Modelling Multi Regression with Particle Swarm Optimization Method to Food Production Forecasting

Adyan Nur Alfiyatin[1], Wayan Firdaus Mahmudy[2], Yusuf Priyo Anggodo[3]
[1,2]Faculty of Computer Science, Brawijaya University, Malang 65145, Indonesia
[3]Data Scientist, Ilmuone Data, Jakarta 12190, Indonesia
adyannur.alfiyatin@gmail.com

***Abstract*—Tempe was one of the perishable foods with a durability of 2 to 3 days. Tempe home-based industry must take into account its production in order to avoid losses. Suitable planning and forecasting can determine the ways for the production process is implemented. Previously, regression analysis was used as the method to improve the process. This research proposed the particle swarm optimisation (PSO) and multiple regression for production forecasting. PSO is used for optimisation value of regression variable of the tempe productions while multiple regression is used to determine the best coefficients for forecasting. The result of the research showed that the combination of multiple regression and particle swarm optimisation method performed quite well, indicated by the RMSE value of 2.081641.**

***Index Terms*—Forecasting: Multiple Regression; Particle Swarm Optimization; Production.**

I. INTRODUCTION

Indonesia is the largest producer of tempe in the world and as many as 50% of its population consumed tempe as a main dish. The Survey Social Economy National (SUSENAS) results conducted by Central Statistics Agency (BPS) in 2015 shows that the average consumption of tempe every year is 6.99 kg [1]. Tempe customer not only come from all over the country but also globally, evidenced by the standardisation of tempe [2]. Tempe has a short shelf life and stored for a short time [3]. So, tempe producer must have a proper planning to produce tempe according to the needs of the customers. Barbosa, et al. [4] mentioned that one of the biggest challenges in the food and beverage industry is to adjust production and minimise stock product that damage properties. To solve the problem, Barbosa et. al use forecasting analysis model.

Several functions of forecasting, such as production planning, precise production scheduling, are there to help in determining decisions especially for future pricing [4]. There are several approaches to analyse forecasting models. For example, a qualitative approach is a subjective opinion and expert judgement. The quantitative approach is predicting the future by the function of past data while the naïve approach is an objective forecasting model but this approximation approach for different periods is considered equal. A causal approach is an approximation approach that identifies variables which affect the estimate [5]. From several studies stated that the causal approach gets higher accuracy results, primarily using regression analysis [6]-[9].

Some methods that can be used for forecasting are regression, neural network [8], support vector regression [9], backpropagation [10]. Based on the research, the regression method is capable to produce effective results, especially in multiple regression [11]. Multiple regression is used for the determination of the coefficients [12]. The disadvantage of multiple regression is it has less precision in choosing the influential variable for the forecasting process thus requires a method that can overcome the problem [11]. PSO can be used to supplement deficiencies of the multiple regression. Several studies proved that PSO was effective in selecting the affected variables between global and local search [13]. PSO is an evolutionary algorithm that works in a population-based way and has several characteristics that are simple, easy to implement, and fast in convergent [14]. It is suitable to be applied in various fields [15].

This study aims to model tempe production forecasting by using hybrid PSO and multiple regression. PSO is used to find the value of the regression variable to get a minimum error. When the best results are obtained, then the selection coefficient is made using multiple regression. The focus of this research is to know the performance of the resulting model and to help the home-based industry in the process of purchasing raw material stock for their tempe production process in the future.

II. RELATED WORK

Previous studies on forecasting used statistical approaches such as multiple regression [16], ARIMA [17]. Han and Halpin [16] multiple regression is used to determine the dependent and independent variables. Murphy et al. [18] used a comparison method for prediction and multi-regression to produce maximum result. Amin et al. [17] use the ARIMA method for long-term forecasting, with time series data and optimised yield.

Wahyuni and Mahmudy [19] used fuzzy Tsukamoto approach to predict the future value. In [9], a genetic algorithm is used to optimise membership function to get the maximum result because the algorithm utilized convergence to find a solution. The main quality of genetic algorithm is fast-finding solution in a large space. When using the fuzzy approach, rule based are needed from experts which will influenced the results. Another problem is when there are many perceptions about the rule bases that are varies among many experts in the case. Fuzzy Tsukamoto approach will be difficult to implement.

