Length of Rainy Season Prediction Based on Southern Oscillation Index and Dipole Mode Index Using Support Vector Regression

Abdul Basith Hermanianto[1], *Agus Buono*[2], *Karlina Khiyarin Nisa*[2]

1 Department of Computer Science, Bogor Agricultural University, Bogor, Indonesia (abhermanianto@gmail.com)

2 Department of Computer Science, Bogor Agricultural University, Bogor, Indonesia

Abstract

The location of Indonesia which is between the Pacific and Indian Ocean causes the climate to be influenced by the global condition in both oceans. The goal of this research is to use modelling to predict the length of the rainy season. To do this we use support vector regression (SVR), and the southern oscillation index (SOI) from the Pacific Ocean and the dipole mode index from the Indian Ocean as the predictor variables. The predictive value of this model is evaluated with determination coefficient (R^2) and root mean square error (RMSE). The data used in this research is the length of rainy season data from three weather stations in Pacitan district (Arjosari, Kebon Agung, Pringkuku) between 1982/1983 and 2011/2012 periods as the observation data. SOI and DMI between 1982 and 2011 used in this research as the predictor data. The result of this research is a prediction model for each climate station. The best R^2 for Arjosari, Kebon Agung, and Pringkuku weather stations are 0.73, 0.63, and 0.58 respectively. Meanwhile, the best RMSE for Arjosari, Kebon Agung, and Pringkuku weather stations are 2.45, 3.23, and 2.86 respectively.

Keywords: Dipole Mode Index, length of rainy season, Southern Oscillation Index, Support Vector Regression

Introduction

One of the effects of climate that can be analyzed is the length of rainy season of an area. The length of the rainy season is very influential for many sectors such as agriculture, fisheries, and the construction sector. As an example, Buono et al. (2012) said that the length of the rainy season affects the rice crop production, especially in the second period of planting season. The second season will have a greater chance to experience drought than the first planting season if the length of the rainy season is too short. In the end, it could lead to crop failure. Knowledge about the length of the rainy season in a period needed to determine what the most appropriate policy will be taken to adapt with the climate and the weather condition there.

Indonesia is an archipelago that lies between two oceans, the Pacific Ocean and the Indian Ocean. Therefore, any climate anomalies that occur in both oceans will also affect the climate in Indonesia itself. There are several variables that can represent the global climate change or climate anomalies that occur in both the Pacific Ocean and the Indian Ocean, including the Southern Oscillation Index (SOI) and Dipole Mode Index (DMI). SOI is an index that reflects the condition of the Pacific Ocean as compared to the state of the seas around Indonesia. The selection of the index is based on the fact that the season in Indonesia is influenced by the condition of the Pacific Ocean (Buono et al 2014). SOI is used to indicate the change and the intensity of the El-Nino or La-Nina in the Pacific Ocean. The El-Nino causes Indonesian climate becomes drier and rainfall in Indonesia decreases, while La-Nina causes more rain occurred in Indonesia. Meanwhile, DMI indicates differences between sea surface temperature anomalies in the western and eastern ends of the Indian Ocean (Behera et al 2013). DMI positive value indicates movement of winds in the Equatorial Indian Ocean from east to west, causing Indonesia to have drier climate conditions. Instead DMI negative value indicates movement of winds in the Equatorial Indian Ocean from west to east, thus cause more rain occurred in Indonesia (Chandrasekar 2010). Because SOI and DMI can represent climate conditions in the Pacific Ocean and the Indian Ocean, both global climate variables can be used as predictors to predict the climatic conditions in Indonesia, including its length of rainy season.

Support Vector Machine (SVM) is a machine learning method which is developed based on statistical learning theory. SVM can be used to perform classification and regression. In the case of regression, the output will be real or continuous numbers. SVM that is used for regression is often known as Support Vector Regression (SVR). SVR is a method that can overcome the overfitting, so it will produce a good performance (Smola and Schölkopf 2004). This study aims to perform predictive modeling of the length of rainy season using Support Vector Regression (SVR) with Southern Oscillation Index (SOI) and Dipole Mode Index (DMI) as the predictors.

