Statistic Approach versus Artificial Intelligence for Rainfall Prediction Based on Data Series

doi:10.3991/ijxx.vxnx.xxx

Indrabayu#1, N. Harun*2 , M.S. Pallu#3 and A. Achmad*4

# Civil Engineering Hasanuddin University
Makassar, Indonesia
[1] indrabayu@unhas.ac.id
[4] salehpallu@hotmail.com

* Electrical Engineering.Dept. Hasanuddin University
[2] n_harun@unhas.ac.id
[3] andani@unhas.ac.id

***Abstract*— This paper proposed a new idea in comparing two common predictors i.e. the statistic method and artificial intelligence (AI) for rainfall prediction using empirical data series. The statistic method uses Auto-Regressive Integrated Moving (ARIMA) and Adaptive Splines Threshold Autoregressive (ASTAR), most favorable statistic tools, while in the AI, combination of Genetic Algorithm-Neural Network (GA-NN) is chosen. The results show that ASTAR gives best prediction compare to others, in term of root mean square (RMSE) and following trend between prediction and actual.**

Keyword- Rain Prediction, ARIMA, ASTAR and GA-NN

I. INTRODUCTION

Makassar is the fourth largest city in Indonesia, and the largest city in eastern Indonesia. Like other cities in Indonesia, Makassar seasons are affected by the rainy season from October to April period and the dry season period from May to September. Weather and climate information is very crucial to support the activities in the various sectors, especially with regard to water resources management. Such information may include rainfall prediction. Nowadays various prediction methods have been developed. The methods developed include qualitative and quantitative prediction methods [1],[2],[3],[4].

A wide range of quantitative approach using AI and Statistic but none have shown comparison in term of root mean square error (RMSE) and comparative of trend between prediction and actual [5],[6],[8],[9]. Statistic approach are very common implemented for stationary data like power consumption [7], however the implementation in non-stationary data like daily empirical rain data have not been tested.

In this paper, the rainfall prediction obtained by compared two methods, statistic approach and artificial intelligences. The GA-NN is generating using Matlab 14 while ARIMA and ASTAR is conducted over several software i.e. MiniTab, SPSS and MARS. The meteorological data is obtained from BMKG Indonesia. The comprises land surface temperature (LST), humidity (H), wind speed direction (W) and empirical rainfall (RF) data, This paper comprises of 5 parts i.e. Introduction, prediction methods, system design, results of processing and conclusions.

II. PREDICTION METHODS

A. *Rainfall Prediction Method*

In forecasting rainfall, commonly used methods are:

- Regression Methods
- ARIMA
- Artificial Neural Network (ANN)

Forecasting rainfall is also used to determine the next rainy conditions for a specified time, it is also used for climate monitoring, detection of drought, bad weather (storms, etc.), warning and flood forecasting monitoring and controlling watershed. In this paper other statistic approach, ASTAR, is implemented for its powerful feature in dealing with non-stationary data like rain series. For better prediction in AI, NN itself will have lesser accuracy, hence in this paper, NN is combined with GA to boost the prediction.

1) ARIMA

ARIMA is used to predict a value in a response time series as a linear combination of its own past values, past errors, and current and past values of other time series. The ARIMA procedure provides a comprehensive set of tools for uni-variate time series model identification, parameter estimation, and forecasting, and it offers great flexibility in the kinds of ARIMA or ARIMAX models that can be analyzed.

The ARIMA procedure supports seasonal, subset, and factored ARIMA models; multiple regression analysis with ARMA errors; and rational transfer function models of any complexity. In general, the ARIMA procedure can be subtle as follows [9]:

Step 0) A class of models is formulated assuming certain hypotheses.

In this step, a general ARIMA formulation is selected to model the rain fall data. This selection is carried out by careful inspection and selection of the main characteristic of the daily rain fall and other meteorological data. The corresponding data are: humidity, air pressure, surface land temperature and wind velocity (corresponding to daily respectively), among others.

Step 1) A model is identified for the observed data.

