Data Mining Techniques for Predicting Cassava Yields in Lower Northern Thailand

Anamai Na-udom[1], Jaratsri Rungrattanaubol[2]
[1]*Department of Mathematics,*
[2]*Department of Computer Science and Information Technology,*
Faculty of Science, Naresuan University, Phitsanulok, Thailand.
anamain@nu.ac.th

Abstract—**This paper investigates the factors influencing the cassava yields and develops the predictive models to predict the cassava yields in lower northern Thailand. The main objective is to compare the prediction accuracy between data mining technique namely Artificial neural network model and the conventional model namely Stepwise regression model. The root mean square error and mean absolute error values are used to validate the prediction accuracy. The results show that the significant factors are plantation area, cassava variety, cultivation period, and quantity of fertilizer. Further Artificial neural network performs better than stepwise regression model in terms of prediction accuracy. The results obtained from this study will assist farmers to improve their practices in order to increase the cassava yields.**

Index Terms—**Cassava Yields; Artificial Neural Network; Stepwise Regression Models.**

I. Introduction

Cassava (Manihot esculenta) is one of the most important economic crops in Thailand. Cassava is known as the third largest source of food carbohydrate in the tropics following rice and maize [1]. In the past cassava was usually grown in small areas around the house for household consumption such as preparing starch for making Thai desert. Since the 1990s the cassava is grown for exportation especially to the European Union (EU) and China. It has been reported by Thai tapioca starch association that the quantity of cassava exported is increasing during 2011-2014 [2]. In 2014, about 10.90 million tons of cassava products are exported [3]. Cassava products play a major role in both of agricultural sector and industrial sector as Thailand is a major cassava exporter of the world. The exports of cassava products consist of chips, pellets, and starch respectively. Cassava has been cultivated in various parts of Thailand. The total plantation area is about 11,200 to 12,800 million m^2 where the North Eastern part is the biggest cassava plantation area [4].

Though, the plantation area of cassava in Thailand is bigger than other countries such as Cambodia and Indonesia. However, it was found that the yield per rai of Thai cassava is lower than those two countries [3]. Currently, the EU requires higher amount of cassava imported from Thailand. Thai farmers are not able to increase the cassava yield due to lack of effective practices to grow cassava. Ratanawaraha et al. [5] reported that major cassava production problems in Thailand consist of declining soil fertility, soil erosion and limited genetic diversity of the crop. Hence the efficient agronomic practices such as soil preparation, planting method, planting time, quantity of fertilization, and proper spacing should be studied.

There are many cassava varieties recommended in Thailand. The most popular variety is the local variety namely Rayong 1, released in 1975. The research studies conducted by the Department of Agriculture indicated that Rayong 1 provides high-yielding of cassava and also high starch contents. Since the early 1990s, there are many varieties developed and released, such varieties adopted to plant are Kasetsart 50, Rayong 5, and Rayong 90 [6].

The cassava plantation area in lower northern Thailand consists of 7 provinces, which are Kampaeng-Phet, Nakhon-sawan, Phetchabun, Phichit, Phitsanulok, Sukhothai, Tak and Uthai-thani. This area is suitable for cassava cultivation as the landscape is very rich and moisture. The harvest area is about 2 million rai. It was reported that lower northern Thailand produces the least cassava yield comparing to other parts in Thailand. Growing cassava is very easy since it is robust to poor condition of soil and water. Hence it would be good alternative to convince farmers to grow the cassava in case of drought and lack of water occurs. Hence the challenging is to investigate the factors influencing the cassava yield so a suitable plan can be made prior to the cultivation time. This will benefit the farmers to improve the cassava yields and earn more income.

There have been a few applications of modeling methods to predict the cassava yields published in the literature so far. In the past, mathematical model such as linear regression model was used to predict the crop yield [7,8]. However, a linear relationship assumption must be assumed prior to fit the model. Hence this method might not suitable for more complex relationship between input variables and output response. Bello [9] applied a response surface methodology to modeling the cassava yield by considering some input factors such as levels of fertilizer, crop spacing, and variety of cassava. The results revealed that a constructed model is adequate and crop spacing has a direct effect on cassava yield. Other non-linear approaches such as artificial neural network (ANN) and Bayesian classification are also used to overcome the complex situation [10,11]. Various modeling methods have been used to find an accurate predictive model. For instance Ji et al. [12] compared the performance of ANN and Regression models for rice yield prediction in mountainous regions and the results showed that ANN is superior over Regression. Paswan and Begum [13] compared and discussed the performance of ANN and regression models in predicting the crop production. Raorane et al. [11] claimed that reliable and accurate forecasting techniques are required for decision making in the government office prior to pre-harvest crop. Uno et al. [14] also did a comparison between ANN and stepwise multiple linear regression models in predicting corn yield and the results showed that there was no clear difference between the two methods in terms of prediction accuracy.

