Title	タイのキャッサバ輸出予測に応用した時系列データ予測のハイブリッドモデルに関する研究
Author(s)	Pannakkong, Warut
Citation	
Issue Date	2017-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/14821
Rights	
Description	Supervisor:Huynh Nam Van, 知識科学研究科, 博士

A Study on Hybrid Models for Prediction of Time Series Data with Application to Forecasting Thailand's Cassava Export

Warut Pannakkong

Japan Advanced Institute of Science and Technology

A Study on Hybrid Models for Prediction of Time Series Data with Application to Forecasting Thailand's Cassava Export

by

Warut Pannakkong

Submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Associate Professor Van-Nam Huynh

School of Knowledge Science
Japan Advanced Institute of Science and Technology

September, 2017

Abstract

Time series forecasting is an active research area that plays important role in planning and decision making in several practical applications. The main task of this research area is to improve the prediction accuracy.

This thesis proposes three novel hybrid forecasting models which are significantly extended from the Zhang's model and the Khashei and Bijari's model by involving the clustering algorithm (i.e. k-means) and the discrete wavelet transform (DWT) for inputs pre-processing. Additionally, instead of including only the lagged values of time series as the input variables, additional variables such as moving averages and annual seasonal index are included into the proposed model. The experiments are conducted comprehensively with several hybridization scenarios in term of structures and variables to find the most suitable forecasting model for Thailand's cassava export.

The first proposed hybrid model (so called ARIMA+ARIMA/ANN+k-means/ARIMA /ANN) is the hybrid forecasting model involving autoregressive integrated moving average (ARIMA), artificial neural network (ANN) and the k-means clustering. These single models and the k-means clustering are used to build the forecasting models in different level of complexity (i.e. ARIMA; hybrid model of ARIMA and ANN; and hybrid model of k-means, ARIMA, and ANN). To obtain the final forecasting value, the forecasted values of these three models are combined with the weights generated from discount mean square forecast error (DMSFE) method.

The second proposed hybrid model (so called DWT/ARIMA/ANN) is the hybrid forecasting model of the DWT, the ARIMA, and the ANN without linear or nonlinear assumption on the approximation and the detail. The proposed model starts with decomposing the time series by the DWT to get the approximation and the detail. Then, the approximation and the detail are separately analyzed by the Zhang's model involving the ARIMA and the ANN in order to capture both linear and nonlinear components of the approximation and the detail. Finally, the linear and nonlinear components are additively combined for the final forecasting value.

These two novel hybrid models are applied to three well-known data sets: Wolf's sunspot, Canadian lynx, and exchange rate (British pound to US dollar) to evaluate the prediction capability in three measures (i.e. MSE, MAE, and MAPE). The prediction performance of the proposed models is compared to both the traditional single and hybrid models. The results imply that the proposed models give the best performance in MSE, MAE, and MAPE for all three data sets.

Then, the proposed hybrid models are implemented to Thailand's cassava export as the case study. In addition, we also propose the third novel hybrid model (so called ARIMA/ANN with pre-processed variables), which is the hybrid model of the ARIMA and the ANN with pre-processed variables for Thailand's cassava export forecasting. The experimental results indicate that the DWT/ARIMA/ANN model is the best model for the native starch and the sago. On the other hand, the ARIMA/ANN with pre-processed variables model is the best model for the modified starch.

In conclusion, all three proposed hybrid models have shown their forecasting capability over both the traditional single and hybrid models. Therefore, they can be used as the alternative models for time series prediction. Moreover, the stakeholders involves in the cassava supply chain can apply the proposed models specified for each type of the cassava export to obtain more accurate prediction results of Thailand's cassava export. The proposed models for cassava forecasting can be applied to other commodity products sharing the similar characteristic with the cassava as well.

Keywords: hybrid time series forecasting model, autoregressive integrated moving average (ARIMA), artificial neural network (ANN), k-means, discrete wavelet transform (DWT)

Acknowledgments

I would like to express the deepest respect and appreciation to my supervisor, Associate Professor Van-Nam Huynh. Without his guidance and persistent help, this thesis would not have been possible.

I am sincerely grateful to Professor Takashi Hashimoto, who is my second supervisor, for his kind help and insightful suggestions.

I also would like to express my sincere thanks to Professor Michitaka Kosaka for his guidance and kind support for my minor research project.

I wish to express my sincere glatitude to Professor Yoshiteru Nakamori, Professor Youji Kohda, Professor Tsutomu Fujinami, Professor Tetsuya Murai, and Associate Professor Takaya Yuizono for being the committee members for my defense. Their critical comments fulfill the weak points and significantly enhance my research.

I also grateful to my colleagues and friends in Huynh-Lab for their help and encouragement in doing research.

I would like to thank JAIST staffs for their kind support and service during three years at JAIST.

Last but not least, I would like to express my gratitude to my family for always understanding and supporting me with love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Time series forecasting is an important research area which has attracted a lot of attention from research communities in numerous practical fields including finance [2, 3], agriculture [4, 5], energy [6, 7], transportation [8, 9], environment [10, 11], etc.

Over the past several decades, many attempts have been made by researchers for the development of efficient forecasting models to continuously improve the forecasting accuracy but the research for developing new forecasting models to improve the prediction accuracy has never stopped.[12].

Cassava (*Manihot esculenta Crantz*) is a main source of calories, after rice and maize, for the world's population particularly in developing countries. The report of Food and Agriculture Organization of the United Nations in 2012 has indicated that the cassava is the ninth rank in term of production quantity. In addition, animal foods and ethanol production for alternative energy industries use the cassava as a raw material [13]. Most available cassava in the international market is exported from Thailand. However, currently, Thailand's cassava export forecasting relies on only a simple and traditional forecasting model.

The rest of this chapter provides statement of problem, research objective, and chapter organization.

1.1 Statement of problems

Recently, a hybridization of forecasting models, so-called hybrid model. The hybrid model can be a combination of same or different models. One of the most popular hybrid model category is linear and nonlinear hybrid model. For instance, a combination between autoregressive integrated moving average (ARIMA) and artificial neural network (ANN) has been introduced by Zhang [14] in order to take advantage of unique strength of the ARIMA and the ANN in linear and nonlinear modeling. The experimental results indicated that the Zhang's model can be an effective approach to improve forecasting accuracy rather than using either the ARIMA or the ANN separately. However, a drawback of the Zhang's model is assuming additive relationship between linear and nonlinear components of the time series. Thus, in some circumstances, the Zhang's model underperformed its components (i.e. the ARIMA and the ANN) [15].

Furthermore, regarding the mixed results of the Zhang's model, Khashei and Bijari [16] extended the Zhang's model by defining time series as a function of linear and nonlinear components. In the Khashei and Bijari's model, firstly, the linear component is extracted from the time series by the ARIMA. Then, they assumed that the nonlinear component still remains in the residuals of the ARIMA and the time series data. Secondly, the ANN is applied to investigate the function of the results of the ARIMA, the lagged of ARIMA residuals, and lagged values of the time series. In fact, the both hybrid models (Zhang's model and Khashei and Bijari's model) are applied in several real-world applications such as stock index [17], agricultural commodity price [18], soil water content [19], value of agricultural imports [20], irrigation water demand [21], sugar and alcohol [22], and Goldman Sachs Commodity Index (GSCI) futures price [23]. Usually, these hybrid models promise better prediction, but only in the average of accuracy measure. Thus, in some prediction periods, the single models can give higher accuracy. Moreover, the input variables are only lagged values of the time series. According to these limitations, there is a potential to improve the forecasting accuracy with a proper combination of single and hybrid models, as well as inputs preprocessing.

