Prediction of Tangerine Yield Using Artificial Neural Network (ANN)

Pichaya Boonprasom[1] and Gumpanart Bumroongitt[2]*

[1]*Department of Food Engineering, Faculty of Agro-Industry, Chiang Mai University, Chiang Mai 50200, Thailand*

[2]*Electricity Generating Public Company Limited (EGCO), Bangkok 10210, Thailand*

**Corresponding author. E-mail: gumpanart.bum@egco.com*

ABSTRACT

This work was aimed at studying and investigating the application of an Artificial Neural Network (ANN) to forecast the tangerine yield, taking into account some influential factors, especially the weather conditions. Data and information relating to tangerine yield of Chiang Mai Province from 1992 to 2001 were collected. Weather data such as average monthly rainfall and average monthly temperature of relevant periods were also obtained. Five ANN models, with different input factors or different structures, were constructed for three sections of the experiment. The results indicated that the ANN had high potential and ability to forecast tangerine yield accurately despite small set of data available. The amount of rainfall was observed to have strong influence and contributed important informations to the ANN in forecasting tangerine yield.

Key words: Tangerine, Artificial Neural Network (ANN), Forecasting, Simulation

INTRODUCTION

Agricultural statistics of 2002 showed that 36,000 rai of cultivated land in Chiang Mai Province were utilized to produce 11,400 tons of tangerine which have become the most important commercial fruit crop of the province. Among many varieties of tangerine, Sai Nam Pung is the most famous and tasteful one. Fang District of Chaing Mai is the leading Sai Nam Pung producer with peak harvest period from September through December. The performance of production predictions is an important aspect in order to establish a reasonable and coherent policy in the citrus sector. In addition, tangerine producers and traders will be able to use detailed and timely yield predictions to optimize purchasing and selling activities in order to prevent marketing problems resulting from the imbalance between production and market possibilities of consumption.

The application of the artificial neural network (ANN) approach to forecasting task has gained a great deal of interest from researchers in many fields of study for quite sometime. The main advantage of the ANN approach is that it does not require any assumption of any functional relationship between its input variables and the corresponding output. Instead, it is able to learn and build its own non-linear model from a relationship between input variables and output during the training process. Thus, high human expertise is not needed and the difficulty in modeling process is not encountered anymore. It is widely accepted that

the application of ANN is quite appropriate when the functional relationship between input variables and output is complex. Several researchers have reported the attempt to apply the ANN approach in the forecast of agricultural yield (Davies and Albrigo, 1994; Haskett et al., 1998; Camago et al., 1999; Jackson and Looney, 1999; Shearer et al., 1999). The majority of the ANN model used in the previous researches for forecasting is a multi-layered feedforward ANN using backpropagation learning algorithm which is of a supervised learning type. One very important task in the study using ANN is the selection of the training set and input variables, since the performance of the ANN relies heavily on the data presented to it. The training set and the input variables should contain the significant information or the principal relationship contributing to the desired output variables.

The Artificial Neural Network used in this study is a Multi-Layered feedforward network using a backpropagation learning algorithm. In particular, the architecture of the ANN chosen here consists of three layers: an input layer, a hidden layer and an output layer as shown in Figure 1.