This paper aims to study the factors influencing the cassava yields in the lower northern part of Thailand. All important factors are then taken to develop the predictive models. It is expected that the results obtained from this study will benefit the farmers to modify their practices in order to increase the cassava yields. A comparison of the prediction accuracy between the two popular modeling methods namely stepwise regression and artificial neural network models will be made. The method to select the input factors will be presented and then the significant factors will be brought into the predictive model. The prediction accuracy of each model is validated by using root mean square error (RMSE). In the next section, we present the research method including details of statistical models used in this study. The results based on prediction accuracy will be given in Section 3 and the conclusion is delivered in Section 4 respectively.

II. Research Methods

As mentioned before, two techniques for construction of predictive models are chosen for predicting cassava yields. One is a classical technique, which is Regression model, and the other is Artificial Neural Network (ANN), one of well-known data mining techniques. The data set of cassava cultivation and production was collected in areas of Lower Northern Thailand. First, the raw data in excel files were preprocessed. Some attributes of data were combined and transformed such as area of cassava grown transformed to a measure of rai, cassava yield to ton per rai, fertilizer usage to kilogram per rai and so on. Originally the raw data set contained 1,928 records. After that, the data cleaning process to improve the quality of the data is performed by getting rid of records with a missing value or misreported and some outliers. After the data preprocessing, the total number of records in the data set is 1,802. The important factors influencing the cassava yield are analyzed using nonparametric correlation analysis. The important factors are listed with details in Table 1.

Table 1
The selected significant input variables and output

Variable name	Meaning	Range	*r*	P-value
Province	Area of growing	{1, 2, 3, 4, 5, 6, 7} Kampaeng-Phet, Nakhon-sawan, Phetchabun, Phichit, Phitsanulok, Sukhothai, Tak and Uthai-thani	0.240	<0.000
Varieties	Cassava varieties	{1, 2, 3, 4, 5, 6, 7} Rayong 5, Rayong 60, Rayong 90, Kasetsart 50, HuaiBong 60, Rayong 7, HuaiBong 80	-0.224	<0.000
Period	Period of cultivation, months	Min=5 Max=24 Avg.=11.65	0.210	<0.000
FertQ	Fertilizer usage, kg per rai	Min=2.53 Max=200 Avg.=36.08	0.009	<0.000
Yield	Cassava yield, Ton per rai	Min=0.5 Max=9.26 Avg.=3.83		

Then the key factors are applied to fit Regression and ANN models. The process of constructing the predictive models based on these two techniques will be explained in section 2.3. The prediction accuracy is validated with RMSE and MAE criteria. This section presents the details of two techniques applied for the cassava predictive models and how to implement the predictive models with good accuracy.

A. Regression Model

Regression analysis is one of the most effective methods that have been successfully used in the context of yield prediction since it is simple to construct and provides information on input variables sensitivity [15]. This method is based on the assumption of random error arising from a large number of insignificant input factors. Given an output response, y, and input variables$= (x_1, \dots, x_d)$, the relationship between y and x can be mathematically written as:

$$y = f(x) + \varepsilon \tag{1}$$

where ε is a random error which is assumed to be normally distributed with mean zero and variance σ^2. Since the true response surface function $f(x)$ is unknown, a response surface $g(x)$ is created to approximate $f(x)$. Therefore the predicted values are obtained by using $\hat{y} = g(x)$, which $g(x)$ can be treated as a polynomial function of $(X_1, X_2, \dots, X_d)$. The observed data set can be expressed in the matrix form using the data matrix X as:

$$y_0 = X\beta + \varepsilon \tag{2}$$

where $y_0 = (y_1, y_2, \dots, y_n)^T$, x is a $n\times\alpha$ design matrix, β is a $(\alpha\times 1)$ vector of the regression coefficients, and ε is a $(n\times 1)$ vector of random error. The number of unknown parameters in equation (2) is determined by α, where $\alpha = 2d + \binom{d}{2} + 1$. The vector of least squares estimators $\hat{\beta}$, can be determined subject to the minimization of:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = (y_0 - X\beta)^T (y_0 - X\beta) \tag{3}$$

Minimization of equation (3) yields:

$$X^T X\hat{\beta} = X^T y_0 \tag{4}$$

Hence, the least squares estimator of β is:

$$\hat{\beta} = (X^T X)^{-1} X^T y_0 \tag{5}$$

provided that $(X^T X)$ is invertible.