Data and Research Method

The data used in this study is the rainfall observation data from several weather stations in Pacitan District (East Java Province, Indonesia), Southern Oscillation Index (SOI) data, and Dipole Mode Index (DMI) data. The rainfall observation or daily precipitation data was taken from three weather stations in Pacitan (Arjosari, Kebon Agung, and Pringkuku) between 1982 and 2012. These data obtained from the Center for Climate Risk and Opportunity Management in Southeast Asia Pacific (CCROM - SEAP), Bogor Agricultural University. The daily precipitation data is then used to calculate the length of the rainy season. To obtain the length of the rainy season, daily precipitation data must be first converted into a form of dasarian (ten days). For example, January is consist of three dasarian, namely dasarian I (day 1 to 10 of January), dasarian II (day 11 to 20 of January), and dasarian III (day 21 to the end of the month). The division also applied to the others month so that in one year there will be 36 dasarians. The length of rainy season is calculated based on the number of dasarians from the start of the rainy season until the end of the rainy season. The criteria of the start of the rainy season -according to the Indonesian Bureau of Meteorology, Climatology and Geophysics (BMKG)- is the dasarian when the precipitation of a dasarian more than 50 mm and occurs as two consecutive dasarian. While the end of the rainy season occurs at the dasarian before the start of the drought season. The definition of start of drought season, -according to the Indonesian Bureau of Meteorology, Climatology and Geophysics (BMKG)- is the dasarian when the precipitation on it less than 50 mm, which occurred consecutively for two dasarian.

Figure 1 show the illustration of length of rainy season determination in Arjosari weather station on 2011/2012 period. The red straight line acts as the threshold of the determination, whose value is 50 mm. From the figure, it can be seen that the start of the rainy season occurs at the third dasarian of October and the start of the drought season occurs at the third dasarian of May. So from there it can be concluded that the length of the rainy season in Arjosari on 2011/2012 period is the number of dasarians from the third dasarian of October until the the second dasarian of May (one dasarian before the third dasarian of May) whose value is 21 dasarians.

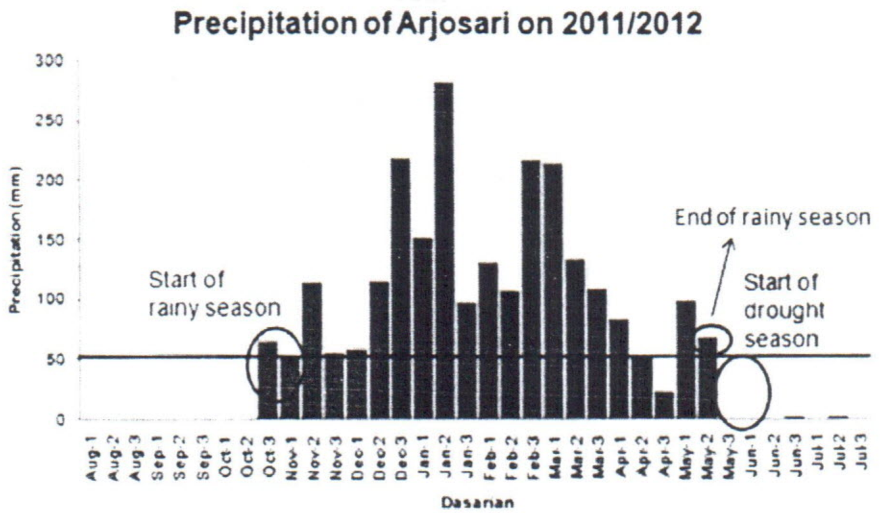

Fig. 1 Illustration of length of rainy season determination

The SOI data obtained from the site of Bureau of Meteorology (BOM), Australia (http://www.bom.gov.au/climate/current/soihtm1.shtml). Clarke (2008) described that SOI is measured by reducing the atmospheric pressure anomalies at Tahiti area (which has been divided by its standard deviation) with atmospheric pressure anomalies in the area of Darwin, Australia (which has been divided its the standard deviation). Calculation of SOI can be formulated as follows:

$$SOI = 10 \times \frac{P_{diff} - P_{diff\,ave}}{Stdev\ P_{diff}}$$

with

P_{diff} = Monthly sea level pressure difference at Tahiti and Darwin

$P_{diff\,ave}$ = average of Monthly sea level pressure difference at Tahiti and Darwin