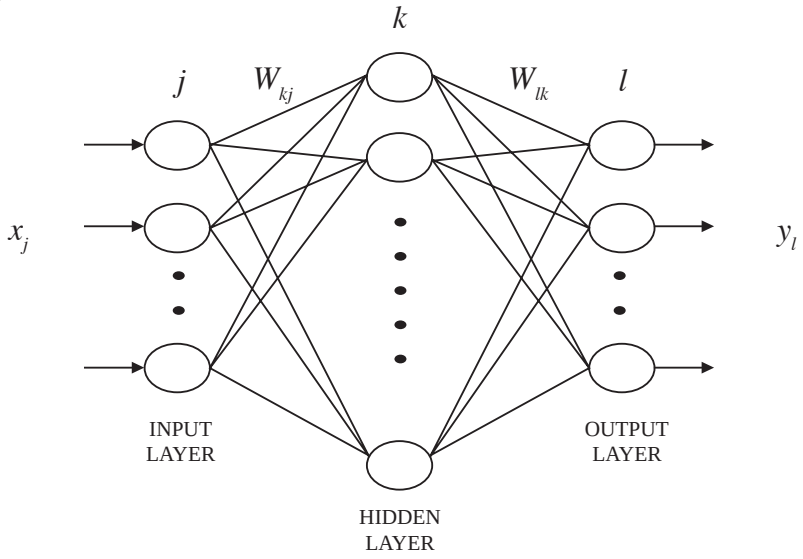

Figure 1. Three-Layered Artificial Neural Network.

The work in this report is aimed at studying and investigating the important factors that influence tangerine yield, especially the weather conditions and the potential of ANN to perform the forecasting task of tangerine yield. It could lead to the way for tangerine production planning and marketing, relevant data collection planning or some influential controlling factors, which would benefit the agriculturist in the future.

COMPUTER MODELING AND SIMULATION

Data

Data and information relating to tangerine yield were collected from various sources in order to construct training set and testing set for ANN learning process. The scope of the data is limited within Chiang Mai area since it is the major production area of tangerine in northern part of Thailand. In general, there would be several major factors which have strong

influence on tangerine yield such as weather, soil fertilization and tangerine genetic information. However, only the weather data were selected and studied as a main factor affecting tangerine yield in this experiment. The following informations were collected for the experiment in this report.

- Tangerine yield and producing area of 12 districts in Chiang Mai Province from 1992 to 2001 were obtained from the Agricultural Department of Chiang Mai. Those districts included Fang, Mae Ai, Chiang Dao, Mae Wang, Samoeng, Mae Taeng, Mae Rim, Jom Thong, Doi Saked, Sansai, Chaiprakarn and Phrao. Only 6 out of 12 districts contained data for all years between 1992 to 2001.
- Average monthly rainfall of 7 districts in Chiang Mai Province from 1992 to 2002 was obtained from the Meteorological Department. Those districts included Fang, Mae Ai, Chiang Dao, Mae Wang, Samoeng, Mae Taeng and Phrao.
- Average monthly temperature of Chiang Mai Province from 1992 to 2002 was obtained from the Meteorological Department.
- Average monthly relative humidity of Chiang Mai Province from 1992 to 2002 was obtained from the Meteorological Department.

Structure of the ANN

The ANN used in this study is a Multi-Layered feedforward network, using a backpropagation learning algorithm. In particular, the architecture of the ANN chosen here consists of three layers: an input layer, a hidden layer and an output layer. Several ANN models with different input factors and different number of neurons in a hidden layer were experimented. The neural network software Qnet v2000 package by Vesta Services, Inc. was used as an ANN tool in this experiment.

The experiment was carried out in 3 sessions to cover the determination of ANN's potential to predict the tangerine yield. The sessions includedd

Session 1 ANN Model A, B and C were used and the suitable set of input data was observed.

Session 2 ANN Model B, D and E were used and the effect of the number of neurons in a hidden layer was observed.

Session 3 ANN Model D and DD were used and the effect of the information resource was observed.

ANN Model A

Model A has 6 neurons in input layer, 2 neurons in hidden layer and 1 neuron in output layer. A data record consists of 6 inputs and one output. The inputs include average monthly rainfall of June, July and August of the current year, producing area of the previous year, tangerine yield of the previous year and producing area of the current year. The output is tangerine yield of the current year.

ANN Model B

Model B has 6 neurons in input layer, 2 neurons in hidden layer and 1 neuron in output layer. A data record consists of 6 inputs and one output. The inputs include average monthly rainfall of August, September and October of the current year, producing area of the previous year, tangerine yield of the previous year and producing area of the current year. The output is tangerine yield of the current year.

ANN Model C

Model C has 9 neurons in input layer, 2 neurons in hidden layer and 1 neuron in output layer. A data record consists of 6 inputs and one output. The inputs include average monthly rainfall and temperature of August, September and October of the current year, producing area of the previous year, tangerine yield of the previous year and producing area of the current year. The output is tangerine yield of the current year.