To our best knowledge, the researches of the cassava export forecasting are limited to using the ARIMA model [24, 25], and in [26], we have applied the ANN model for cassava export forecasting. So far, there is no study of a hybrid model originated for Thailand's

cassava export forecasting, in spite of the cassava is one of the most important source of calories for the world's population, after rice and maize [13], and Thailand is the first rank cassava exporter in the world. The cassava export quantity from Thailand influences cassava trading in international market such that the Thailand's future cassava export quantity can support decision makers involved in the cassava supply chain to improve production planning, policies making for helping cassava farmers, profit of cassava trading in futures market, etc.

For these reasons, it seems interesting to consider the Thailand's cassava export as a case study in this thesis. Hence, we propose a novel hybrid forecasting model for the Thailand's cassava export. The proposed model is significantly extended from the Zhang's model and the Khashei and Bijari's model by involving the clustering algorithm (i.e. k-means [27]) and the wavelet transform [28] for inputs preprocessing. Additionally, instead of including only the lagged values of time series as the input variables, additional variables such as moving averages and annual seasonal index are included into the proposed model. The experiments are conducted comprehensively with several hybridization scenarios in term of structure and variables to find the most suitable forecasting model for the cassava export.

1.2 Research objectives

- To improve accuracy of time series prediction by developing novel hybrid forecasting models
- To develop hybrid models for Thailand's cassava export forecasting
- To recommend forecasting models for Thailand's cassava export forecasting

1.3 Chapter organization

- **Chapter 1** provides introduction of time series forecasting and its applications, statement of problems, research objectives, and the chapter organization of this thesis.

- **Chapter 2** introduces background and literature review of autoregressive integrated moving average (ARIMA), artificial neural network (ANN), k-means clustering algorithm, discrete wavelet transform (DWT), hybrid models, and forecasting accuracy measures.

- **Chapter 3** presents the hybrid model of k-means clustering, ARIMA and ANN for time series forecasting. This hybrid model improve prediction accuracy by applying the k-means clustering to classify time series before further analyses.

- **Chapter 4** describes the hybrid model using discrete wavelet transform to convert time series into approximation and detail. Then, hybrid ARIMA and ANN model is applied to them separately.

- **Chapter 5** explains the hybrid models for Thailand's cassava export forecasting. The hybrid models in Chapter 3 & 4 are implemented for this prediction. In addition, another hybrid model of ARIMA and ANN with pre-processed variables for the cassava export is proposed as well.

- **Chapter 6** includes the thesis contribution involving practical implication, theoretical implication, and contribution to Knowledge Science.

- **Chapter 7** contains conclusion and future works.

Chapter 2

Background and literature review

In this chapter, a background and literature review of time series forecasting are described to present fundamental knowledge related to the content in this thesis. First, traditional time series forecasting models (e.g. autoregressive integrated moving average (ARIMA) and artificial neural network (ANN)) are introduced. Second, k-means clustering algorithm and discrete wavelet transform (DWT), which are techniques used for time series pre-processing, are presented. Third, hybrid forecasting models of ARIMA and ANN are introduced. Forth, prediction accuracy measures used to evaluate the forecasting models in this thesis are presented.

2.1 Autoregressive integrated moving average (ARIMA) model

The ARIMA [29] is a popular statistical model for forecasting both stationary and non-stationary time series during several past decades. Typically, this model is an integration of autoregressive (AR) and moving average (MA) models, including data transformation term called differencing (I).

The formulation of the ARIMA comes up with autoregressive moving average (ARMA) model, which is a special case of the ARIMA as:

$$Z_t = c + \sum_{i=1}^{p} \phi_i Z_{t-i} + a_t - \sum_{j=1}^{q} \theta_j a_{t-j} \tag{2.1}$$

The ARMA model predicts a time series at period t (Z_t) by using lagged values of time series ($Z_{t-1}, \ldots, Z_{t-p}$) and lagged random errors ($a_{t-1}, \ldots, a_{t-q}$) where ϕ_i and θ_j are the model parameters; c is a constant; a_t is a random error with mean of zero and a constant variance of σ^2; p and q are orders of the lagged values included in the model.

In order to simplify the mathematical formula, backward shift operator (B), defined as $B^i Z_t = Z_{t-i}$, is substituted for the ordinary algebraic symbol in (2.1), thus, the ARMA model can be formulated as (2.2) below:

$$Z_t = c + \sum_{i=1}^{p} \phi_i Z_t B^i + a_t - \sum_{j=1}^{q} \theta_j a_t B^j \tag{2.2}$$

Then, by rearranging the terms related to Z_t in (2.2), we obtain the ARMA model as (2.3).

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right) Z_t = c + \left(1 - \sum_{j=1}^{q} \theta_j B^j\right) a_t \tag{2.3}$$

which is compactly rewritten as:

$$\phi_p(B) Z_t = c + \theta_q(B) a_t \tag{2.4}$$

where

$$\phi_p(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$$

$$\theta_q(B) = 1 - \sum_{j=1}^{q} \theta_j B^j$$

are called the autoregressive operator ($\phi_p(B)$) and the moving average operator ($\theta_q(B)$), respectively.

Note that, nevertheless, the ARMA model has no capability to deal with non-stationary time series, in this case, the differencing is required to transform the non-stationary into stationary time series by substitute $(1 - B)^d Z_t$ for Z_t in (2.4), where d is the degree of differencing. Then, we can obtain the ARIMA as follows:

$$\phi_p(B)(1 - B)^d Z_t = c + \theta_q(B) a_t \tag{2.5}$$

In a situation where there is seasonality in time series, the seasonal model components such as seasonal autoregressive operator ($\Phi_P(B^s)$) and seasonal moving average operator ($\Theta_Q(B^s)$) respectively defined by:

$$\Phi_P(B^s) = 1 - \sum_{k=1}^{P} \Phi_k B^{ks}$$
$$\Theta_Q(B^s) = 1 - \sum_{l=1}^{Q} \Theta_l B^{ls}$$

are included in (2.5) in order to capture the relationship of the seasonality. Then, the seasonal ARIMA, or SARIMA, can be formulated as:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Z_t = c + \theta_q(B)\Theta_Q(B^s)a_t \tag{2.6}$$

where s is the time span for a season, P is the seasonal order of autoregressive, Q is the order of seasonal moving average, and D is the degree of seasonal differencing. In addition, the seasonal time span can be determined by autocorrelation analysis. The time span giving the highest and significant autocorrelation is considered as the seasonal time span.

However, the ARIMA model has several limitations due to the linearity assumption which is hard to be fully satisfied in real-world applications, or the use of only historical time series as the model's inputs.

2.2 Artificial neural network (ANN)

The artificial neural network (ANN), which is a kind of artificial intelligent technique mimicking biological neurons mechanism, is a well-known tool in time series forecasting in term of usage flexibility because there is no assumption on its inputs and it has self-learning ability as human brain neurons [30].

The structure of the ANN, graphically depicted in Fig. 2.1, consists of nodes which are located in three types of layers: input layer; hidden layer; and output layer. Usually, there is only one input layer and one output layer, while the number of the hidden layer can be more than one. Nevertheless, the ANN with one hidden layer is capable to approximate any continuous function [31].