$Stdev\ P_{diff}$ = Standard deviation of Monthly sea level pressure difference at Tahiti and Darwin

The DMI data obtained from calculating the difference of sea surface temperature (SST) between the western end (60E-80E, 10S-10N) and the eastern end (90E-110E, 10S-0S) of Equatorial Indian Ocean (Saji *et al.* 1999). DMI can be calculated using the following formula:

$$DMI = \frac{P_{diff} - P_{diff\,ave}}{Stdev\ P_{diff}}$$

with

P_{diff} = SST difference between the western and eastern parts of the Indian Ocean

$P_{diff\,ave}$ = average of SST difference between the western and eastern parts of the Indian Ocean

Stdev P_{diff} = Standard deviation of SST difference between the western and eastern parts of the Indian Ocean

SST data itself obtained from the site of the International Research Institute for Climate and Society (IRI), Columbia University, USA. SST data obtained by opening the extended reconstructed sea surface temperature (ERSST) link from IRI Data Library (IRIDL) (http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.ERSST/.version3b/.sst/). At the link then select SST (Data selection section) based on the time span and the desired region. SOI and DMI data used in this study is the SOI and DMI from 1982 to 2012.

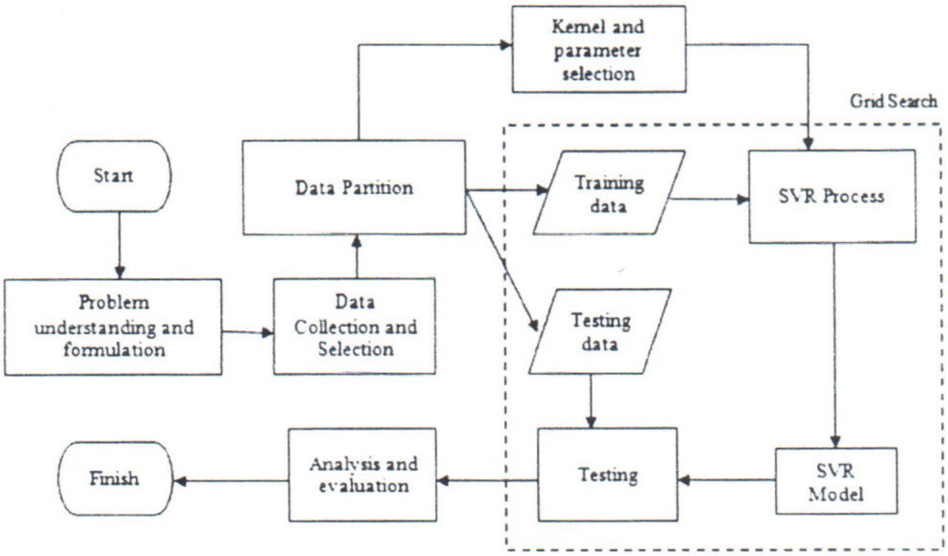

Fig. 2 Research method block diagram

Figure 2 show the research method flowchart. The data selection process involves predictor data selection. The predictors, SOI and DMI, being selected using correlation analysis with the observed length of rainy season. Months of SOI and DMI that satisfy the threshold of α = 10% then selected as the predictors in this study. The selected predictors and the observed length of rainy season then being divided into two parts, training data and testing data, using percentage split whose fraction is 66.67% as training data and 33.33% as testing data.

After the data partition, the training data used to train the support vector regression model. Suppose we have a training data set λ, $\{(x_1, y_1), \ldots, (x_i, y_i)\} \subset X \times R$, where X represents the input space (e.g. $X = R^d$). SVR aims to find a regression function *f(x)* which has the greatest deviation ε of the actual targets y_i for all the training data. Regression function *f(x)* can be expressed by the following formula (Smola and Schölkopf, 2004):

$$f(x) = w \cdot \emptyset(x) + b$$

with x is input that is mapped into a feature space by a nonlinear function $\varphi(x)$. The coefficient w and b predicted by minimizing the risk function defined in equation (Smola and Schölkopf 2004):

$$\min \frac{1}{2} \|w\|^2$$

$$\text{subject to} \begin{cases} y_i - \langle w \cdot \emptyset(x_i) \rangle - b \leq \varepsilon, i = 1,2,\ldots\lambda \\ \langle w \cdot \emptyset(x_i) \rangle - y_i + b \leq \varepsilon, i = 1,2,\ldots\lambda \end{cases}$$