Once β is estimated, equation (5) can be used to predict the cassava yield value at any untried settings of input variables.

B. Artificial Neural Network Model

Artificial neural network (ANN) was designed to mimic the complex learning systems of the human brain, which consists of millions of closely interconnected sets of neurons. ANN has been commonly used in various areas such as machine learning, image recognition and complex decision making systems [16, 17]. A basic artificial neuron model consists of a set of inputs (p_i), a combination function or summation function ($\sum$) and activation function (*f*). A set of input is combined by a summation function with a set of weight (w_i) corresponding to each input (p_i). In technical term, this summation point is normally referred as a node, and each node is assigned a bias to it (i.e. b_1). The output from a node is passed to an activation function (*f*), and then eventually an output response (*y*) is obtained as shown in Figure 1.

An activation function is a nonlinear function, which the one used here is a sigmoid function.

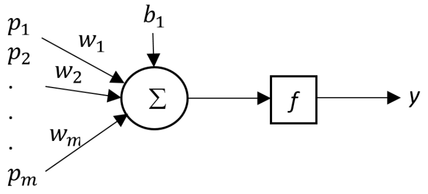


Figure 1: A basic artificial neuron model

The processing unit of ANN presented in Figure 1 can be formulated as (6)

$$y = f(w_i p_i + b_1) \tag{6}$$

A structure of ANN typically consists of an input layer, a hidden layer and an output layer. In general there may be more than one hidden layer; however one hidden layer is often sufficient enough. In this study we only used one hidden layer. Every node in a layer is connected to every node in the next layer as depicted in Figure 2.

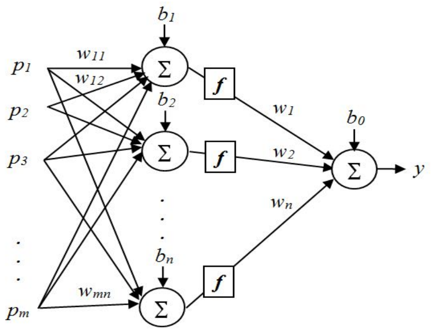


Figure 2: A structure of ANN [18]

ANN in Figure 2 contains *m* inputs, one hidden layer with *n* number of nodes and one output y.

Each input p_i in an input layer is completely connected to each node ($\sum$) in a hidden layer. The weights between each node are adjusted by back propagation method. The process of weight adjustment is controlled by two parameters, namely learning rate and momentum rate. Learning rate influences how large the weight adjustment should be and momentum rate influences the current adjustment to move in the same direction as previous.

The entire processing unit of ANN can be rewritten as:

$$y = \left[\sum\nolimits_{i=1}^{n} w_i \times (f(\sum\nolimits_{i=1}^{n} \sum_{j=1}^{n} w_{ij} p_i + b_i))\right] + b_0 \tag{7}$$

where *n* is a number of nodes in a hidden layer and *m* is a number of inputs.

C. Model accuracy measurement

The accuracy of predictive models is evaluated on the basis of RMSE and MAE values. The formula of RMSE and MAE are defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{k}(y_i\text{-}\hat{y}_i)^2}{k}} \tag{8}$$

$$MAE = \frac{1}{k}\sum_{i=1}^{k}\left|\frac{y_i\text{-}\hat{y}_i}{y_i}\right| \tag{9}$$

where k is the number of test points, y_i is the actual response of the i^{th} test point and $\hat{y}_i$ is the predicted response from the predictive models for the i^{th} test point. Lower values for RMSE and MAE imply a more accurate predictive model.