ANN Model D

Model D has 6 neurons in input layer, 3 neurons in hidden layer and 1 neuron in output layer. A data record consists of 6 inputs and one output. The inputs include average monthly rainfall of August, September and October of the current year, producing area of the previous year, tangerine yield of the previous year and producing area of the current year. The output is tangerine yield of the current year.

ANN Model E

Model E has 6 neurons in input layer, 6 neurons in hidden layer and 1 neuron in output layer. A data record consists of 6 inputs and one output. The inputs include average monthly rainfall of August, September and October of the current year, producing area of the previous year, tangerine yield of the previous year and producing area of the current year. The output is tangerine yield of the current year.

Training Set and Testing Set

The training set and testing set were created according to input of each ANN model. Since Fang and Mae Ai are the two major districts of tangerine production in Chiang Mai while other districts produce only small amount of tangerine under small area; therefore, only the data from Fang and Mae Ai were utilized. Despite of being one of the major districts for tangerine production, Mae Ai also had small amount of production in the early years. Therefore the total number of data selected for the experiment was 13. The other corresponding data for each ANN model such as amount of rainfall, average monthly temperature were selected to build the training set and testing set according to the structure of each model. It was observed that data set for the experiment contained wide variation of tangerine yield, ranging from below 10,000 tons to nearly 100,000 tons. This helped in the evaluation of the potential of the ANN in forecasting tangerine yield.

From the thirteen data records selected, one data record was set aside for testing purpose, called testing set, and this data record was not presented to ANN during the training session but it was presented to the ANN after training session had been completed. Hence, the total number of data record in a training set was only twelve. For the purpose of the experiment, each data record had equal chance and was assigned as a testing set once. By performing this, potential of ANN for forecasting at every level of yield was evaluated. Therefore, thirteen ANN's of each model were trained and tested.

Each ANN was trained until the Root Mean Square (RMS) error decreased and reached 0.01 or the number of iteration exceeded 100,000, when the training session stopped. Then the testing set was presented to the trained ANN in order to determine the accuracy of the forecast. The measurement for the accuracy of each testing data record was Absolute

Percentage Error (APE), as shown in Equation (1). The forecast was done for thirteen testing data records and Mean Absolute Percentage Error (MAPE), as stated in Equation (2), of those results was also determined.

$$APE = (|t_i - y_i| / ti) x 100 \quad (1)$$

$$MAPE = (1 / N_T) \sum APE \quad (2)$$

where t_i is actual yield, y_i is the predicted output of the ANN, and N_T is the total number of tested data.

RESULTS AND DISCUSSION

Results of the Experiment Session 1

ANN Model A, B and C were completed in the training session and then tested. These models have different input factors but have relatively similar network structure where the hidden layer consists of two neurons. The forecast performance of each ANN model is presented in Figure 2 and Table 1. From the figure, it can be demonstrated that the ANN of each model is able to perform a forecasting task in a satisfactory manner. Despite of wide variation of tangerine yield used in the experiment, ranging from below 10,000 tons to nearly 100,000 tons, the forecast outputs trace the actual production very well for all level of yields.