A trial model must be identified for the rain fall data. First, in order to make the underlying process stationary (a more homogeneous mean and variance), a transformation of the original rain fall data and the inclusion of factors of the form may be necessary. In this step, the checking process can be done using Autocorrelation function (ACF) or unit root test. A further check for lag residual and lag dependent tested from partial ACF.

Step 2) The model parameters are estimated.

After the functions of the model have been specified, the parameters of these functions must be estimated. Good estimators of the parameters can be computed by assuming the data are observations of a stationary time series (Step 1). If a Moving Average (MA) pattern is identified then further optimization process needed by using maximum likelihood or least square estimation.

A conditional likelihood function is selected in order to get a good starting point to obtain an exact likelihood function. Also, an option to detect and adjust possible unusual observations is selected. As these events are not initially known, a procedure that detects and minimizes the effect of the outliers is necessary. With this adjustment, a better understanding of the series, a better modeling and estimation, and, finally, a better forecasting performance is achieved.

Step 3) If the hypotheses of the model are validated, go to Step 4, otherwise go to Step 1 to refine the model.

In this step, a diagnosis check is used to validate the model assumptions of Step 0. This diagnosis checks if the hypotheses made on the residuals (actual prices minus fitted prices, as estimated in Step 1) are true. Residuals must satisfy the requirements of a white noise process: zero mean, constant variance, uncorrelated process and normal distribution. These requirements can be checked by taking tests for randomness, such as the autocorrelation and partial autocorrelation plots. If the hypotheses on the residuals are validated by tests and plots, then, the model can be used to forecast prices. Otherwise, the residuals contain a certain structure that should be studied to refine the model in Step 1.

Step 4) The model is ready for forecasting.

In Step 4, the model from Step 2 can be used to predict future values of daily rainfall data. Due to this requirement, difficulties may arise because predictions can be less certain as the forecast lead time becomes larger. Based on the natural of data, time series forecasting is suit to short term forecasting (hourly or daily). For a long term period, a structural forecaster is more comply for the situation.

The flowchart of corresponding steps above can be seen in Fig. 1. Several historical daily data is collected from BMKG Makassar over 10 year periods (2001-2010) [6].

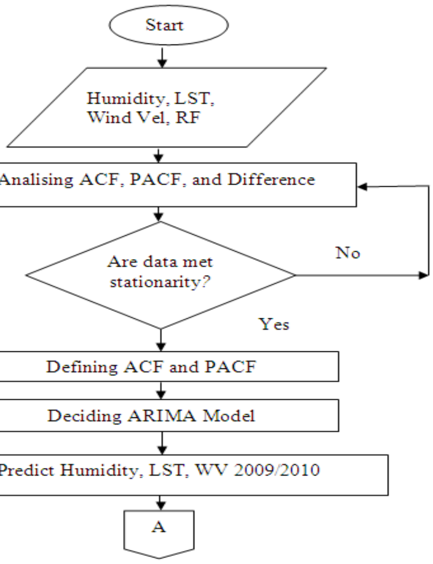

Fig. 1. ARIMA Steps

2) Double Regression

Double regression as part of multivariate analysis aiming on revealing the substantial relationship between two variable. A dependent variable Y is influenced by subsequent independent variable X. General view of the process is shown in Fig. 2.

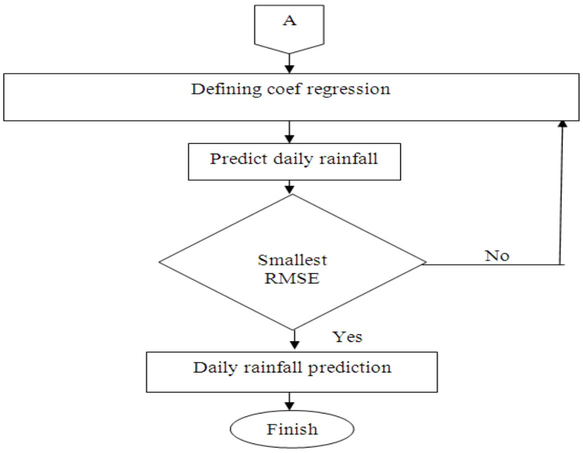

Fig. 2. Double Regression Steps

The coefficient of independent variable Y (rain fall) and subsequent dependent variables Xn are formulated as follows:

Eq.1

$$Y = \beta_0 + \beta_1 X_1 + \beta 2 X_2 + \beta_3 X_3 \ldots + \beta_n X_{ni}$$

Y : Dependent/response Variable

X_1 : Independence parameter 1

X_2 : Independence parameter 2

X_n : Independence parameter n

β: Regression Coefficient

B. Adaptive Spline Threshold Autoregression (ASTAR)

ASTAR is a method of modeling nonlinear time series threshold as Multivariate Adaptive Regression method development Splines (MARS) with predicting the value of lag (zt-d), as in the time series [5]. ASTAR has advantage in its ability to form a model with limit cycles when the data model time series exhibit periodic behavior.

One example ASTAR models are as follows:

$$Zt = c + \quad {}_1 (Zt - d_1 - t_1) + \quad {}_2 (Zt - d_2 - t_2) + \quad {}_3 (Zt - d_1 – t_1) (Zt - d_2 - t_2) + ... + \varepsilon t \qquad \text{Eq.2}$$

C = constants

= coefficient

t_1, t_2 = knots each value of the variable Zt-d1, and

Zt-d_2, d_1, d_2 = a lag values 1 and 2. Knots value is the amount for which there is a change function.

Most of the research in time series data by using modeling and analysis in the model, assumes a linear shape. Though not all of the data in the form of linear, including climate data. Therefore we need a nonlinear time series modeling. One method is developed for nonlinear time series data is Splines Adaptive Threshold Autoregressive (ASTAR). Lewis and Steven (1991) using MARS method with predictor variables lagged values of time series data. So the model obtained contains autoregressive model. ASTAR model is a development of the MARS regression with {Zt} as the response variable and {Zt-j} as the predictor variable.

Formation of knots in the same ASTAR knots forming in MARS. Point knots in ASTAR usually referred to as the threshold. For example Zt, where t = 1,2, ..., N is a variable response from ASTAR models and predictor variables lag p = 3, namely Zt-1, Zt-2, and Zt-3. The selection starts with the forward only constant function B0 (x) = 1. At each step forward stepwise algorithm in ASTAR select one of a set of new forms for ASTAR models. Selection of the number and location of knots automatically based on variable values of ASR (Average sum of square residual) is minimum. When the algorithm is complete then proceed stepwise with backward algorithms. Backward algorithm is used to eliminate some of the base functions to function basis where $1 \leq S \leq M$, in order to obtain a model that minimizes the GCV.

C. Genetic Algorithm

Genetic algorithms are algorithms that attempt to apply an understanding of the natural evolution in problem-solving tasks (problem solving). The approach taken by this algorithm is to combine a wide selection of solutions randomly within a population and then evaluate them to get the best solution.

By doing this process repeatedly, these algorithms simulate the process of evolution as the desired number of generations. This generation will represent improvements on previous population. In the end, we will get the best solutions appropriate to the problems faced. To use a genetic algorithm, solutions to problems represented as a set of genes that make up chromosomes. This chromosome was randomly based coding techniques are used. The entire set of chromosomes is observed representative of the population.

Chromosomes will be evolved in several stages iterations called generations. The new generation is obtained by cross breeding techniques (crossover) and mutation (mutation). Crossover includes cutting two pieces of chromosomes based on the desired number of points and then combine half of each chromosome with other couples. While mutations include the replacement value of the gene in a chromosome with the value of other genes from other chromosomes become partner. The chromosomes are then evolved to a suitability criterion (fitness) and the set will be selected the best results while others are ignored. Furthermore, the process repeated until you have a chromosome that has the best fit (best fitness) to be taken as the best solution of the problem.

On Genetic Algorithms, the best solution search techniques performed simultaneously at a number of solutions known as population. Individuals in a population are referred to as chromosomes. This chromosome is a solution that is shaped symbol. Initial population is built randomly, while the next population is the result of the evolution of chromosomes through iterations called generations. In each generation, the chromosomes will go through an evaluation process using a measurement tool called the fitness function. Fitness value of a chromosome will show the quality of the chromosomes in the population.