D. Predictive model construction

After applying the spearman rank correlation coefficient (*r*) as a criterion to select input variables for the model, the input variables that are statistically related to the cassava yield at the significance level of 0.05 are presented in Table 1. There are 4 input variables for the predictive models and one output, which is a cassava yield measured in terms of ton per rai. The data set of 1,802 records is used to construct the predictive model based on 5-fold cross validation method. The fold cross validation technique is used to assess how the results of the prediction accuracy analysis will generalize to an independent data set. It is mainly used in fitting where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

The data set of 1,802 is partitioned into 5 folds (fold0, fold1, fold2, fold3, fold4). With 5-folds, there are 5 predictive models to be constructed for each technique. The training set is formed by 4 out of 5 folds and the left fold is the test set. The training set (seen data) is used to construct the predictive model and the test set (unseen data) is used to evaluate the model. RMSE values of the training set and the data set are recorded to present the accuracy of the predictive models. The training set$_i$ and the test set$_i$ are defined as (10),

$$Training\ set_i = \sum_{j=1}^{4} fold_{(i+j) mod\ 5} \tag{10}$$

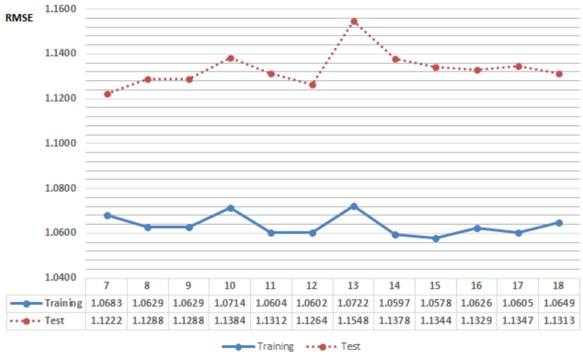
$$Test\ set_i = fold_i \tag{11}$$

where $0 \leq i \leq 4$

The regression model is constructed by transforming two categorical input variables namely province and varieties using dummy variables. The stepwise technique is used here to select the best model. According to 5-fold evaluation, we repeatedly created the model five times using SPSS for Windows. The average RMSE values for training and test set are 1.12992 and 1.13695 respectively and the most accuracy model is obtained from Training set$_2$.

The process of constructing a regressive model is quite simple; in contrast constructing ANN models is more complicated since ANN has some parameters to be considered when fitting a model. The parameters in this study are learning rate, momentum rate and number of nodes used in a hidden layer. Similarly to a regression model, the two category input variables, namely province and varieties has been defined as binary input in order to obtain better predictive ANN models. Hence, 4 input variable originally, now the total input node for ANN is 16.

In this study we varied learning rate with values of 0.1, 0.2 and 0.3 and momentum rate with 0.1 and 0.2. The first experiment is to investigate the most effective values of these two parameters by creating and evaluating the models by combining these set of values on 9 to 11 nodes, which is 90 models in total. As a result, we found that the most accuracy model (i.e. with less average RMSE value) were obtained from learning rate and momentum rate with a value of 0.1and 0.2.



	7	8	9	10	11	12	13	14	15	16	17	18
Training	1.0683	1.0629	1.0629	1.0714	1.0604	1.0602	1.0722	1.0597	1.0578	1.0626	1.0605	1.0649
Test	1.1222	1.1288	1.1288	1.1384	1.1312	1.1264	1.1548	1.1378	1.1344	1.1329	1.1347	1.1313

Figure 3: Average RMSE values of ANN models with various number of nodes used

Then, we considered a number of nodes used in a hidden layer by varying number of nodes from 7 to 18 nodes with a value of 0.1and 0.2 for learning rate and momentum rate respectively. We repeatedly reproduced the model for 60 times and calculated average RMSE values based on 5-folds. As depicted in Figure 3, the model with 15 nodes on average has the best accuracy, i.e. the least RMSE. All ANN models were implemented by WEKA [19] with 16 input nodes and varied number of nodes from 7 to 18. The initial weights are randomly generated with seed number of 0 and backpropagation is a learning technique to adjust weights. The learning process is repeated for 500 times. The dataset can be downloaded from https://sites.google.com/a/nu.ac.th/advmining/download.

III. Results and Discussion

After the best model of each method is found, the most prediction accuracy Regression model is obtained from Training set$_2$ and the most accuracy ANN model is obtained from the model with 15 nodes, which is formed from Training set$_4$ and Test set$_4$. Their RMSE and MAE values are presented in Table 2.

Table 2
RMSE and MAE of the best accurate models from Regression and ANN

Methods	Training data		Test data	
	RMSE	MAE	RMSE	MAE
Regression	1.11350	0.28688	1.20464	0.27734
ANN	1.06900	0.25749	1.05840	0.26523

It can be clearly seen from Table 2 that ANN performs better than regression as the RMSE and MAE values obtained from both of training and test set are lower than that of regression model. For RMSE criterion, the percentage improvement of ANN over regression model is about 4% in the case of training set and it increases up to 12% in test set. When MAE values are considered, the percentage improvement is about 10% and 4% for training and test set respectively. As the structure this data set is quite complex and most of the input factors are qualitative data, hence it is not unusual to observe that the assumption free approach like ANN is more accurate than regression model.