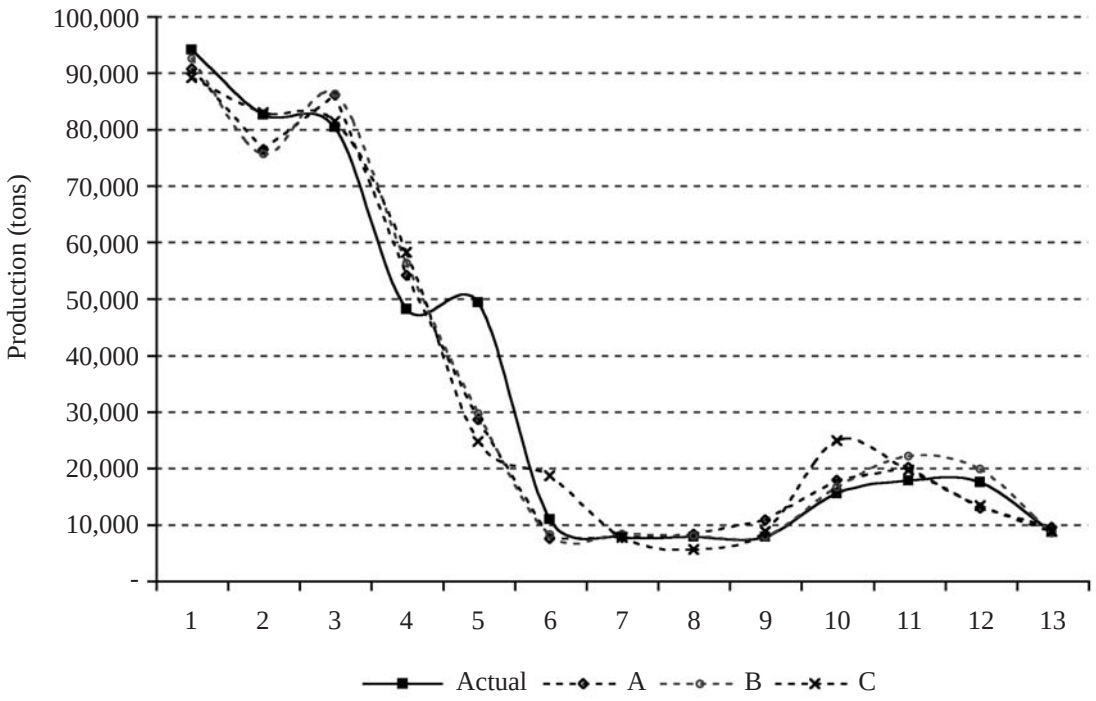

Figure 2. Forecast performance of ANN Model A, B and C.

Table 1. Forecast performance of ANN Model A, B and C.

	ANN Model		
	A	**B**	**C**
MAPE (%)	16.50	12.18	22.02
Average Error (tons)	4,534	4,122	5,174

From Table 1, Model B has superior performance over the other two. Its MAPE and average error in tons are lowest. On the contrary, Model C shows the poorer performance than the other two models. Figure 3 shows the comparison of forecasting error in APE sorted in ascending order between each model. As seen from the error profile of each model, it is confirmed that Model A and B can perform better than Model C.

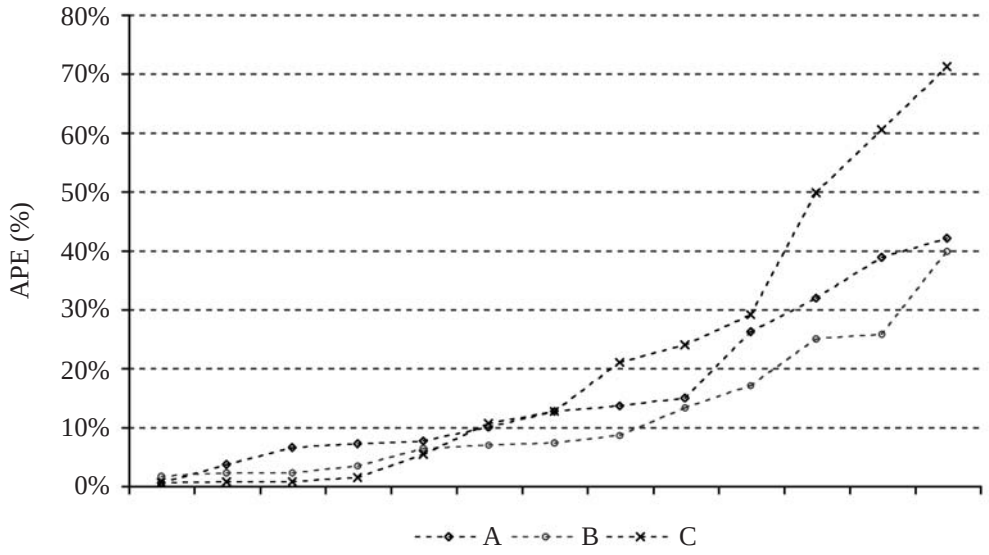

Figure 3. APE Comparison of ANN Model A, B and C, sorted in ascending order.