Jiang and Wu [9] proved that the support vector regression

(SVR) still has not obtained optimal results, so it needs an additional method. The method that was used to improve the accuracy was PSO and it was proven that the addition of PSO has obtained optimal results. In another case, SVR is optimised using PSO and simulated annealing (SA) [20]. Hybrid PSO and SA are used to find a solution to the complex problem. SA helps to find a solution with the focus in global searching. SA is very capable to get a maximum solution but need more time for the computation process.

Neural network approach is one of the methods to solve the prediction problem. Back-propagation is the popular method in the neural network approach. It has many neural to 'learn' all the problem. Paswan and Begum [21] mentioned that neural network is very useful to identify the relationship between the variables, which is difficult to solve using a statistical approach. In forecasting problem, back-propagation succeeded in forecasting the Indonesia inflation rate [10] where the back-propagation is capable to get the maximum result. However, learning in neural network needs a long time because it must determine and test a lot of parameters such as the number of the hidden layer and the number of the node in all layers. In this matter, a statistical method is used because it has an explicit function to determine the value of the parameter. Ghiasi et. al. [22] used the dynamic artificial neural network to improve the performance of the neural network. The result has improved however with less approximate optimal accuracy [15].

Furthermore, there are several hybrid research methods that combine PSO and neural network methods such as Xiao et al. [13]. PSO parameters are flexible that enable balanced exploration and exploitation process of particles. Kosan and Kantaantha [18] utilized hybrid SVM and PSO method to predict production that decrease risk. Yarushev [8] predicted house prices using neural network, multiple regression, fuzzy regression and economic methods. The conclusions of his research indicated that the neural network method are well matched with any method, but superior results are obtained when combined with multiple regression. Hsieh [18] utilized the hybrid PSO-SVR where parameters that are selected using the PSO method obtained better results than the parameters with conventional SVR.

All methods to solve the above-mentioned problems must need optimisation to get the maximum result. Many researchers use a heuristic method or a neural network approach. In many cases, the neural network needs a long time for process computation. So, the heuristic method is preferred to quickly solve the problem. Rahmi and Mahmudy [12] explained the method to solve the forecasting problem. Regression modelling with genetic algorithm for optimisation produced the best result. In this case, the genetic algorithm can work at maximum to find a solution. However, the genetic algorithm does not get a optimum solution. Usually, genetic algorithm often is trapped in global searching, then it cannot find a solution in the local area.

With regards to this matter, although PSO is focusing in finding a solution in the local area, at the same time PSO can balance local search and global search. Based on that, in a small space solution, PSO can provide a maximum solution. Several studies show that the PSO can implement variable optimisation with a combination of various methods and proven to get optimal results. So, the focus of this research is to combine the method of particle swarm optimisation and multiple regression to obtain optimal results and in accordance with the field data.

III. RESEARCH METHODOLOGY

The data in this research was obtained from the home industry 'ABC' Malang from January 2015 until December 2016. This data was processed in the system to look for patterns of relationship that can be used to forecast the tempe production. Data obtained are the number of raw materials, the number of requests, the number of productions, and time series data. The research flow diagram is shown in Figure 1.

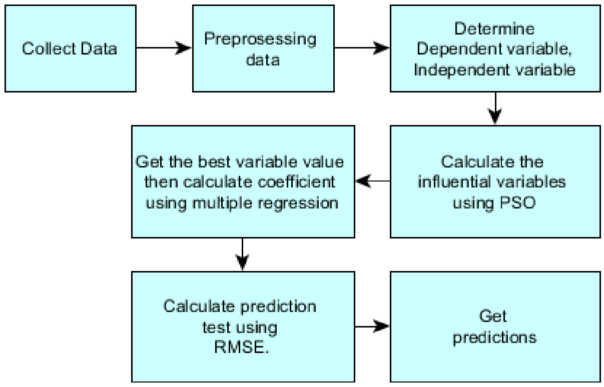

Figure 1: Research flow diagram

A. Regression Models

Regression analysis was introduced by Galton. Regression analysis was the method to determine relationship cause-result between one variable with another variable [23]. Usually, a regression model was used as prediction method. It has many models, for example linear regression, multiple regression, and quadratic regression. Linear regression is the only model which only used one variable that was influenced.