Assume that there is a function f that can approximate all points (x_i, y_i), with precision ε. In this case it is assumed that all existing points are feasible in the range of $f \pm \varepsilon$. In the infeasible case, there is a possibility that some points are out of the range of $f \pm \varepsilon$. The addition of slack variables ξ, ξ* can be used to solve problems of infeasible constraint in the problem optimization. Furthermore, the optimization problem can be formulated as the following:

$$\min \frac{1}{2} w^2 + C \sum_{i=l}^{\lambda} (\xi + \xi^*)$$

$$\text{subject to} \begin{cases} \boldsymbol{y_i - (w \cdot \emptyset(x_i) + b) \leq \varepsilon + \xi_i} \\ \boldsymbol{(w \cdot \emptyset(x_i) + b) - y_i \leq \varepsilon + \xi_i^*} \\ \boldsymbol{\varepsilon_i, \varepsilon_i^* \geq 0, \; i = 1, 2, \ldots, \lambda} \end{cases}$$

In the dual formulation, the optimization problem of SVR is as follows:

$$\max - \frac{1}{2} \sum_{i=1}^{\lambda} \sum_{j=1}^{\lambda} (\boldsymbol{\alpha_i + \alpha_i^*}) (\boldsymbol{\alpha_j + \alpha_j^*}) \boldsymbol{K \langle x_i, x_j \rangle} + \sum_{i=1}^{\lambda} (\boldsymbol{\alpha_i + \alpha_i^*}) \boldsymbol{y_i} - \varepsilon \sum_{i=1}^{\lambda} (\boldsymbol{\alpha_i + \alpha_i^*})$$

$$\text{subject to} \begin{cases} \sum_{j=1}^{\lambda} (\boldsymbol{\alpha_i + \alpha_i^*}) = \boldsymbol{0} \\ \boldsymbol{0 \leq \alpha_i, \alpha_i^* \leq C}, i = 1,2,\ldots\lambda \end{cases}$$

with α and α^* are the Lagrange multipliers and $\boldsymbol{K \langle x_i, x_j \rangle}$ is kernel dot-product. By using Langrange multipliers and optimality conditions, the regression function is explicitly defined as follows:

$$\boldsymbol{f(x)} = \sum_{i=1}^{\lambda} (\boldsymbol{\alpha_i + \alpha_i^*}) \boldsymbol{K \langle x_i, x \rangle + b}$$

Kernel function can solve the non-linear separable case. The kernel will be projecting the data into a high dimensional feature space to increase the computational abilities of linear learning machines. The kernel used in this study are linear, polinomial, and radial basis function (RBF). All Support Vector Regression processes are done using Library LIBSVM by Chang and Lin (2011).

In this study, grid search is applied in order to find the best model of length of rainy season prediction. The optimized parameters are the SVR parameter *C* and the kernel parameter.

The accuracy and error measurement of prediction results obtained by the SVR models use the coefficient of determination (R^2) and root mean square error (RMSE). The best model occurs when the R^2 value come near to 1 and the RMSE come near to 0. Coefficient of determination indicates the strength of relationship between two variables. Walpole (1992) formulated the coefficient of determination as follows:

$$R^2 = \frac{\left[\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Where y_i are the actual data and $\hat{y}_i$ are the predicted data.

Calculation errors using root mean square error (RMSE) defined by Walpole (1992) as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (X_t - F_t)^2}{n}}$$

Where X_t are the actual value and F_t are the predicted value.

Results and Discussion

Correlation analysis were conducted between the SOI and length of rainy season and between the DMI and length of rainy season. Both correlation analysis was conducted on each of the weather stations used in this study (Arjosari, Kebun Agung, and Pringkuku). Significance level (α) used in the present study was 10%. Table 1 show the selected months of SOI and DMI of all stations which later used as the input of the support vector regression.