The next generation is known as the child (offspring) are formed from the combination of two generations of chromosomes that act as the parent (parent) using the crossover operator. Besides crossover operator, a chromosome can also be modified by using mutation operators. The population of the new generation is formed by selecting the fitness value of parent chromosome and the fitness value of the chromosomes of children, and discard the other chromosomes so that the population size (the number of chromosomes in a population) constant. After several generations, the algorithm will converge to the best chromosome.

Genetic Algorithms steps for generating initial weight as follows

- Create an initial population randomly of meteorological data.

- Evaluate each individual in the population.
- Generate new population using genetic operations.
- Determine the final result at the time of termination criteria.

III. System Design

The research methodology is described in Fig. 3. The meteorological data are retrieved i.e. humidity (H), wind velocity (W), temperature (T), and rainfall (R) for 8 years data (2001-2008). The consecutive year, 2009-2010 are used for validation of the system accuracy. Data selection, interpolation, and normalization are conducted before processing into the system. The process are optimized till it reach best RMSE, in this case the lowest RMSE. The ARIMA and ASTAR was simulated by minitab and SPSS while the GA-NN was conducted by Matlab 14 in this paper.

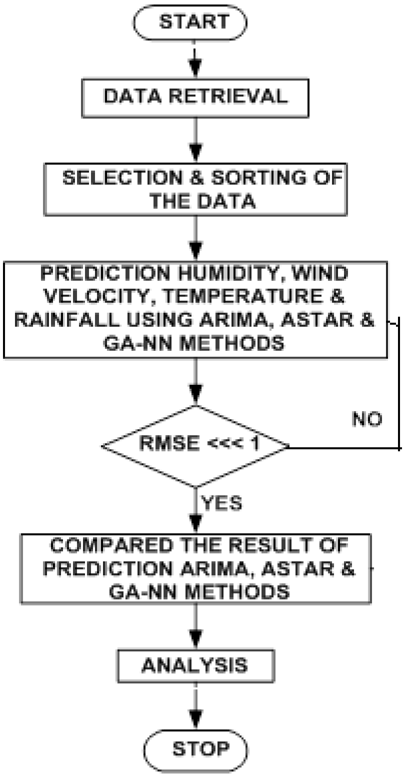
START
DATA RETRIEVAL
SELECTION & SORTING OF THE DATA
PREDICTION HUMIDITY, WIND VELOCITY, TEMPERATURE & RAINFALL USING ARIMA, ASTAR & GA-NN METHODS
RMSE <<< 1
NO
YES
COMPARED THE RESULT OF PREDICTION ARIMA, ASTAR & GA-NN METHODS
ANALYSIS
STOP

Fig. 3. Research Methodology

The main task in the system is the data preprocessing. Due to some imperfect data record from BMKG, data interpolation is needed in some parameters. Several testing on data lagging is also conducted to find best correlation among data.

IV. Results

Autoregressive Integrated Moving Average (ARIMA) is a technique to find the most suitable pattern from a group of data (curve fitting), thereby fully utilizing the data ARIMA past and present to make accurate short-term forecasting. The pictures show a comparison of the results from the prediction of rain per day in February of 2009 and 2010 against the actual data BMKG in 2009 and 2010 are shown in figure 4 and 5.

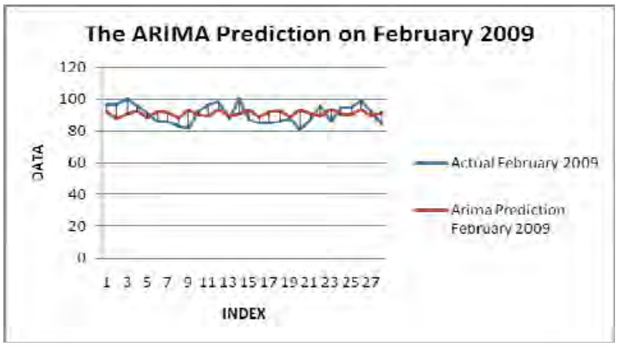

Fig. 4. Comparison of Actual Rainfall and Rainfall Prediction Month February 2009