Multiple regression deals with the study of the dependency of one variable (the dependent variable) to one or more other variables (explaining variables) with the intention to estimate and forecast the average value of the dependent variable if the explanatory variable is known. In the regression model, the dependent variable depends on two or more explanatory variables, commonly called multiple regression.

Multiple regression was a method of two variables; those are independent variable (y) and the dependent variable (x) [21]. Based on the data obtained, an independent variable in this research is the amount of production. The dependent variables are classified as the amount of raw material, the amount of request product, and the time-series data. This method aims to determine the influential variable and obtained the regression model for predictive production.

The formula that states inter-relationship between variables is shown in Equation (1).

$$y` = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + + \beta_n X_n \qquad (1)$$

where Y' is the dependent variable, $\beta_0, \beta_1, \beta_2, \beta_3, \ldots \beta_n$ are a parameter of free variable regression coefficient, and $X_1, X_2, X_3, \ldots, X_n$ are the independent variable

Based on Equation (1), the resulting regression model is shown in Equation (2).

$$y = 63{,}1 + 0{,}080\, x^1 + 0{,}240\, x^2 + 0{,}043\, x^3 - 0{,}138\, x^4 \qquad (2)$$

```
Predictor      Coef   SE Coef       T      P
VIF
Constant      63,07     13,14    4,80  0,000
X1           0,0798    0,1020    0,78  0,436
1,0
X2          0,23974   0,09162    2,62  0,010
1,0
X3           0,0426    0,1109    0,38  0,702
1,2
X4          -0,1377    0,1124   -1,22  0,224
1,2


S = 30,1971   R-Sq = 9,2%   R-Sq(adj) = 5,0%


Analysis of Variance

Source           DF        SS       MS      F
P
Regression        4    8100,4   2025,1   2,22
0,073
Residual Error   88   80244,2    911,9
Total            92   88344,6


Durbin-Watson statistic = 1,94851
```

Based on the regression model that has been marked positive on variables x_1, x_2, x_3 indicated that the variable has affected to the production process. The negative sign on the variable x_4 indicated the variable does not affect the process production.

B. Particle Swarm Optimization

The concept of PSO was a stochastic optimisation method where problem solution represented by particle [24]. The particle is generated randomly as much as n particles; each particle consists of several dimensions. There are position *xi* and velocity *vi*. The cost value for such particle will be calculated by using Equation (3).

$$\text{Cost} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - p_i)^2} \quad (3)$$

n showed the total data, $\boldsymbol{d_i}$ was actual data, $\boldsymbol{p_i}$ was the prediction data. The first step in the PSO method was to represent the particle by an integer, ranging from 0 – 1 for all variables. There are four variables; variable *x1* is actual data $n-2$, *x2* is actual data $n-1$, *x3* is the number of material, and *x4* is the number of request.

After finalizing the particle representation, the next step is to determine the size of the herd = N randomly. Then, the initial population is generated within the range a (b) and a (a) randomly, in order to obtain *x1*, *x2* , ..., *xN*. Particle *j* and speed on *i* iteration notated as *x(i)j* and *v(i)j*. The initial particles are denoted as the coordinate vectors of the particles. Subsequently, the cost values are calculated. the velocity of each particle is calculated, moving from one position to another influenced by speed. Thus, the best position ca be obtained through an adaptive velocity formulation by using Equation (4) [25].

$$vit+1 = w.vit + c1.r1(pbesti - xi) + c2.r2(pgbesti - xi \quad (4)$$

vi represents the velocity value for the particle dimension *i* to *n*, *t* denotes time of iteration, *w* denotes the value of the inertia vector obtained dynamically using Equation (5) [26]. *pbesti* was the best position obtained to each particle, while *pgbesti* was the best position obtained from the whole particle. *c1* and *c2* are social and cognitive Constanta, which in this research 2,5 for *c1* value and 0,5 for *c2* value. *r1* and *r2* are randomly generated range of [0,1]. This research used 0,5 and 2,5. Equation (6) is used to update the position.