Table 1 Selected SOI and DMI months

Station	SOI Months	Bulan DMI
Arjosari	April, May, July, August, September, November, December	September, October
Kebon Agung	August, September, October, November, December	September, October
Pringkuku	May, June, July, August, September, October, November, December	January, February, March, April, July, September, October, November

This study was done by making the model at each weather station using three kernels (linear, polynomial, and RBF). The best model obtained using grid search. Figure 3 show the comparison of R^2 for each kernel and each station.

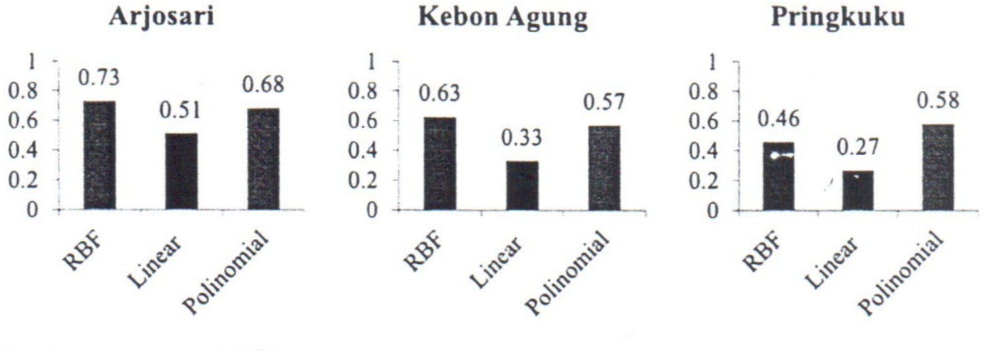

Fig. 3 Comparison of R^2 for each kernel and each station

From Figure 3, it can be seen that the best coefficient determination at Arjosari weather station obtained when RBF kernel used in the modeling with the value of R^2 is 0.73. The best coefficient determination at Kebon Agung weather station also obtained when RBF kernel used in the modeling with the value of R^2 is 0.63. On the other hand, the best coefficient determination at Pringkuku weather station obtained when polynomial kernel used in the modeling with the value of R^2 is 0.58. The figure also shows that the worst kernel based on its coefficient determination is the linear kernel. It show that modeling using linear kernel produced results which have the weakest relationship with the actual data.