The above observation when comparing the performance of Model A and B to that of Model C, it shows that the amount of rainfall has strong influence on yield of tangerine while the average temperature has much less influence. However, as indicated earlier that temperature data available were from only a single location at Chiang Mai International Airport which could be different from the areas of tangerine production used in the training set. Therefore the use of average temperature in this particular experiment does not contribute positive effect to the forecast.

Furthermore, when comparing between Model A and B, the results are fairly in the same trend although the period of rainfall data used as input was shifted slightly. This result is consistent with our knowledge that the amount of water provided to tangerine is very crucial to the yield. Model B, which uses rainfall data in the period of August to October, has shown superior performance. In case when more data are available in the future, longer period may be used as an input for the ANN.

Results of the Experiment Session 2

From the previous experiment, it was found that the most accurate forecasting was obtained by using Model B. Therefore Model D and E were developed, based on the same input pattern as that of Model B, to study the effect of number of neurons in a hidden layer on the accuracy of the forecast. ANN Model D and E were completed in the training session and then tested in the same manner as the previous experiment.

The forecast performance of each ANN model is presented in Table 2. Figure 4 shows the comparison of forecasting error in APE sorted in ascending order between each model.

Table 2. Forecast performance of ANN Model B, D and E.

	ANN Model		
	B	**D**	**E**
Neurons in a hidden layer	2	3	6
MAPE (%)	12.18	13.14	16.34
Average Error (tons)	4,122	3,861	4,409

From Table 2, Model B shows that it can perform best in terms of MAPE. The MAPE of the forecast increases as the number of neurons in a hidden layer increases. The ANN tends to memorize the training set rather than generalizing it and that would affect the accuracy of forecasting the unseen data. It has been discussed that the number of neurons in a hidden layer tend to relate with size of a training set, number of input factors and complexity of the problem. Nevertheless, the proper number of such neurons could be found through experimental method.

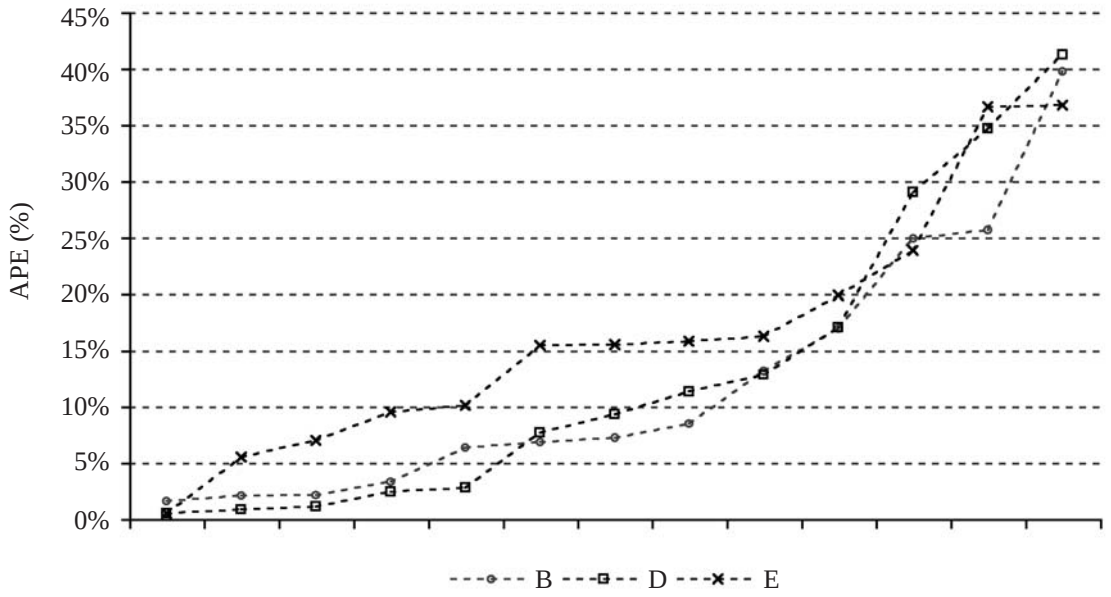

Figure 4. APE Comparison of ANN Model B, D and E, sorted in ascending order.