IV. Conclusion

This paper presents the method to investigate the influence of various input factors on the cassava yield. The performance of the two popular predictive models, stepwise regression and ANN for predicting cassava yield in the lower northern Thailand is also presented. The results on training and predicting the cassava yield reveal that ANN performs better than stepwise regression model. The advantage of ANN model is that it is flexible to set all related parameters. Furthermore, ANN is robust to different structure of complex data. Hence ANN model would be recommended to use for modeling the cassava yield especially when the priori information of factors under study is unknown.

Acknowledgement

This research project was fully supported and sponsored by the National Research Council of Thailand under grant R2558B054 at Naresuan University.

References

[1] Poramacom, N. Ungsurattana, A. Ungsurattana, P. Supavititpattana, P. 2013. Cassava Products, Prices and Related Policy in Thailand. *American International Journal of Comtemporary Research*. 3(5): 43-51.

[2] Thai Tapioca Starch Association, http://www.thaitapiocastarch.org

[3] Office of Agricultural Economics, http://www.oae.go.th

[4] Usubharatana, P. Phungrassami, H. 2015. Carbon Footprint of Cassava Starch Production in North-Eastern Thailand. *Procedia CIRP*. 29: 462-467.

[5] Ratanawaraha, C. Senanarong, N. Suriyapan, P. 2000. Status of cassava in Thailand: Implications for future research and development. Proceedings of the Validation Forum On The Global Cassava Development Strategy. 3:63-102.

[6] Kaweewong, J. Kongkaew, T. Tawornprek, S. Yampracha, S. and Yost, R. 2013. Notrogen requirements of cassava in selected soils of Thailand. *Journal of Agriculture and Rural Development in the Tropics and Subtropics*. 114(1): 13-19.

[7] Zhange, G. et al. 1998. Predicting with artificial neural network. *International Journal of Predicting*, 14: 35-62.

[8] Kaul, M., Hill, R. L., Walthall, C. 2005. Artificial Neural Network For Corn And Soybean Prediction. *Agricultural System*, 85: 1-18.

[9] Bello, A. O. 2014. Modeling Cassava Yield: A Response Surface. *International Journal on Computation Sciences and Applications*. 4(3): 61-70.

[10] Khairunniza-Bejo, S., Mustaffha, S. and Ismail, W. I. W. 2014. Application of Artificial Neural Network in Predicting Crop Yield: A Review. *Journal of Food Science and Engineering* 4: 1-9.

[11] Raorane, A. A. and Kulkarni, R. V. 2013. Review-Role of Data Mining in Agriculture. *International Journal of Computer Science and Information Technology*. 4(2): 270-272.

[12] Ji, B. Sun, Y. Yang, S. and Wan, J. 2007. Artificial Neural Networks For Rice Yield Prediction In Mountainous Regions. *The Journal of Agricultural Science*. 145(3): 249-261.

[13] Paswan, R. P. and S. A. Begum. 2013. Regression and Neural Networks Models for Prediction of Crop Production. *International Journal of Scientific & Engineering Research*. 4(9): 98-108.

[14] Uno, Y., Prasher, R., Laeroix, R., Goel, P.K., Karimi, A. and Viau, et al. 2005. Artificial Neural Network To Predict Corn Yield From Compact Airborne Spectrographic Imager Data. *Computers and Electronics in Agriculture*, 47. 149-161.

[15] Montgomery, D. C., Peck, E. A. and Vining, G. G. 2012. Introduction to Linear Regression Analysis. Fifth Edition, John Wiley & Sons, New Jersey.

[16] Bozdogan, H. 2003. Statistical Data Mining and Knowledge Discovery. Chapman & Hall/CRC, New York.

[17] Ripley, B.D. 1993. Statistical aspects of neural networks. In: Barndoff-Nielsen, O.E., Jensen, J.L., Kendall, W.S., editors. Networks And Chaos-Statistical And Probabilistic Aspects, Chapman & Hall, New York. 40-123.

[18] Na-udom, A. and Rungrattanaubol, J. 2015. A Comparison of Artificial Neural Network and Regression Model for Predicting the Rice Production in Lower Northern Thailand. Information Science and Applications, *Lecture Notes in Electrical Engineering*. 339:745-752.

[19] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. 2009. The WEKA Data Mining Software: An Update; SIGKDD Explorations. 11.