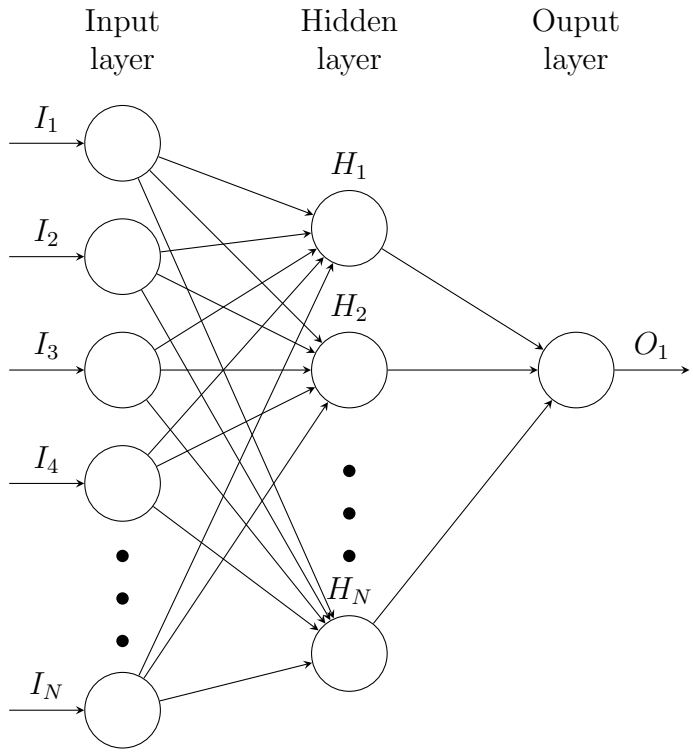

Figure 2.1: Feed-Forward Artificial Neural Network [1]

In general, the numbers of nodes and layers depend on experience and level of understanding in the problem of the architect because so far, there is no theory for selecting the best parameters for the ANN. Hence, a fashionable approach is trial and error until obtaining the appropriate parameters [32, 30].

At each node (Fig. 2.2), the relationship of the inputs and the output can be expressed by:

$$\begin{aligned} a_v^l &= f\left(\sum_{u=1}^{U} a_u^{l-1} w_{uv}^l + b_v^l\right) \\ &= f(n_v^l) \end{aligned} \tag{2.7}$$

where U denotes the total number of nodes in layer $l-1$, $l \leq L$ is the integer representing the currently considered layer, L denotes the total number of layers in the ANN, and f denotes the transfer function.

For instance, at node v in layer l, the inputs (a_u^{l-1}) which are the output from nodes $u = 1, \ldots, U$ in the previous layer $(l-1)$ are aggregated by weights (w_{uv}^l) and combined with the bias (b_v^l) to generate the net input (n_v^l). Then, the net input (n_v^l) is passed

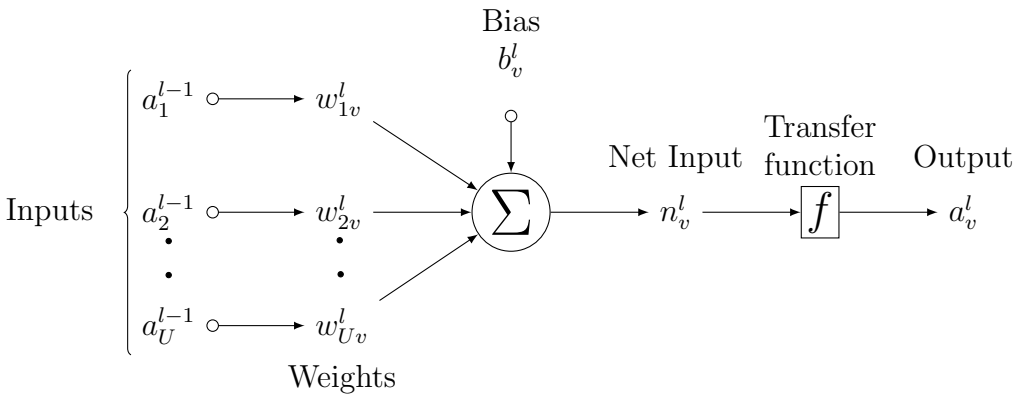

Figure 2.2: A Node in Artificial Neural Network [1]

through the transfer function (f) to compute the output (a_v^l) which will become the input for the next layer.

After the output of the ANN is generated from the output layer, the output is compared to the target in order to measure prediction accuracy. The ANN can learn for minimizing the forecasting error via an training algorithm attempting to determine weight (w_{uv}^l) and bias (b_v^l) that fit to relationship between the inputs and the target.

In this thesis, a feed-forward ANN [1] is applied. The training algorithm is Levenberg-Marquardt algorithm with Bayesian regularization [33]. The following sections explain how the structure of the ANN is identified.

Input layer

The input layer consists of the input nodes representing input variables (e.g. historical time series). Theoretically, the input variables can be divided into two types: technical variables and fundamental variables [32]. The technical variables are lagged values (e.g. time series value at time $t-1$) or processed values (e.g. moving average and seasonal index) of time series. The fundamental input variables are other variables (e.g. month in year) believed that there is existing relationship between them and the dependent variable (e.g. future time series).

Output layer

The output layer is the layer producing the results of the ANN by aggregating the outputs from the hidden layer as:

$$
\begin{aligned}
a_k^l &= f_{lin}\left(\sum_{j=1}^{J} a_j^{l-1} w_{jk}^l + b_k^l\right) \\
&= f_{lin}(n_k^l)
\end{aligned}
\tag{2.8}
$$

where j and k are nodes in the hidden layer (previous layer) and the output layer respectively, a_j^{l-1} is the output from node j in the hidden layer, w_{jk}^l is the weight between node j of the hidden layer and node k of the output layer, b_k^l is the bias of node k, n_k^l is the net input of node k and a_k^l is the output of node k. In order to obtain the output (a_k^l), the net input (n_k^l) is passed through the pure linear transfer function (f_{lin}), then the outputs (a_k^l) are stored in a memory for further analyses.

Hidden layer

The hidden layer, the layer located between the input and the output layers, generates the outputs by aggregation of the outputs from the previous layer with the weights, the biases, and the transfer function which is usually nonlinear, as the following:

$$
\begin{aligned}
a_j^l &= f_{tan\text{-}sig}\left(\sum_{i=1}^{I} a_i^{l-1} w_{ij}^l + b_j^l\right) \\
&= f_{tan\text{-}sig}(n_j^l) \\
&= \frac{2}{1+e^{-2n_j^l}} - 1
\end{aligned}
\tag{2.9}
$$

where i and j are nodes in the previous layer and the hidden layer respectively, a_i^{l-1} is the output from node i in the previous layer, w_{ij}^l is the weight between node i of the previous layer and node j of the hidden layer, b_j^l is the bias of node j, n_j^l is the net input of node j. The output of the hidden layer at node j (a_j^l) is the result of substituting the net input (n_j^l) into the tan-sigmoid transfer function ($f_{tan\text{-}sig}$), then, the output (a_j^l) is adopted as the input of the next layer.

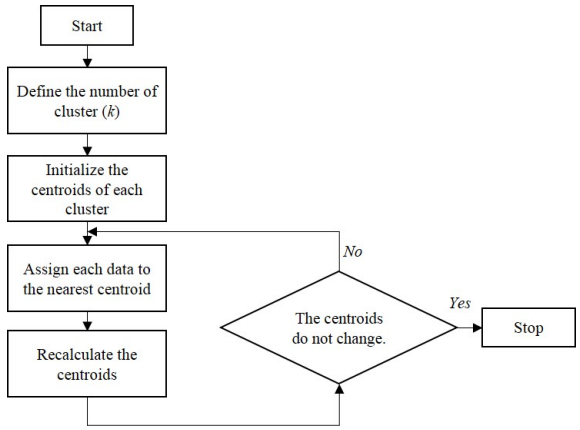

Figure 2.3: k-means clustering algorithm

2.3 k-means clustering algorithm

The k-means algorithm is an unsupervised learning algorithm commonly used for grouping the data set into k clusters [27]. The algorithm starts with defining the number of cluster (k) and the position of centroids for each cluster. Second, each data point is assigned to the nearest centroid using the Euclidian distance (d_{ij}) as below:

$$d_{ij} = \sqrt{\sum_{v=1}^{V} (x_{iv} - c_{jv})^2} \tag{2.10}$$

where x_{iv} is the value of attribute v of the data i, and c_{jv} is the value of the attribute v of the centroid of the cluster j.