The experiments in this section reveal that Model B and D, which have smaller number of neurons in their hidden layer, have better performance than Model E (Table 2). This could be the result of small size of training set used in this experiment. If larger set of data and more useful parameters are available in the future, the training set may be larger and contain higher number of input factors, then the number of neurons in a hidden layer may have to be increased to gain higher accuracy of the forecast.

However, it should be noted that Model D has the lowest average error in tons while its maximum error in tons is highest. This leads to the observation that the rest of forecast errors in Model D, other than the highest one, could be fairly low. From detailed examination of the forecast, it was found that the highest forecast error in tons occurred when the forecast was performed on the data record number 5. This very-high forecasting error on the data record number 5 occurred for all ANN models.

It was found that data record number 5 belonged to the production year 1997 of Fang District. There was abrupt increase of yields between year 1996 and 1997 when tangerine yield jumped from 10,876 tons to 49,140 tons, or approximately 350% increase. It showed that the ANN could not adjust itself to learn such kind of sudden change as can be seen in Figure 2 that the forecasts fell short of the actual yield. This is one point that should be investigated in future research for more effective technique of ANN training to cope with the sudden change in yield of tangerine. It is possible that some other factors should be investigated and included into the ANN input.

Consequently, the MAPE of Model B, D and E were calculated if all the forecast errors, except for that of data record number 5, were taken into consideration. The MAPE of Model B, D and E were 9.89%, 10.79% and 14.66% respectively. The average error in tons of those models were 2,839 tons, 2,494 tons and 3,278 tons. This observation will lead to the experiment in session 3.

Results of the Experiment Session 3

To check the effect of data record number 5, Model D was selected for another new training session. Data record number 5 was deleted from a training set and Model DD, as called in this experiment section, was trained and tested in the same manner as in the previous experiment (meaning that Model DD, which has the same structure as Model D, was trained with the training set from which data record number 5 was deleted).

Table 3. Forecast performance of ANN Model D and DD.

	ANN Model		
	D	**D**	**DD**
Consideration of record #5	Yes	No	No
MAPE (%)	13.14	10.79	11.96
Average Error (tons)	3,861	2,494	2,690

From the results shown in Table 3 and Figure 5, Model DD shows a better forecasting performance than the original Model D. Nevertheless, when the performance of Model D is

evaluated without the forecast of data record number 5, it can be seen that Model D generally outperforms Model DD. Such observations indicate that with one less data in the training set, the ANN has less resource of knowledge to learn and its performance could deteriorate, even though the missing data was the one which yielded high forecasting error.