$$\mathrm{w} = (w\max - w\min)^{iterasi-t}_{iterasi} + \mathrm{w}\min \quad (5)$$

$$xi\,t+1 = xi + vi.t+1 \quad (6)$$

Generally, for moving position particle, PSO is considered too fast. Thus in the early stages, it undergoes convergent and does not find the optimum solution. It can be solved using speed control [27]. The speed control mechanism is implemented by performing conditions for the velocity of each particle as follows.

$$if\ (vij\,.t+1 > vj\max)\ then\ vij.t+1 = vj\,max,$$
$$if\ (vij\,.t+1 > vj\min)\ then\ vij.t+1 = vj\,min$$

vj max value generated using Equation (7) and *vj* min was negative value from *vj* max.

$$Vj\max = k\ \frac{(x_{j\,max} - x_{j\,\min})}{2} + k\ \in [0,1] \quad (7)$$

The $\boldsymbol{vi}$ is speed velocity calculation cycle, and the *xi* is updated position continued until the complete iteration or convergent.

C. Random Injection

This method was first proposed by Mahmudy et al. [28] to solve problems of early convergence in genetic algorithms. A common problem of PSO that at certain iteration, the value of the solution is equal to each particle before reaching the optimum solution. According to Utomo et al. [29], the use of k-means and PSO do not produced the optimum solution result. To solve the problem, the handling of early convergence is done using a random injection. Random injection is a simple mechanism and improve accuracy results. It works by inserting random *p* particle in *q* iteration. The best determination of *p* particle values and *q* iterations should be tested first.

IV. Testing and Analysis

Testing of regression and particle swarm optimisation modelling on the prediction of tempe production is done on three testing parameters. These parameters are particle testing, iteration testing and inertia weight testing. Testing is done to determine the most optimal model in this research. Each parameter required ten times to run, and the cost value was calculated from the average cost [20].

A. Particle Testing

In this test, the number of particle test are multiplied by 10. And stop at the amount of 100 particles because the decreased results. The combined inertia weight (0.9 and 0.4) are based on Eberhart and Shi [30]. The combination obtained optimal results as shown in Figure 2.

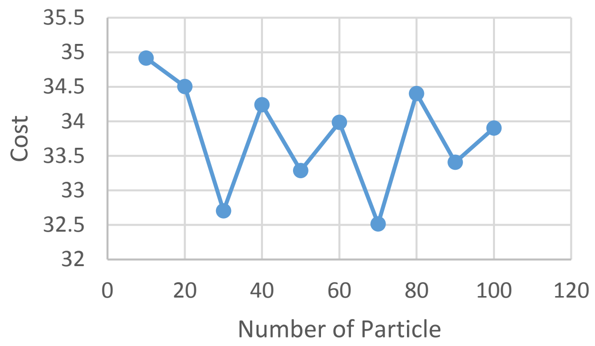

Figure 2: Number of particle testing

Based on the results of particle testing, the highest cost value on the particle is 70 with a value of 32.51. The test was discontinued in the 100 population because the result of cost value decreased, so it is useless if the test is continued.

B. *Iteration Testing*

The iteration test was performed using the best cost result on the particle testing which is 70. The test was conducted by multiplication of 10 with ten times of testing. The test result is shown in Figure 3.

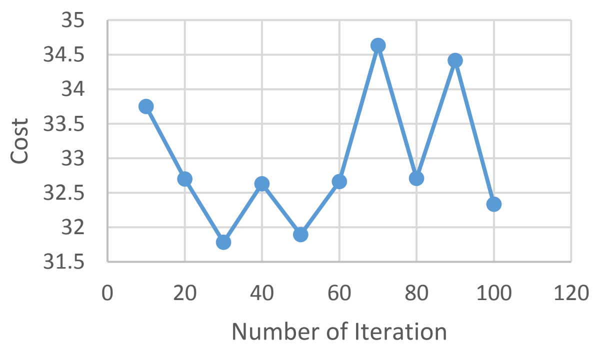

Figure 3: Number of iteration testing

Referring to Figure 3, the highest cost value on iteration 30 with a value of 31.78. Since iteration 100 does not find the best cost value, so the test is stopped.

