root of mean square errors (RMSE) at testing datasets. The results show that an individual ARIMA method yields more accurate forecast in five rainfall data, whereas ensemble averaging multi model yields better forecast in one rainfall data. In general, these results in line with M3 competition results that more complicated model not always yield better forecast than simpler one.***

Keywords: ARIMA, ANFIS, Hybrid, Ensemble.

1. Introduction

Rainfall is one of factor that effects the food production in Indonesia. ARIMA is a traditional method that still used in prediction techniques, especially in climate prediction. However, there are several problems frequently arise in ARIMA modeling, i.e. the stationary of the data is often not met and determination of the order of [p,d,q] is often difficult. Additionally, computational intelligence methods were also used for rainfall prediction. These methods have no assumption such classical methods. ANFIS is a combination between artificial neural networks and fuzzy inference system [1].

In Indonesia, statistical models which are currently developed and applied for rainfall prediction have not given adequate results. Recently, the combined method is develop in forecasting. It intended to improve forecasting accuracy. Many researches showed that combined model yielded better forecast accuracy than an individual forecasting model. The purpose of this study is to develop combining techniques for rainfall prediction in certain area in Indonesia based on ARIMA and ANFIS.

Makridakis and Hibbon [2] stated that based on M3-Competition, more complicated model not always yield better forecast than simpler one. On the other hand, there is a statement shows that when the various methods are being combined, the accuracy is better than the individual method. Ten-daily rainfall data in Tlekung, Tinjumoyo, Temas, Pendem, Ngujung, and Ngaglik are used as case study. Four main forecasting methods will be developed and compared, i.e. ARIMA, ANFIS, Hybrid methods, and Ensemble methods. The best model is measured by the smallest root of mean square errors (RMSE) at testing datasets.

2. Material and Method

A. ARIMA

ARIMA model are introduced by Box and Jenkins in 1976. It is a linear time series models that can be used for modeling many different types of seasonal as well as non seasonal time series. The mathematics form of the ARIMA model is [3][4] [5]:

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D y_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t \quad (1)$$

where

$$\begin{aligned}\phi_p(B) &= 1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p\\ \Phi_P(B^S) &= 1-\Phi_1 B^S-\Phi_2 B^{2S}-\ldots-\Phi_P B^{PS}\\ \theta_q(B) &= 1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q\\ \Theta_Q(B^S) &= 1-\Theta_1 B^S-\Theta_2 B^{2S}-\ldots-\Theta_Q B^{QS},\end{aligned}$$

and S is the seasonal period, B is the backshift operator, and ε_t is a sequence of white noise with zero mean and constant variance. Box and Jenkins proposed four iterative steps for building ARIMA model, known as Box-Jenkins procedure, i.e. identification, parameter estimation, diagnostic checking, and forecasting.

B. ANFIS

ANFIS is a combination between artificial neural networks and fuzzy inference system. It has a hybrid algorithm to estimate the parameters, least square to estimate the linear parameter and error back propagation to estimate nonlinear parameter. There are five layers in ANFIS. The linear parameter will be estimated in first layer and the nonlinear parameter in fourth layer.

There are some technical terms in ANFIS modeling, such as fuzzy set and fuzzy inference systems. Both terms are the basis of ANFIS modeling. Fuzzy set is the set where the membership of each element does not have clear boundaries [1]. Fuzzy inference system is a method that interprets the

values in the input vector and, based on user-defined rules, assigns values to the output vector. Fuzzy set has a concept that difference with the classical set. The input space is mapped into a given weight or degree of membership through a function called membership function. It defines how each point in the input space is mapped into weights or degrees of membership between 0 and 1. In this study, we will use three types of membership functions and two types of clustering method, i.e. Triangular, Gaussian, and Gbell membership function, grid partition and fuzzy cluster mean as clustering methods. The input space will be mapped into 2 and 3 clusters. Figure 1 is an example of ANFIS structure with two inputs and two rules.