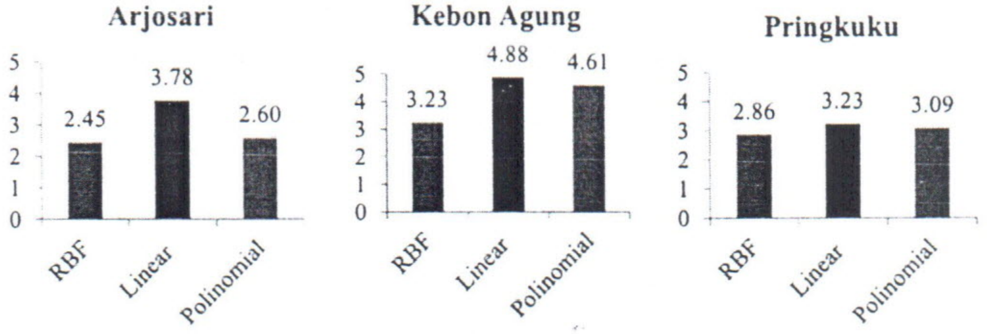

Fig. 4 Comparison of RMSE for each kernel and each station

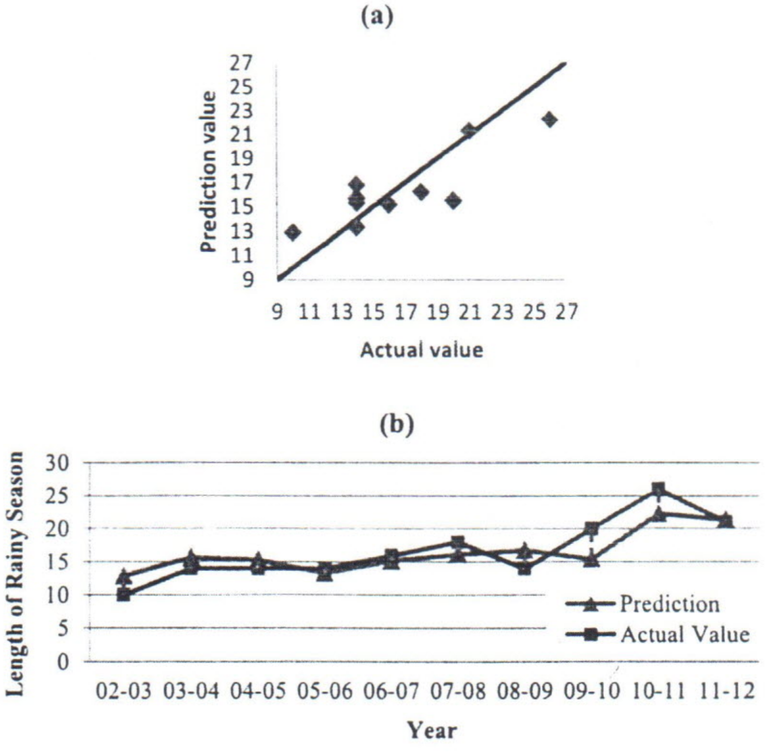

Fig. 5 Comparison between predicted value and actual value in Arjosari station using RBF kernel: (a) scatter plot and (b) line chart

Figure 4 show the comparison of RMSE for each kernel and each station. It show that the best RMSE at Arjosari, Kebon Agung, and Pringkuku weather station obtained when the model used RBF kernel, with the value of RMSE respectively are 2.45, 3.23, and 2.86. The figure show that the best kernel between the three kernel is the RBF kernel and the worst kernel is the linear kernel. It means that modeling using RBF kernel gives the least error in predicting the length of rainy season.

Figure 5 show the comparison between the length of rainy season value obtained from SVR with RBF kernel and the actual length of rainy season in Arjosari weathers station. It shows that at several periods the predictions are almost accurate while in some periods the prediction are still overestimated and underestimated.

Conclusion

The conclusion of this study is the modeling of the length of rainy season based on Southern Oscillation Index and Dipole Mode Index successfully done for each weather station. The best model at Arjosari and Kebon Agung weather station obtained when RBF kernel used in the modeling. The best accuracy values obtained are R^2 of 0.73 and RMSE of 2.45 for Arjosari weather station, whereas the best accuracy values obtained at Kebon Agung weather station are R^2 of 0.62 and RMSE of 3.23. For Pringkuku weather station, the best determination coefficient (R^2) obtained when polynomial kernel used in the modeling with the value of R^2 is 0.58. Meanwhile, the best RMSE value at Pringkuku weather station obtained when RBF kernel used in the modeling with the RMSE value is 2.86. The study also concluded that for modeling length of the rainy season, the best kernel to be used is RBF kernel while the worst kernel is the linear kernel.

Acknowledgement

The authors thank the Directorate of Higher Education, Indonesia for financial support via BOPTN scheme and the Centre for Climate Risk and Opportunity Management in Southeast Asia and Pacific, Bogor Agricultural University (CCROM-SEAP IPB), Indonesia for providing the rainfall data.

References

A. Buono, M. Mukhlis, A. Fakih, R. Boer (2012), Pemodelan jaringan syaraf tiruan untuk prediksi panjang musim hujan berdasar sea surface temperature. Seminar Nasional Aplikasi Teknologi Informasi, Yogyakarta, June, 2012.

A. Buono, I.S. Sitanggang, Mushthofa, and A. Kustiyo (2014), Time-delay cascading neural network architecture for modeling time-dependent predictor in onset prediction, Journal of Computer Science, 10(6), pp. 976-984, 2014.

A. Chandrasekar (2010), Basics of Atmospheric Science, PHI Learning Private Limited, New Delhi, India.

A.J. Clarke (2008), An Introduction to the Dynamics of El Nino & the Southern Oscillation, Elsevier, London, UK.

A.J. Smola, B. Schölkopf (2004), A tutorial on Support Vector Regression, Statistics and Computing. 14, pp. 199-222, 2004.

C. Chang, C. Lin (2011), LIBSVM : a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, 2(3), 27, 2011.

S. Behera, P. Brandt, G. Reverdin (2013), The tropical ocean circulation and dynamics, International Geophysics Series: Ocean Circulation and Climate. 103, pp. 385-404, 2013.

R.E. Walpole (1992), Pengantar Statistika (Introduction to Statistics), Gramedia, Jakarta, Indonesia.