C. *Inertia Weight Testing*

Testing a combination of inertia weight between range 0.2–0.9. This parameter test used the result of the previous parameter, that the best parameter at number 70 and the best iteration number 30. Figure 4 shows the result of inertia weight combination test.

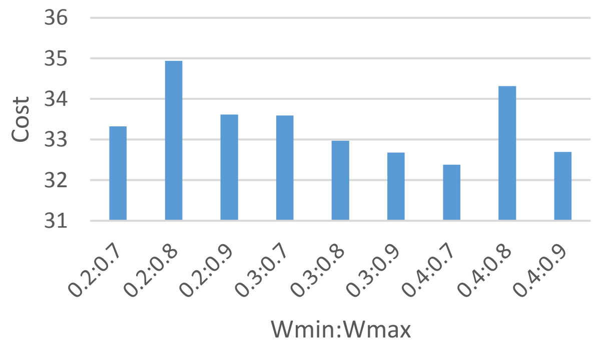

Figure 4: Inertia weight testing

Figure 4 showed that the combination of inertia weight whose gained the best cost value in the combination of 0.4 and 0.7 with the value 32.38. Tests were conducted between range 0.2-0.4 and 0.7-0.9 based on the Eberhart and Shi study obtaining optimum results in that range [30]. Moreover, it is proven in the testing of inertia weight combination that the optimum result is obtained in range 0.4 and 0.7.

After testing those parameter, Figure 5 shows the comparison between prediction and actual data. In general, the predicted results have followed the pattern of the actual value. The result of multi-regression and multi-regression PSO each have 33.68 and 31.07 respectively. The reason for a high RMSE value is because of the lack of parameters used in affecting the production process, hence fewer data are utilized. Therefore, an approach is required to get the archive predictions that close to the actual data. The parameters used in the tempe production data can be divided into new subcategories and afterward, prediction of the production can be done by using multiple regression-PSO models. Finally, the exact predictions value can be obtained. One of the approaches is using fuzzy that can produce maximal results with minimal data in the forecasting problem [31].