Third, the centroids are recomputed. Then, the second and the third steps are repeated until all centroids do not move anymore.

A problem of the k-means algorithm is selecting an optimal number of clusters (k). To find the optimal k, the Silhouette [34] can be used as a measure for clustering quality evaluation; it can be computed as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{2.11}$$

where a_i is the average dissimilarity of data i to all other data within the same cluster, b_i is the minimum dissimilarity of data i to all data in the other clusters. According to (2.11), the higher value of s_i implies the better matching between the data and its cluster. Thus, the number of cluster (k) giving the highest average of s_i for the whole data set is

considered as the optimal number of cluster (k).

2.4 Discrete wavelet transform (DWT)

The wavelet transform is a technique analyzing both time and frequency of signals simultaneously [28]. This technique decomposes an input signal into two parts: low frequency information (approximation) and high frequency (detail) by using low and high frequency pass filters. For multiple decomposition level, the approximation of the previous decomposition level is the input of higher decomposition level. Actually, there are two main types of the wavelet transform: continuous and discrete wavelet transforms. However, in practical, the time series are discrete and suitable to be decomposed by the discrete wavelet transform (DWT) as:

$$\begin{aligned} Z_t &= A_J(t) + \sum_{j=1}^{J} D_j(t) \\ &= \sum_{k=1}^{K} c_{J,k}\phi_{J,k}(t) + \sum_{j=1}^{J}\sum_{k=1}^{K} d_{j,k}\psi_{j,k}(t) \end{aligned} \tag{2.12}$$

where Z_t denotes the time series at period t; $A_J(t)$ denotes the approximation of the highest decomposition level (J); $D_j(t)$ denotes the detail of decomposition level j; $c_{j,k}$ and $d_{j,k}$ denote the coefficient of approximation and detail respectively, at decomposition level j and period k; $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ denote low (approximation) and high (detail) pass filters respectively, at decomposition level j and period k; K denotes the total number of time series J denotes the total level of decomposition.

2.5 Hybrid models

Traditionally, the ARIMA is one of the most popular statistical techniques for time series forecasting because of its ability in dealing with both stationary and non-stationary time series. The ARIMA is suitable for linear modeling and has the linearity assumption of the time series property [29]. Actually, the linearity assumption of time series made in the ARIMA is difficult to meet in many practical applications. To overcome this drawback,

numerous nonlinear forecasting models have been developed in literature; among them the ANN has received increasing attention for nonlinear time series forecasting due to its benefits of being considered as a universal function approximator and data-driven model without any prior assumptions on the time series [31].

The ANN has been applied in several practical situations, and comparative studies between the ANN and the ARIMA have been conducted as well. In most cases, the comparative results show that the ANN has better prediction accuracy than the ARIMA; nonetheless, there are still some cases that the ARIMA can outperform the ANN [30].

The mixed results implied that neither the ARIMA nor the ANN is the best model for all problems. Particularly, it seems to be inappropriate to use the ARIMA to fit nonlinear data, and apply the ANN model to linear data as well.

Moreover, it is difficult to clearly identify exact properties (e.g. linear or nonlinear) of the data from real world applications. In fact, pure linear and pure nonlinear time series rarely exist in the real world [14].

In practical, there is no single forecasting model giving the best performance in any situation. To decrease the risk of selecting inappropriate model, currently, hybrid models, which can be an integration of either same or different type of the single models [15], have been proposed to enhance prediction accuracy by taking unique capabilities of the single models in a complementary manner, such as combinations of the ARIMA and the ANN that has capability in linear and nonlinear modeling, respectively [14, 16].

The hybrid models of the ARIMA and the ANN, which have capability in both linear and nonlinear modeling, use the ARIMA to capture linear component from the time series, and then the ANN model is used to capture nonlinear component from residuals of the ARIMA.

The hybridization has been proved that it can give a better prediction accuracy than applying individual models alone in several applications. Faruk [35] proposed a hybrid ANN and ARIMA for water quality time series prediction. Pai and Lin [36] developed a hybrid ARIMA and support vector machines (SVM) model for stock price forecasting. Chun-Ling Lin and Shyu [37] presented an application of hybrid multi-model forecasting system involving ARIMA and ANNs focusing on market demand of display market in Taiwan. He et al. [38] proposed a hybrid ARIMA and SVM for short term load forecasting

in Hebei province, China. Bouzerdoum et al. [39] also proposed a hybrid model of seasonal autoregressive integrated moving average (SARIMA) and SVM for short-term power forecasting but it is for a small-scale grid-connected photovoltaic plant.

In day-ahead electricity price forecasting, Zhang et al. [40] introduced a new hybridization of ARIMA and least squares support vector machine (LSSVM) for Australian national electricity market. Shafie-Khah et al. [41] presented a hybrid model based on ARIMA and radial basis function neural networks (RBFN) for mainland Spain electricity price. These two papers used wavelet transform to decompose the time series and applied particle swarm optimization (PSO) to optimize the models. Furthermore, Chaâbane [42] developed a hybrid model of autoregressive fractionally integrated moving average (ARFIMA) and neural network model for electricity price prediction in Nordpool, Norway.

Chen and Wang [43] proposed a hybrid SARIMA and SVM to predict the production value of Taiwan machinery industry. Chen [44] Combining linear and nonlinear model using ARIMA, ANN and SVM in forecasting tourism demand of Taiwanese outbound tourists. Aslanargun et al. [45] compared performance of ARIMA, ANN and their hybrid models in forecasting amount of monthly tourists visiting Turkey.

Lo [46] conducted a study of applying ARIMA and SVM models to predict the number of failure in software execution. Ruiz-Aguilar et al. [47] presented hybrid approaches based on SARIMA and ANN to forecast the number of inspection goods in customs and border controls at the Port of Algeciras Bay in Spain, which is the top ten of Europeans ports.

Shi et al. [48] assessed capability of hybrid approaches of ARIMA, ANN and SVM models in forecasting wind speed and power in Colorado. Cadenas and Rivera [49] developed a hybrid ARIMA and ANN model to predict wind speed in three districts of Mexico. Díaz-Robles et al. [50] proposed a hybrid model of ARIMA and ANN models in order to forecast particulate matter occurring in metropolitan using Temuco in Chile as the case study. Zhu and Wei [51] presented a novel combination of ARIMA and LSSVM for carbon price forecasting. Barak and Sadegh [52] introduced ARIMA-ANFIS hybrid algorithm, combination of ARIMA and ensemble adaptive neuro fuzzy inference system (ANFIS), to forecast energy consumption in Iran.

The rest of this section describes the formulation of the Zhang's model and the Khashei

and Bijari's model. These two hybrid models are further extended in order to improve time series prediction accuracy as demonstrated in the next chapters.

2.5.1 Zhang's model

Zhang [14] firstly invented a hybridization of the ARIMA and the ANN in purpose of obtaining unique capability in linear and nonlinear modeling of these two single models. The concept of this approach is applying ARIMA to capture linear component ($\hat{L}_t$) from the time series at first. Then, the residuals of the ARIMA results are considered as the source of the nonlinear component ($\hat{N}_t$) which is captured by the ANN. Finally, the linear and nonlinear components are aggregated to predict the future time series ($\hat{Z}_t$) as:

$$\hat{Z}_t = \hat{L}_t + \hat{N}_t \tag{2.13}$$

The linear component ($\hat{L}_t$) is the result of the ARIMA in predicting the actual time series (Z_t). In addition, the ARIMA residuals (e_t) are defined as:

$$e_t = Z_t - \hat{L}_t \tag{2.14}$$

In case of the nonlinear component ($\hat{N}_t$), it is the outcome of applying the ANN to the ARIMA residuals (e_t) as following:

$$\hat{N}_t = f(e_{t-1}, e_{t-2}, \ldots, e_{t-n}) \tag{2.15}$$

where f is a nonlinear function fitted by the ANN, n is total lagged periods included in the model.