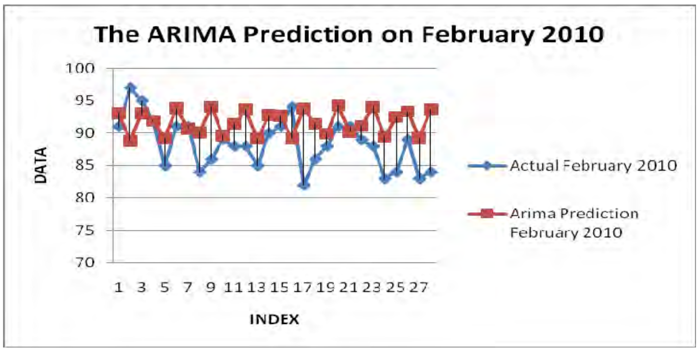

Fig. 5. Comparison Charts Rainfall Actual and Predicted Rainfall in February of 2010

The picture above shows a comparison chart results predicted rain per day in February of 2010 with actual data BMKG in 2010, where the red line is the predicted results and the blue line is the actual data. It can be seen that prediction and actual data are mostly differs to each other compare to 2009 prediction. This is due to the fact that in 2010 rain event were more stochastic and fluctuated. It seems ARIMA can only deal with stationary data and failed to read non-stationary data.

A. *Validation of Prediction Accuracy (RMSE)*

To determine the accuracy of the prediction performance, the importance of knowing Root Mean Square Error (RMSE). The smaller the RMSE the prediction accuracy will get better.

RMSE is calculated as follows:

$$RMSE = \frac{\sqrt{\frac{1}{N}\sum_{t=h}^{N}(y_t - \hat{y}_t)^2}}{y\max - y\min} \text{.....................................}(3)$$

Where: N is the number of data used and y represent individual data

The RMSE value for 2009 and 2010 respectively are 0,24 and 0,301. As stated before in the year of 2010 more fluctuated rain are occurred where over the years rain event has spread even in the dry season.

Adaptive Methods Splines Threshold Autoregression (ASTAR) is a technique to find the best model from a group of data, thus fully utilizing the data ASTAR past and present to make accurate short-term forecasting. To design rainfall prediction in 2011 in ASTAR method, the 2010 data are validated and interpolated. Figure 7 and 8 shows comparison result between actual and prediction.

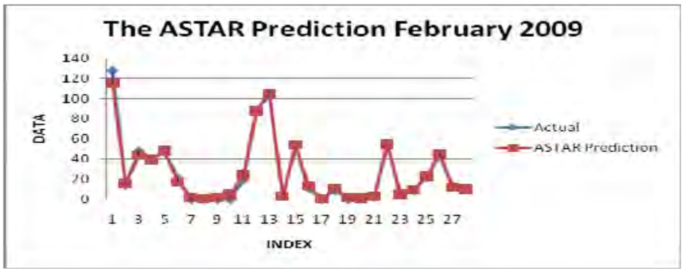

Fig. 6. Comparison of Actual Rainfall and Predicted Month February 2009

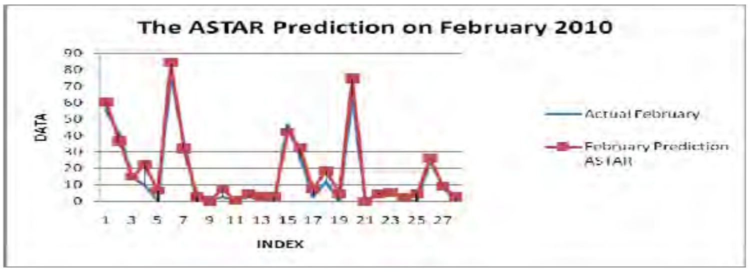

Fig. 7. Comparison Charts Rainfall Actual and Predicted in February of 2010

From Fig. 6 and 7, data prediction are very close to actual data. The trend is also outfit each other for both year 2009 and 2010. It seems ASTAR can handle non-stationary data better than ARIMA.

B. *Overall RMSE for Statistic Methods*

From the results of these predictions is obtained RMSE as following in Table 1:

TABLE I

	RMSE	
	ARIMA	ASTAR
2009	0.24	0.02
2010	0.30	0.06

From the calculation error can be seen that the RMSE Year 2009 and Year 2010 has been qualified predefined the RMSE values much less than 1. ASTAR clearly outperform ARIMA in term of RMSE.