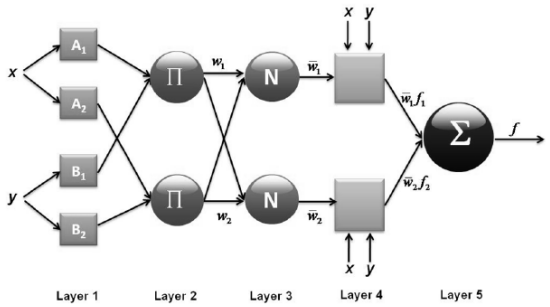

Figure 1. Architecture of ANFIS with two inputs and two rules

C. Hybrid Model

Hybrid model is a forecasting technique that combined other individual methods. In this method, there are two level procedure to get the forecasting value, i.e. linear model on first level and nonlinear model on second level. Zhang [6] used hybrid ARIMA-ANN where ARIMA as linear model and ANN as nonlinear model. It purposed to both of them can hold linear and nonlinear pattern from the datasets.

$$Y_t = L_t + N_t \tag{2}$$

where L_t is linear component and N_t is nonlinear component. Firstly, we make the ARIMA model for linear component, and then residual of linear component will contain nonlinear relationship. For example, e_t is residual notation at t from linear model?

$$e_t = Y_t - \hat{L}_t \tag{3}$$

and the mathematics equation from nonlinear model where is in this study we use ANFIS model is,:

$$e_t = f(e_{t-1}, e_{t-2}, ..., e_{t-n}) + \varepsilon_t \tag{4}$$

f is nonlinear function and ε_t is random error. We can write the mathematics equation from hybrid model as :

$$\hat{Y}_t = \hat{L}_t + \hat{N}_t \tag{5}$$

dengan $\hat{N}_t$ is the forecast value from equation (4).

D. Ensemble Model

Ensemble forecasting is a forecasting technique that combines several outputs of forecasting methods. Recent studies have shown that the robustness and reliability of an combining several models into an ensemble [7]-[12]. In ensemble model, there are two methods usually be used for combining the difference outputs from membership ensemble, i.e. averaging and stacking [11]. Figure 2 shows a review of the two methods in ensemble model:

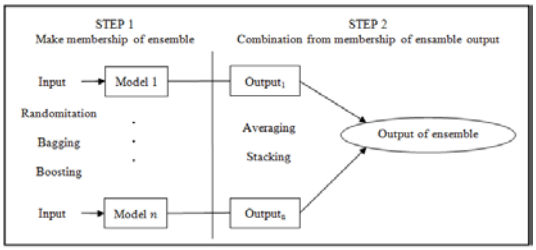

Figure 2. Architecture of ensemble

In averaging method, the output of the ensemble is obtained by computing the mean of the output of the member networks. Assume that N is the number of individual ANFIS members in an ensemble, the combination function (*f*) is:

$$\hat{y}_i = f(\hat{y}_i^k) \quad , i = 1, 2, ..., m \tag{6}$$

where $\hat{y}_i$ is a forecast value of the instance i obtained from the k^{th} network and the form of the function f is

$$f(\hat{y}_i^k) = \frac{1}{N} \sum_{k=1}^{N} \hat{y}_i^k \tag{7}$$

The averaging approach is easy, and it has been shown to be an effective approach to improve the performance of the individual forecasting model like Neural Network and ANFIS [13].

Stacking is a general method with the combination of a higher-level model and the lower-level models with the purpose of achieve a greater predictive accuracy. Breiman [14] suggested minimizing the function G that can give better generalization the model, is

$$G = \sum_{i=1}^{m} \left[y_i - \sum_{k=1}^{N} c_k \hat{y}_i^k \right]^2 \quad , c_k > 0, \sum_{k=1}^{N} c_k = 1 \cdot \tag{8}$$

The coefficients $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_N$ in (8) are predicted in order to create the final output of the ensemble:

$$\hat{y}_i = \sum_{k=1}^{N} c_k \hat{y}_i^k \quad , i = 1, 2, ..., m \tag{9}$$

3. Result and Discussion

Six rainfall data from January 1996 until June at six area in Indonesia, i.e. Tlekung, Tinjumoyo, Temas, Pendem, Ngujung, and Ngaglik are used as case study. The data are divided in two parts, i.e. training and testing datasets. Rainfalls data from January 2011–June 2012 are used as testing datasets. Time series plot for each datasets in same graph is shown in Figure 3.

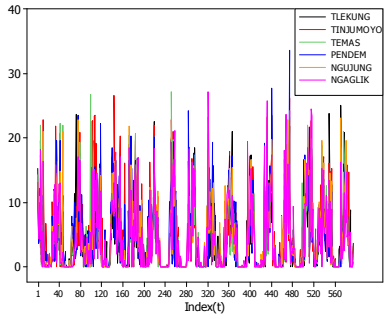

Figure 3. Time series plot of six rainfall data

Figure 3 shows that the six datasets have same pattern. There are some extreme values in January 1996 until June 2012. Four methods are used to forecast the six rainfall data, i.e. ARIMA and ANFIS as individual methods, hybrid and ensemble model as combined methods. Accuracy of data is measured by root mean square error for each method. The best model is chosen to predict rainfall for each area based on smallest RMSE.