However, the Zhang's model assumed additive relationship between linear and nonlinear components; such assumption cannot be satisfied in all circumstances.

2.5.2 Khashei and Bijari's model

Khashei and Bijari [16] developed a new hybridization of the ARIMA and the ANN, representing time series as a function of linear and nonlinear components to overcome the additive relationship assumption of the Zhang's model. The Khashei and Bijari's model

can be formally expressed by:

$$\hat{Z}_t = f(\hat{L}_t, \hat{N}_t) \tag{2.16}$$

where $\hat{L}_t$ and $\hat{N}_t$ denotes the linear and nonlinear components, respectively.

There are two stages in this hybrid approach. In the first stage, the linear component ($\hat{L}_t$) can be estimated from the results of the ARIMA. The residuals of the ARIMA (e_t) can be determined as:

$$e_t = Z_t - \hat{L}_t \tag{2.17}$$

where $\hat{L}_t$ denotes the result of the ARIMA and Z_t denotes the time series at time t.

In the second stage, the purpose is to capture the nonlinear components from the residuals of both the ARIMA ($\hat{N}_t^1$) and the time series ($\hat{N}_t^2$) as described in (2.18) and (2.19) respectively below:

$$\hat{N}_t^1 = f^1(e_{t-1}, \ldots, e_{t-n}) \tag{2.18}$$

$$\hat{N}_t^2 = f^2(Z_{t-1}, \ldots, z_{t-m}) \tag{2.19}$$

where f^1 and f^2 are the nonlinear functions determined by the ANN, n and m are integers representing the number of maximum previous periods included in the model.

Finally, the linear and the nonlinear components are combined as:

$$\begin{aligned} Z_t &= f(\hat{L}_t, \hat{N}_t^1, \hat{N}_t^2) + \epsilon \\ &= f(\hat{L}_t, e_{t-1}, \ldots, e_{t-n_1}, z_{t-1}, \ldots, z_{t-m_1}) + \epsilon \end{aligned} \tag{2.20}$$

where f is the nonlinear function fitted by the ANN, $n_1 \leq n$ and $m_1 \leq m$ are integers that are determined in the design process, and ϵ is the error term.

2.6 Forecasting accuracy measures

In order to measure the forecasting accuracy, three popular measures, namely mean square error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE), are computed as in (2.21), (2.22) and (2.23) respectively. Let Z_t denotes the actual value at period t, $\hat{Z}_t$ denotes the forecasted value at period t and N denotes amount of total forecasting period. The smaller value of these measures implies the better prediction accuracy.

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(Z_t - \hat{Z}_t)^2 \tag{2.21}$$

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}|Z_t - \hat{Z}_t| \tag{2.22}$$

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\frac{|Z_t - \hat{Z}_t|}{Z_t} \tag{2.23}$$

Chapter 3

Hybrid model of k-mean clustering, ARIMA, and ANN for time series forecasting

3.1 Introduction

Traditionally, the linear model such as the autoregressive integrated moving average (ARIMA) dominated other methods because the ARIMA is capable to deal with non-stationary time series as well as stationary time series. Nevertheless, the ARIMA made a prior assumption on relationship between the historical and the future time series as a linear function which is very difficult to be satisfied in practical situations [29].

The artificial neural network (ANN), nonlinear model mimicking human brain neurons mechanism, became popular due to their advantages over the ARIMA. The ANN can fit the relationship between historical and future values without prior assumption, and they can be considered as an universal approximator for any continuous function [30, 53]. Comparative studies in several applications were done, and the results implied that the ANNs usually outperformed the ARIMA [54, 55, 56, 57, 58, 59]. Recently, the ANN has been improved by including the predicted values and the residuals of the ARIMA as the inputs in order to obtain unique strength in linear and nonlinear modelling (so called ARIMA/ANN model) [16] but there is no guarantee that it can capture all of the linear pattern as the ARIMA because the nature of the ANN is the nonlinear model.

Currently, clustering techniques (e.g. self-organizing map (SOM) and k-means clustering) have been applied for time series preprocessing in purpose of classifying the time series that have the similar statistical distribution into the same cluster in order to capture non-stationary pattern before using the ANN to forecast the future values for each cluster [60, 61, 62]. This approach can increase the forecasting accuracy because the ANN focuses on the pattern of the data in each cluster instead of considering the whole data set, however, this can lead to the overfitting problem. In addition, even though the time series is well separated into the clusters, we cannot actually know the cluster of the future values. In this case, the suitable way to use the predicted values generated from each cluster is the problem that should be concerned.

A recent work proposed to use the summation of the prediction values from every cluster [61], but in fact, it would be more logical if the future values are produced from the ANN dedicated to their cluster. However, the cluster of future values is unknown. For this reason, the ANN can be selected based on the predicted cluster of the future values. In order to predict the cluster of the future values, firstly, the boundaries separating the different clusters are identified. Then, the predicted values are plotted, and their cluster can be determined according to the boundaries. In this case, the clustering technique providing the straight clear cut boundary between the clusters such as the k-means clustering is required. Therefore, k-means/ARIMA/ANN model is developed.

Normally, hybridization of the forecasting models have been constructed by combination of single models. The hybrid models can outperform the single models, but only in the average of prediction performance. The single models can still give the better results than hybrid models in some forecasting period. Moreover, the forecasting models having difference of complexity (e.g. single, hybrid, and hybrid with clustering) may extract different information from the time series. Thus, there is a potential to improve prediction accuracy if we can properly combine such models together.

This chapter presents a novel hybrid forecasting model that takes advantage of unique strength of the three approaches such as linear modeling, hybrid linear-nonlinear modeling, and hybrid linear-nonlinear modeling with clustering. In addition, a new method for using the predicted values from each cluster to predict the actual future values is proposed as well. The rest of this chapter is organized as follows. The proposed hybrid model is

described in section 3.2. In section 3.3, the experiments are explained and the results are interpreted. Finally, the conclusions are given in section 3.4.

3.2 Proposed model

In this section, the proposed model is presented (Fig. 3.1). The objective of the proposed model is to take unique advantages of the different forecasting models in term of type (i.e. linear and nonlinear) and complexity (i.e. single and hybrid).

The proposed model consists of four stages: I) linear modeling, II) linear-nonlinear modeling, III) linear-nonlinear modeling with clustering, and IV) final forecasting. At stage I, the linear modeling stage uses the ARIMA to extract linear component from time series. At stage II, the linear-nonlinear modeling stage applies Khashei and Bijari's model [16] which is the ANN with the inputs: linear component (results of the ARIMA), residuals of the ARIMA, and lagged values of the time series. The model of the stage II can be referred as ARIMA/ANN model. At stage III, the linear-nonlinear modeling with clustering stage performs k-means clustering algorithm to classify time series into different clusters. Then, the ARIMA/ANN model is built for each cluster. Meanwhile, the cluster of one-step ahead time series is predicted. The results of this stage come from the ARIMA/ANN model of which the predicted cluster. The model of the stage III can be called k-means/ARIMA/ANN model. At stage IV, the final forecasting is done by aggregation of the results from the previous three stages with the weights generated by discount mean square forecast error (DMSFE). The detail of these four stages are explained stage by stage, and the aggregation of the results of stages I, II, and III in the final forecasting is described.

3.2.1 Stage I: the linear modeling stage

The ARIMA model is applied to the whole data set to obtain predicted values ($\hat{L}_t$) and their residuals (e_t). The results are passed through the final forecasting stage and also used as the inputs for the ARIMA/ANN and the k-means/ARIMA/ANN models.