C. *GA-NN System Design*

To determine the input variables that will be used to predict rainfall, the data input of meteorological parameter are first tested on its correlation to rainfall data. This is to gain knowledge of how deep the influence of parameter to the rain event.. Figure 9 and 10 show a comparison of the results from the GA-NN prediction of rain per day in January of 2009 and March of 2009 against the actual data BMKG.

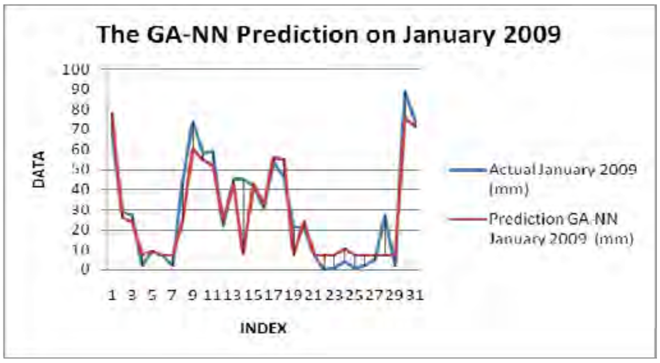

Fig. 8. Comparison of Actual Rainfall and Rainfall Prediction Month January 2009

Fig. 8 shows the trend between prediction and actual data. GA-NN shows better performance than ARIMA but have slightly drawback compare to ASTAR. The RMSE calculated for year 2009 and 2010 are respectively 0.07 and 0.09.

V. RESULTS

The proposed GA-NN method for rainfall prediction was compared to the statistic method which is ARIMA and ASTAR. The results show that statistic approach and artificial intelligence are quite balance in performance. ARIMA has the worst performance due to its weakness in dealing with non-stationary data. ASTAR has the best performance with slightly gain form GA-NN in both RMSE and following trend of prediction to actual. Future research will incorporate other powerful techniques in AI like Support Vector Machine or wavelet.

REFERENCES

[1] Indrabayu, Neural Network and Fuzzy methods for rainfall prediction, *Proc. The 1st Fortei Conference, Makassar*, Indonesia, 2011, pp135 (In Indonesian)

[2] Indrabayu, N. Harun, M. S. Pallu, and A. Ahmad, Constructing Auto-Regressive Integrated Moving Average (ARIMA) as Expert System for Daily Precipitation Forecasting, *The 2nd MICEEI, Makassar*, Indonesia, 2011, pp89.

[3] Indrabayu, N. Harun, M. S. Pallu, and A. Ahmad, Performance of ASTAR for Rainfall Forecasting, *Proc. The 3rd MICEEI, Makassar*, Indonesia, 2012, pp327.

[4] I. Sonjaya, T. Kurniawan, “Uji Aplikasi HyBMG Untuk Prakiraan Curah Hujan Pola Monsunal”, Ekuatorial dan Lokal. BULETIN METEOROLOGI KLIMATOLOGI DAN Vol. 5 No. 3 SEPTEMBER 2009.

[5] Fangqiong Luo, Chunmei Wu and Jiansheng Wu, "A Novel Neural Network Ensemble Model Based on Sample Reconstruction and Projection Pursuit for Rainfall Forecasting", ICNC, IEEE, 2010.

[6] J.F. Nong, "Application of Nonparametric Methods in Short-range Precipitation Forecasting", International Joint Conference on Computational Sciences and Optimization IEEE, 2009.

[7] J. Contreras, “ARIMA Models to Predict Next-Day Electricity Prices”, IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 18, NO. 3, AUGUST 2003.

[8] Lewis, P.A.W and Stevens, J.G., Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Splines (MARS), Journal of the American Statistical Association Vol. 86, No. 416, Dec., 1991, pp. 864-877.

[9] Sutikno. **“**Prakiraan Cuaca dengan Metode ARIMA, NN, dan ASTAR di Stasiun Juanda Surabaya**”.** Jurnal Jurusan Statistik ITS, Surabaya. 2010.