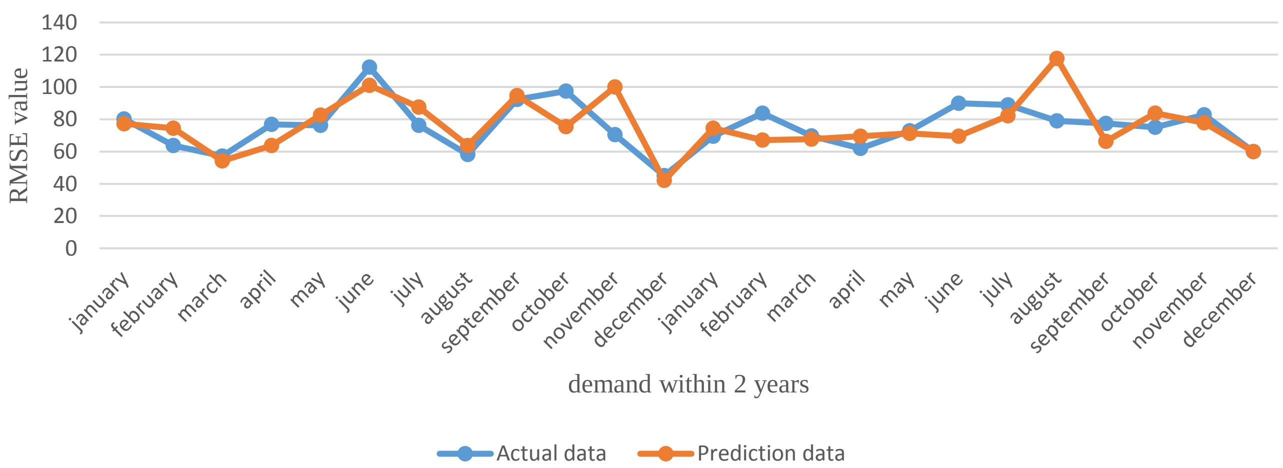

Figure 5: Comparison between the actual and the prediction result

V. Conclusion

This research is conducted several tests using multiple regression and particle swarm optimisation. There parameters are tested on particle swarm optimisation which are particle testing, iteration testing and combination inertia weight. The results indicated that 70 particles, 30 iterations, combination 0.2 and 0.7 are required to get the best cost value. Particle swarm optimisation is used to determine the influential variable on tempe production process while multiple regression is used to optimise the coefficient contained in particle swarm optimisation to obtain the optimal result. This research proved that by using multiple regression and particle swarm optimisation, many results are not suitable for actual data and prediction data. Therefore, in the next research, a combination of fuzzy approach and multiple regression-PSO models will be tested to obtain the prediction results that close to actual data [31]. Moreover, the researcher can add influential variables for products such as the price of soybeans, consumers interest [32] and the data for tempe production as more data utilization will produce more accurate results.

References

[1] Suwandi, *Outlook of Agricultural Commodities Sub Sector of Food Crops*. Jakarta: Center for Agricultural Information System and Data of the Ministry of Agriculture, 2016.

[2] Badan Standardisasi Nasional, *Tempe : Indonesia's Offering to the World*. 2012.

[3] D. Di and K. Rajeg, "Chemical and Organoleptic Composition of Formula Nugget Based on Tempe Flour and Ricebran Flour," *Indones. J. Hum. Nutr.*, vol. 1, no. 2, pp. 135–148, 2014.

[4] N. P. Barbosa, E. S. Christo, and K. A. Costa, "Demand forecasting for production planning in a food company," *ARPN J. Eng. Appl. Sci.*, vol. 10, no. 16, pp. 7137–7141, 2015.

[5] E. C. Daniel, O. U. Patrick, and A. State, "Application of Forecasting Methods for the Estimation of Production Demand By," vol. 3, no. 2, pp. 184–202, 2014.

[6] F. S. Gharehchopogh, T. H. Bonab, and S. R. Khaze, "A Linear Regression Approach to Prediction of Stock Market Trading Volume: A Case Study," *Int. J. Manag. Value Supply Chain.*, vol. 4, no. 3, pp. 25–31, 2013.

[7] S. K. Sheikh and M. Hapse, "Unit Costs Estimation in Sugar Plant using Multiple Regression Least Squares Method," *Int. J. of Adv. in Eng. & Tech.*, vol. 1, no. 4, pp. 299–306, 2011.

[8] С. А. Ярушев, S. A. Yarushev, A. N. Averkin, and A. A.N., "Review of studies on time series forecasting based on hybrid methods, neural networks and multiple regression," *Международный Журнал "Программные Продукты И Системы,"* vol. 31, no. 113, pp. 75–82, 2016.

[9] F. Jiang and W. Wu, "Hybrid Genetic Algorithm and Support Vector Regression Performance in CNY Exchange Rate Prediction," vol. 2, no. Esm, pp. 136–139, 2016.

[10] N. R. Sari, W. F. Mahmudy, and A. P. Wibawa, "Backpropagation on Neural Network Method for Inflation Rate Forecasting in Indonesia," *Int. J. Adv. Soft Comput. Appl.*, vol. 8, no. 3, 2016.

[11] R. Koen and J. Holloway, "Application of multiple regression analysis to forecasting South Africa ' s electricity demand," *J. Energy South. Africa*, vol. 25, no. 4, pp. 48–58, 2014.

[12] A. Rahmi and W. F. Mahmudy, "Regression Modelling for Precipitation Prediction Using Genetic Algorithms," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 15, no. 3, 2017.

[13] J. J. Liu, "Time Series Forecasting Using a Hybrid Adaptive Particle Swarm Optimization and Neural Network Model," vol. 2, no. 4, pp. 335–344, 2014.