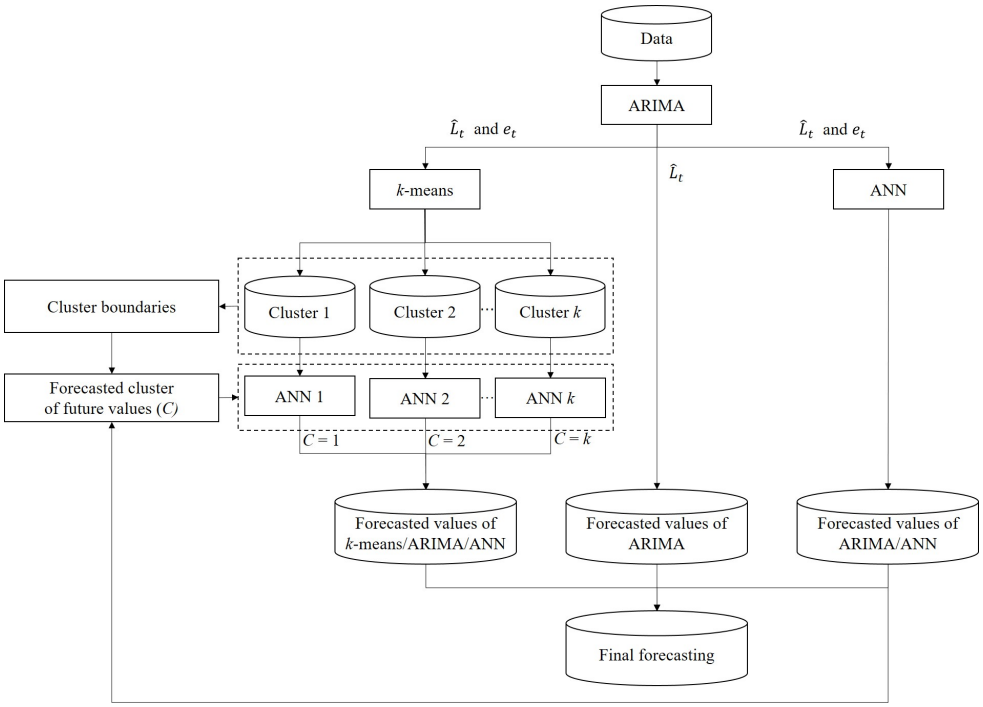

Figure 3.1: The proposed hybrid model

3.2.2 Stage II: the linear-nonlinear modeling stage

The Khashei and Bijari's model (ARIMA/ANN model) [16] is a single hidden layer ANN including the lagged values of the times series, and the predicted values and the residuals of the ARIMA model as the inputs. The benefit of this model is capability in fitting the relationship between the time series and the linear and nonlinear components as a function shown in (3.1) in stead of additive relationship in the Zhang's model [14].

$$\hat{Z}_t = f(\hat{L}_t, e_{t-1}, \ldots, e_{t-n1}, Z_{t-1}, \ldots, Z_{t-m1}) \tag{3.1}$$

where $\hat{Z}_t$ is the forecasted value at period t, f is the function fitted by the ANN, $\hat{L}_t$ is the predicted value of the ARIMA at period t, Z_t is the actual value at period t, e_t is the residual of the ARIMA at period t, $n_1 \leq n$ and $m_1 \leq m$ are integers that are identified in the design process.

The feedforward neural network with one hidden layer is applied. The transfer function between the input and the hidden layer is sigmoid function (3.2). On the other hand, the transfer function between the hidden and the output layer is linear transfer function.

$$\text{Sigmoid}(x) = \frac{2}{1 + e^{-x}} \tag{3.2}$$

The number of hidden nodes, $n1$, and $m1$ are varied for searching the best fitted model. The forecasted values of the best fitted model are moved to the final forecasting stage and also used for the cluster prediction in the k-means/ARIMA/ANN model as well.

3.2.3 Stage III: the linear-nonlinear modeling with k-mean clustering stage

In this stage, firstly, the k-means clustering algorithm is applied for clustering the time series into the clusters. The suitable number of the clusters (k) can be selected from the number of the clusters of which has the highest average of the Silhouette (s_i). To determine the suitable number the cluster (k) is varied from one to five Then, the ANN models (so called k-means/ARIMA/ANN models) are built dedicatedly for each cluster.

In order to forecast the future time series, the cluster of the future time series have to be identified at first. The prediction results from the stage II are assigned into the clusters by using the boundaries between the clusters. The boundaries between the clusters are defined by taking an average of a maximum data of the lower centroid cluster and a minimum data of the higher centroid cluster. The observations falling into the same boundaries will be considered as members of the same cluster. In each forecasting period, the result comes from the k-means/ARIMA/ANN model chosen according to the predicted cluster.

For instance, we suppose that there are two clusters: cluster 1 and cluster 2. The centroid of the cluster 1 is lower than the cluster 2. The maximum value of the cluster 1 is 18. The minimum value of the cluster 2 is 22. In this case, the average of these values (i.e. 18 and 22) is 20 which is the boundary separating the cluster 1 and 2. On the other hand, the lower and higher boundaries of the cluster 1 is 0 and 20 respectively. the lower and higher boundaries of the cluster 2 are 20 and infinity respectively. If the forecasted value of the stage II (ARIMA/ANN model) at period $t + 1$ is 19, then, the forecasted value belongs to the cluster 1 because 19 is between the boundaries of the cluster 1. Thus, the cluster 1 is chosen for the prediction at period $t + 1$. Eventually, the results of the

k-means/ARIMA/ANN model are included in the final forecasting stage.

3.2.4 Stage IV: the final forecasting stage

According to the experimental results of the previous three stages, the different forecasting models give the different prediction values due to unique capability in capturing patterns of the time series. The model giving the best average performance cannot guarantee to outperform the others in every forecasting periods. On the other hand, the model providing the worst average accuracy can have the best result in some periods of prediction. In order to improve forecasting accuracy based on this circumstance, a combination of the three forecasting models is proposed as additive weighting summation of the results from stage I, II, and III:

$$\hat{Z}_t = \sum_{i=1}^{m} w_{i,t} \hat{Z}_t^i \tag{3.3}$$

where $\hat{Z}_t$ is the final forecasted value at period t, $\hat{Z}_t^i$ is the forecasted value of model i at period t, $w_{i,t}$ is the weight of model i at period t, and m is the total number of the forecasting models.

The forecasting model giving more accurate prediction will be assigned to have more weight ($w_{i,t}$). The weight of the first forecasting period ($w_{i,1}$) is determined by applying discount mean square forecast error (DMSFE) method [63] to the training period as:

$$w_{i,1} = \frac{\left[\sum_{tr=1}^{Tr} \gamma^{Tr-tr+1} \left(Z_{tr} - \hat{Z}_{tr}^i \right)^2 \right]^{-1}}{\sum_{i=1}^{m} \left[\sum_{tr=1}^{Tr} \gamma^{Tr-tr+1} \left(Z_{tr} - \hat{Z}_{tr}^i \right)^2 \right]^{-1}} \tag{3.4}$$

where Z_{tr} is the actual value at training period tr, $\hat{Z}_{tr}^i$ is the fitted value of forecasting method i at training period tr, Tr is the total number of the training period, γ is the discount factor assumed 0.8.

To compute the weight of each forecasting model at the first forecasting period ($w_{i,1}$), first, the squared error or SE, which is denoted as $(Z_{tr} - \hat{Z}_{tr}^i)^2$ in (3.4), of each forecasting model in the training period is computed. Second, the value of the SE at each training period is deduced over the time by the discount factor (γ). Third, the discounted SEs in

every training period are summed up to get the sum of discounted SEs for each forecasting model (the top part of (3.4)). Forth, the sum of discounted SEs of all forecasting models are aggregated into the total sum of discounted SEs (the bottom part of (3.4)).