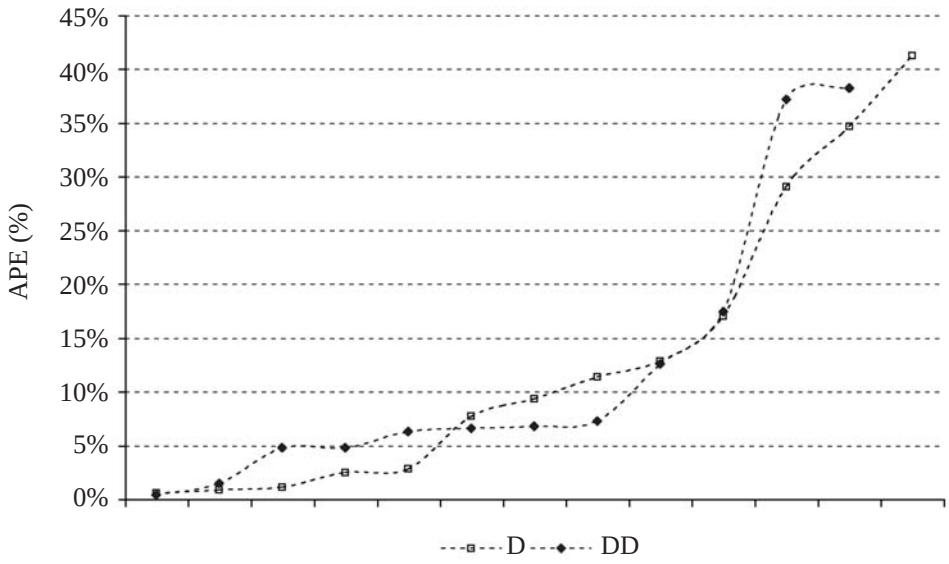

Figure 5. APE Comparison of ANN Model D and DD, sorted in ascending order.

In this particular experiment, it indicated that there was no need to discard the data record which had sudden change since it helped contribute knowledge during training the ANN. Correspondingly, this leads to better performance of the ANN in general, except when forecasting such particular data record where the ANN has difficulty to adjust itself. The techniques for improving ANN learning for such kind of data would be worthwhile to investigate in the future work.

CONCLUSION AND RECOMMENDATIONS

From the results of the experiment, ANN shows its potential and ability to forecast tangerine yield quite accurately. Despite small set of data available, ANN can be trained to perform forecasting. The forecast outputs trace the actual production very well for all level of tangerine yields in those years used in the experiment. The MAPE's of the forecast by several ANN models ranged approximately from 12% to 22%. Then the appropriate ANN model was selected for further experiments.

It is observed from the experiment that the amount of rainfall has strong influence on yield of tangerine, while the average temperature has less influence and does not contribute positive effect to the forecast in this particular experiment. This could be the results of several factors such as the small set of data used for the experiment or the fact that the temperature data available were from a single location at Chiang Mai International Airport, which is not quite relevant to the production areas in the experiment. If temperature data are

available for each district that has tangerine production, the use of temperature may improve the accuracy of the forecast.

The experiment shows that the number of neurons in a hidden layer of an ANN could result in accuracy of the forecast. It can be observed that as a number of neurons in a hidden layer increases, the MAPE of the forecast also increases. It has been discussed that the number of neurons in a hidden layer tend to relate with size of a training set, number of input factors and complexity of the problem. Nevertheless, the proper number of such neurons could be found through experimental method.

The ANN shows that it has some difficulty to learn and adjust itself to data that has the nature of sudden change. It would be interesting to investigate in future research any kind of effective technique of ANN training to cope with the sudden change in yield of tangerine. It is possible that some other important factors may be investigated and included into the ANN input, for instance soil fertilization. At the same time, the experiment indicates that the data which has sudden change also helps contribute knowledge during training the ANN and it is not necessary to discard such data from the training set.

Finally, it can be concluded that the ANN has high potential for the application of tangerine yield forecast according to several advantages. Further investigation, especially the influential factors, data collection or training technique would definitely help improve the performance of the ANN in tangerine yield forecast.

ACKNOWLEDGEMENT

This research would not have been possible without the funding from Chiang Mai University. We would like to thank Chiang Mai University for the support.

REFERENCES

Camago, M., A.A. Ortolani, M.J. Pedro Jr., and S.M. Rosa. 1999. Agrometeorological model for yield prediction of orange cultivar Valencia. Bragantia 58(1): 171–178.

Davies, F.S.,and L.G. Albrigo. 1994. Citrus. CAB International, Wallingford, Oxon, UK. 254 p.

Jackson, D.I., and N.E. Looney. 1999. Temperate and subtropical fruit production. Second edition. CAB International, Wallingford, Oxon, UK. 332 p.

Haskett, J., Y. Pachepski, and B. Acock. 1998. Use of Artificial Neural Networks to simulate soybean crop yield as affected by global change. Tektran, United States Department of Agriculture, Agriculture Research Service.

Shearer, S.A., T.F. Burks, J.P Fulton, S.F. Higgins, J.A. Thomasson, T.G. Mueller, and S. Samson. 1999. Yield prediction using a Neural Network classifier trained using soil landscape features and soil fertility data. ASAE Paper No. 993041. St. Joseph, Michigan.