[14] Y. Zhang and C. Science, "Prediction of Financial Time Series With," no. May, 2004.

[15] A. Hayder M. Albehadili Abdurrahman and N. . Islam, "An Algorith for Time Series Prediction Using," *Int. J. Sci. Knowl. Comput. Inf. Technol.*, vol. 4, no. 6, pp. 26–33, 2014.

[16] S. Han, F. Drive, C. Bldg, and D. W. Halpin, "The Use Of Simulation For Productivity Estimation Based On Multiple Regression Analysis," 2005, no. 1, pp. 1492–1499.

[17] M. Amin, M. Amanullah, and A. Akbar, "Time Series Modeling for forecasting wheat production of Pakistan," *Plant Sci.*, vol. 24, no. 5, pp. 1444–1451, 2014.

[18] H.-I. Hsieh, T.-P. Lee, and T.-S. Lee, "A Hybrid Particle Swarm Optimization and Support Vector Regression Model for Financial Time Series Forecasting," *Int. J. Bus. Adm.*, vol. 2, no. 2, pp. 48–56, 2011.

[19] I. Wahyuni and W. F. Mahmudy, "Rainfall Prediction in Tengger , Indonesia Using Hybrid Tsukamoto FIS and Genetic Algorithm Method," *J. ICT Resource Appl.*, vol. 11, no. 1, pp. 38–54, 2017.

[20] D. Novitasari, I. Cholissodin, and W. F. Mahmudy, "Hybridizing PSO With SA for Optimizing SVR Applied to Software Effort Estimation," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 14, no. 1, pp. 245–253, 2016.

[21] D. A. Lind, W. G. Marchal, and S. A. Wathen, *Business & Economics Fifth Edition*. 2006.

[22] M. Ghiassi, D. Lio, and B. Moon, "Pre-production forecasting of movie revenues with a dynamic artificial neural network," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 3176–3193, 2015.

[23] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and Statistics for Engineers and Scientists*, vol. 3rd. 2012.

[24] M. Clerc, "Particle swarm optimization," p. 244, 200AD.

[25] R. C. Eberhart and Y. Shi, "Comparing inertia weights and constriction factors in particle swarm optimization," *IEEE Congr. Evol. Comput.*, vol. 1, no. 7, pp. 84–88 vol.1, 2000.

[26] A. Ratnaweera, S. K. Halgamuge, and H. C. Watson, "Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients," *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 240–255, 2004.

[27] F. Marini and B. Walczak, "Particle swarm optimization (PSO). A tutorial," *Chemom. Intell. Lab. Syst.*, vol. 149, pp. 153–165, 2015.

[28] W. F. Mahmudy, R. M. Marian, and L. H. S. Luong, "Optimization of Part Type Selection and Loading Problem with Alternative Production Plans in Flexible Manufacturing System using Hybrid Genetic Algorithms - Part 1 : Modelling and Representation," in *International Conference on Knowledge and Smart Technology (KST)*, 2013, pp. 75–80.

[29] M. C. C. Utomo, W. F. Mahmudy, and S. Anam, "Determining the Neuron Weights of Fuzzy Neural Networks Using Multi-Populations Particle Swarm Optimization for Rainfall Forecasting," *J. Telecommun. Electron. Comput. Eng.*, vol. 9, no. 2, pp. 37–42, 2017.

[30] R. C. Eberhart and Y. Shi, "Comparing inertia weights and constriction factors in particle swarm optimization," vol. 1, no. 7, pp. 84–88 vol.1, 2000.

[31] Y. P. Anggodo and W. F. Mahmudy, "A Novel Forecasting Based on Automatic-Optimized Fuzzy Time Series," *TELKOMNIKA (Telecomunication Computing Electronics and Control)*, vol. 16, no. 4, pp. 1809-1817, 2018.

[32] I. H. Kaya, "Motivation Factors of Consumers ' Food Choice," *Food Nutr. Sci.*, vol. 7, no. March, pp. 149–154, 2016.