The weight ($w_{i,1}$) obtained from the DMSFE is used as the initial weight of the forecasting period ($t = 1$), and then the weight ($w_{i,t}$) is updated in every forecasting period ($t = 2, 3, \ldots, T$) by exponential smoothing method with α assumed 0.2 as below:

$$w_{i,t} = \alpha \left[\frac{\left(Z_t - \hat{Z}_t^i \right)^{-2}}{\sum_{i=1}^{m} \left(Z_t - \hat{Z}_t^i \right)^{-2}} \right] + (1 - \alpha) w_{i,t-1} \tag{3.5}$$

3.3 Application of the proposed model to real-world time series

In order to examine the prediction capability of the proposed model, the proposed model is applied to three well-known data sets: Wolf's sunspot, Canadian lynx, and exchange rate (British pound to US dollar). These three data sets are different in term of fields and statistical properties. Several researchers have applied these data sets to demonstrate effectiveness of linear, nonlinear and hybrid models [14, 16].

To investigate the suitable parameters of the proposed model, comprehensive experiments are performed in various scenarios that have different set of parameters: 1-10 lagged periods of actual time series (Z_t), results of the ARIMA ($\hat{L}_t$) and residual of the ARIMA (e_t); 1-10 hidden nodes of the ANN; and 2-5 numbers of cluster (k).

Each scenario is run for five replications to obtain the average of the results (i.e. one step-ahead forecasting). After that, the prediction performance is evaluated by three well-known measures: mean square error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The proposed model is compared to the ARIMA, the ANN, the Khashei and Bijari's model (ARIMA/ANN model), and the k-means/ARIMA/ANN model. The detail experiments and discussion of the results are provided in the rest of this section.

3.3.1 Wolf's sunspot forecasting

The Wolf's sunspot time series is a historical annual record of number of black spots on the sun surface. The sunspot time series involved in this experiment is recorded during 1700-1987 (288 observations) (Fig. 3.2). The records in 1700-1920 (221 observations) are included in model training. The remaining records in 1921-1987 (67 observations) are used for model testing.

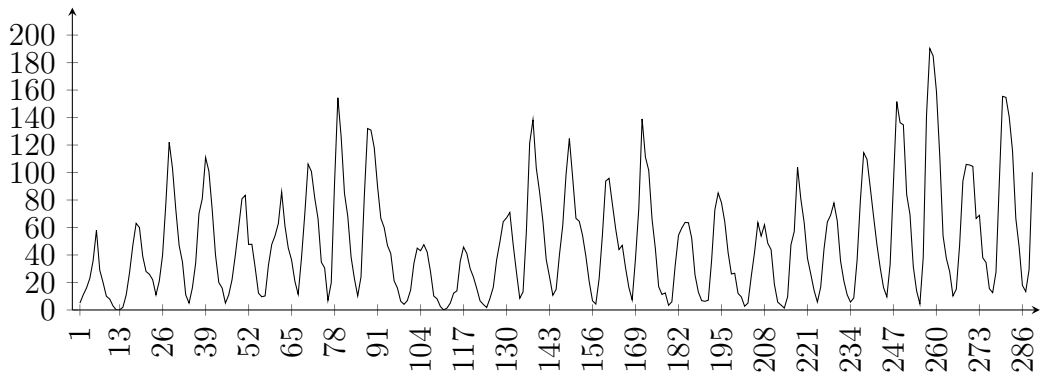

Figure 3.2: The Wolf's sunspot time series (1700-1987)

Stage I. linear modeling: The sunspot time series is analyzed by the ARIMA to obtain the linear component ($\hat{L}_t$). ARIMA(9,0,0) is the best fitted model that has also been chosen by several researchers [16, 14, 64]. The results of the ARIMA(9,0,0) are passed through stage II-IV.

Stage II. linear-nonlinear modeling: The ARIMA/ANN model is applied to deal with linear and nonlinear components of the sunspot time series. The linear components, the residuals of the ARIMA(9,0,0), and the lagged values of the sunspot time series are included in the model. The ARIMA/ANN(7-4-1) model (Fig. 3.3) is the best fitted model.

Stage III. linear-nonlinear modeling with k-mean clustering: Firstly, the training set of the sunspots time series is classified into the clusters by the k-mean clustering algorithm. Two clusters is suitable for the sunspot time series according to the highest average of Silhouette (s_i). The clusters are completely separated without overlapping (Fig. 3.4). Then, the two k-means/ARIMA/ANN models are dedicatedly constructed for each cluster. The k-means/ARIMA/ANN models fitted to cluster 1 and 2 are the k-means/ARIMA/ANN(4-3-1) model (Fig. 3.5) and the k-means/ARIMA/ANN(5-2-1) model (Fig. 3.6) respectively.