For each dataset, ARIMA modeling is done by implementing Box-Jenkins procedure, beginning with identification, parameter estimation, diagnostic checking, and forecasting. All data are not stationary. Consequently, differencing is done to determine the order of ARIMA. Identification step based on autocorrelation function (ACF) and partial autocorrelation function (PACF) of each data yield more than one order of tentative ARIMA. The ARIMA models for each area are shown in Table 1. They have significant parameter and white noise process but normal distribution of residual didn't meet. It happened because there are several extreme values in data. There is a solution if normal distribution of residual didn't meet in ARIMA model, such as entering the outlier to model (ARIMAX).

Table 1: ARIMA model of rainfall data for each area

Area	*Model*
Tlekung	$ARIMA(0,0,0)(0,1,1)^{36}$
	$ARIMA((29),0,(29))(0,1,1)^{36}$
Tinjumoyo	$ARIMA(1,0,1)(0,1,1)^{36}$
	$ARIMA((5),0,1)(0,1,1)^{36}$
Temas	$ARIMA(1,0,1)(0,1,1)^{36}$
	$ARIMA(0,0,0)(0,1,1)^{36}$
	$ARIMA(0,0,1)(0,1,1)^{36}$
Pendem	$ARIMA(0,0,0)(0,1,1)^{36}$
	$ARIMA(1,0,0)(0,1,1)^{36}$
	$ARIMA(0,0,1)(0,1,1)^{36}$
Ngujung	$ARIMA((2),0,2)(0,1,1)^{36}$
	$ARIMA(1,0,1)(0,1,1)^{36}$
	$ARIMA(1,0,0)(0,1,1)^{36}$
Ngaglik	$ARIMA(1,0,0)(0,1,1)^{36}$
	$ARIMA(0,0,1)(0,1,1)^{36}$

To get ANFIS model, we must determine the input and output target. We use significant lag based on the ARIMA models as input variable. For example, there are two kind input variable in Tlekung, i.e. lag 36 derived from model one, and lag 29, 36, 65 from model two.

Table 2 : The Best Model for Each Methods in Six Area

Model	*RMSE in-sample*	*RMSE testing*
Tlekung :		
$ARIMA(0,0,0)(0,1,1)^{36}$	4,358	5,177
$ARIMAX(0,0,0)(0,1,1)^{36}$	2,481	5,474
ANFIS_3input_2cluster_FCM	4,477	5,526
HybridARIMA-ANFIS	4,217	5,243
HybridARIMAX-ANFIS	2,395	5,545
Ensemble Multi Model Averaging	3,382	5,128*
Tinjumoyo :		
$ARIMA(1,0,1)(0,1,1)^{36}$	5,212	4,905*
$ARIMAX(0,0,1)(0,1,1)^{36}$	2,713	5,997
ANFIS_1input_2cluster_Gbell_GP	5,098	5,396
HybridARIMA-ANFIS	4,870	4,967
HybridARIMAX-ANFIS	2,604	6,01
Ensemble single model hibrida averaging	3,736	5,096
Temas :		
$ARIMA(0,0,0)(0,1,1)^{36}$	4,847	4,442*
$ARIMAX(0,0,(1,4,5,8,10,15))(0,1,1)^{36}$	1,784	5,872
ANFIS_1input_2cluster_Gbell_GP	4,909	5,239
Hybrid ARIMA-ANFIS	4,868	4,698
Hybrid ARIMAX-ANFIS	1,819	5,864
Ensemble single model ARIMA averaging	3,781	4,696
Pendem :		
$ARIMA(0,0,1)(0,1,1)^{36}$	4,852	4,067*
$ARIMAX(0,0,(1,31))(0,1,1)^{36}$	3,156	4,613
ANFIS_1input_3cluster_FCM	4,869	5,258
Hybrid ARIMA-ANFIS	4,834	4,144
Hybrid ARIMAX-ANFIS	3,179	4,64
Ensemble single model ARIMA averaging	4,188	4,138
Ngujung :		
$ARIMA(1,0,0)(0,1,1)^{36}$	4,513	4,593*
$ARIMAX(1,0,0)(0,1,1)^{36}$	1,706	5,678
ANFIS_1input_2cluster_Trianguler_GP	4,575	5,671
Hybrid ARIMA-ANFIS	4,477	4,912
Hybrid ARIMAX-ANFIS	1,717	5,662
Ensemble single model ARIMA averaging	3,551	4,785
Ngaglik :		
$ARIMA(0,0,1)(0,1,1)^{36}$	4,726	4,192*
$ARIMAX(0,0,1)(0,1,1)^{36}$	2,2	5,148
ANFIS_1input_2cluster_Trianguler_GP	4,758	4,728
Hybrid ARIMA-ANFIS	4,617	4,462
Hybrid ARIMAX-ANFIS	2,175	5,044
Ensemble single model ARIMA averaging	3,542	4,395

Table 2 is the result of forecasting accuracy for each method. To get the best prediction, we use RMSE from testing datasets. Table 2 show that the best model for predict ten daily rainfall for Tlekung is combination methods, ensemble multi model averaging. But for other area, ARIMA is the best model for predict the rainfall although residual form this model didn't have normal distribution. Hybrid method not always provides better accuracy than the initial model.

4. Conclusion

The rainfall pattern from each area is almost same but the result model is different. The best model for predict ten daily rainfall for Tlekung is combination methods, ensemble multi model averaging. But for other area, ARIMA is the best model for predict the rainfall. It proves that more complex methods are sometimes no better than a simple method (M3-Competition). But sometimes, combination from various methods can increase the accuracy.

References

[1] J.-S.R. Jang, C.-T. Sun, and E. Mizutani, Neuro-Fuzzy and Soft Computing, A Computational Approach to Learning and Machine Intelligence, Upper Saddle River, NJ: Prentice-Hall, Inc., 1997.

[2] S. Makridakis, and M. Hibon, "The M3-Competition: results, conclusions and implications," International Journal of Forecasting, vol. 16, pp. 451-476, 2000.

[3] Bowerman, B.L., and O'Connell, R.T, "Forecasting and Time Series: An Applied Approach. 3rd edition. California: Duxbury Press. 1993.

[4] W.W.S. Wei, Time Series Analysis, Univariate and Multivariate Methods, 2nd ed., Pennsylvania: Pearson Education Inc., pp. 108-134. 2006.

[5] J.D. Cryer, and K.S. Chan, Time Series Analysis with Application in R, 2nd ed. New York: Springer Science+Business Media, pp. 109-141. 2008.

[6] Zhang, G. P. (2003). Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomputing, 50*, hal.159-175.

[7] M. Leutbecher and T.N. Palmer, "Ensemble forecasting," Journal of Computational Physics, vol. 227, pp. 3515-3539, 2008.

[8] D.V. Sridha, R.C. Seagrave, and E.B. Bartlett, "Process modeling using stacked neural network," AIChE Journal, vol. 42, pp. 2529-2539, 1996.

[9] J. Zhang, E.B. Martin, A.J. Morris, and C. Kiparissides, "Inferential estimation of polymer quality using stacked neural networks," Computer Chemical Engineering, vol. 21, pp. 1025-s1030, 1997.

[10] C. Shu, and D.H. Burn, "Artificial neural network ensembles and their application in pooled flood frequency analysis," Water Resource Research, vol. 40, pp. 9, 2004.

[11] I. Zaier, C. Shu, T.B.M.J. Ouarda, O. Seidou, and F. Chebana, "Estimation of ice thickness on lakes using artificial neural network ensembles," Journal of Hydrology, vol. 383, pp. 330-340, 2010.

[12] A.J.C. Sharkey, "Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems (Ed.), Springer-Verlag, New York, 1999.

[13] C.M. Bishop, Neural Network for Pattern Recognition, Oxford: Clarandon Press, 1995.

[14] L. Breiman, "Stacked regression," Machine Learning, vol. 24, pp. 49-64, 1996.

Author Profile

Ria Faulina is Master Student in Department of Statistics from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. She has graduated from bachelor degree of Statistics Institut Teknologi Sepuluh Nopember Surabaya in 2011. She continued to Magister of Statistics in 2011 until now. She has focused her study in forecasting.

Mr. Suhartono received her Bachelor degree in Statistics from Institut Teknologi Sepuluh Nopember Surabaya-Indonesia, Master in UMIST, Manchester and Doctor in Universitas Gadjah Mada Indonesia. Recently, he is a lecturer in Department of Statistics Institut Teknologi Sepuluh Nopember. Surabaya-Indonesia. He has focused her study in time series forecasting, econometrics, neural network, and spatial time series model.