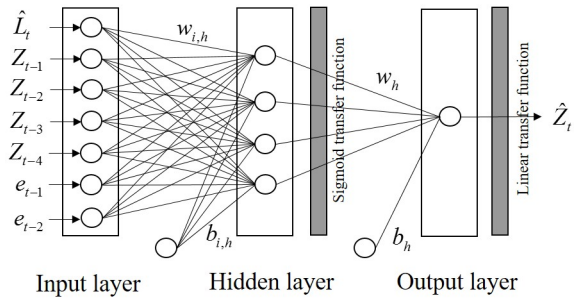

Figure 3.3: ARIMA/ANN(7-4-1) model for the sunspot time series

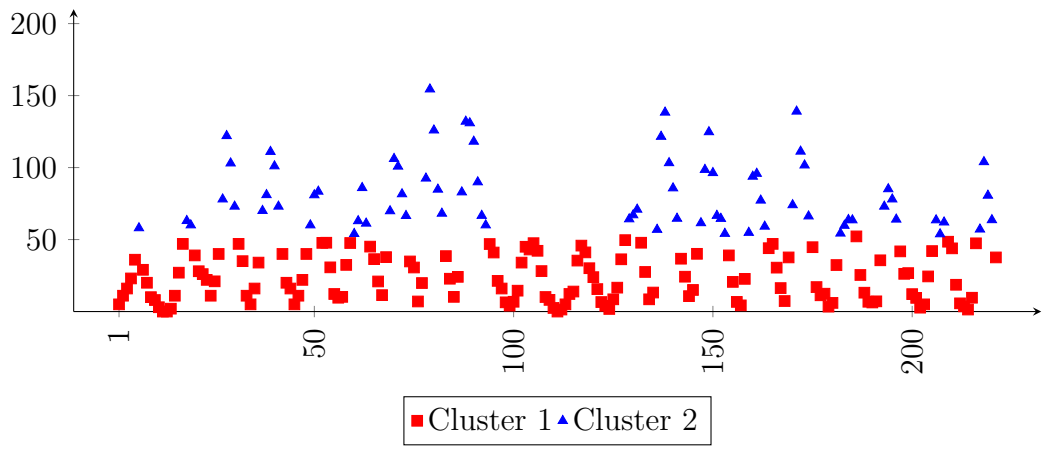

Figure 3.4: Clusters of the sunspot training set

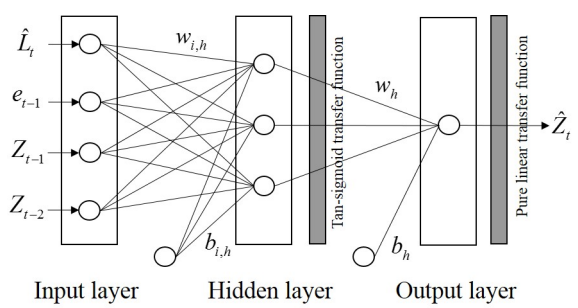

Figure 3.5: k-means/ARIMA/ANN(4-3-1) model for the sunspot cluster 1

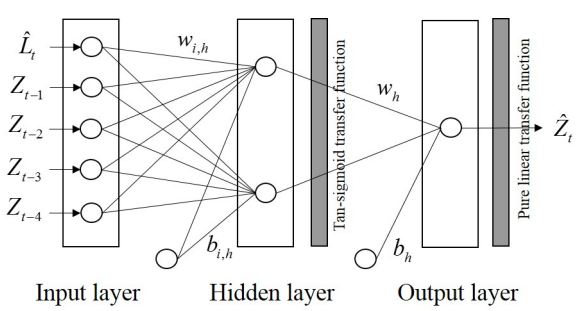

Figure 3.6: k-means/ARIMA/ANN(5-2-1) model for the sunspot cluster 2

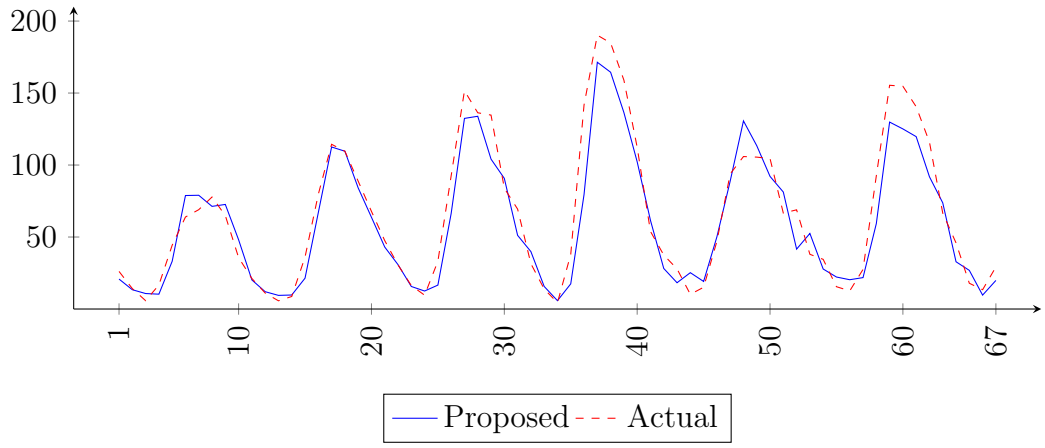

Figure 3.7: Forecasted values of the proposed model for the sunspot time series

Stage IV. final forecasting stage: The forecasted values from the ARIMA, the ARIMA /ANN and the k-means/ARIMA/ANN models are aggregated by the weights from the discount mean square forecast error (DMSFE) method. The proposed model gives 224.84 of MSE, 11.60 of MAE, and 24.47% of MAPE. The forecasted values of the proposed model are shown in Fig. 3.7.

3.3.2 Canadian lynx forecasting

The Canadian lynx time series used for this experiment is the annual number of lynx trapped in the Mackenzie River district of Northern Canada during 1821-1934 (114 observations) (Fig. 4.3). The lynx time series has been studied in several researches [65, 66, 67]. However, the data transformation with based ten logarithm is performed [14, 16]. The training and test sets are 1821-1920 (100 observations) and 1921-1934 (14 observations) respectively.

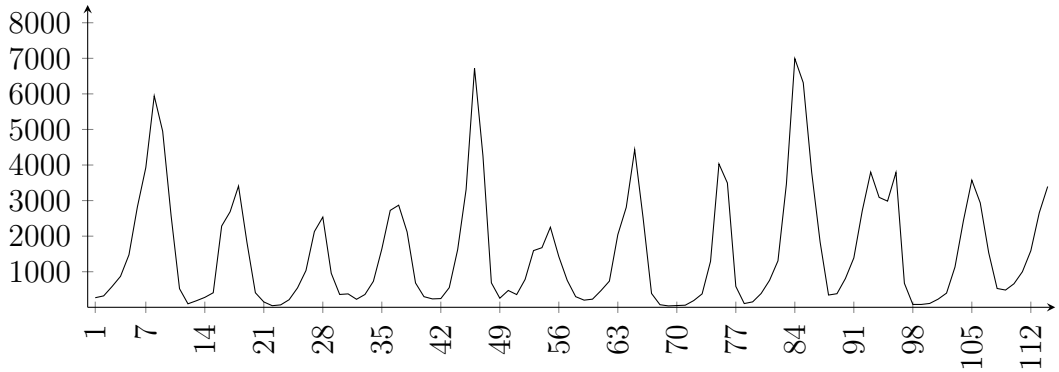

Figure 3.8: Canadian lynx time series (1821-1934)

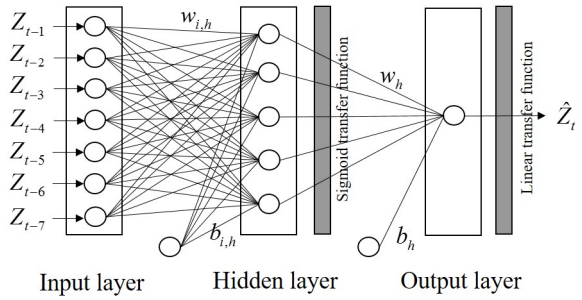

Figure 3.9: ARIMA/ANN(7-5-1) model for the Canadian lyx time series

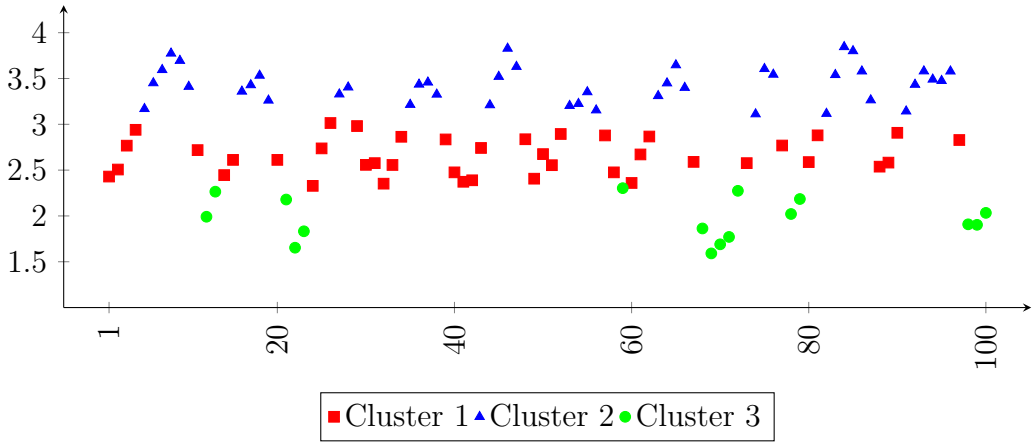

Figure 3.10: Clusters of the Canadian lynx training set

Stage I. linear modeling: ARIMA(12,0,0) is the best fitted of the ARIMA for the lynx time series. It has been also chosen by several researchers [14, 16].

Stage II. linear-nonlinear modeling: The best fitted model is ARIMA/ANN(7-5-1) model that contains seven inputs which are lagged values from one to seven periods and five hidden nodes as shown in Fig. 3.9.

Stage III. linear-nonlinear modeling with clustering: Although the training set of lynx time series is separated into three clusters (Fig. 3.10), but after prediction of the clusters of the test set, there are only cluster 1 and 2. Therefore, the k-means/ARIMA/ANN for cluster 3 is not involved in this stage. The appropriate models for the cluster 1 and 2 are k-means/ARIMA/ANN(2-4-1) (Fig. 3.11) and k-means/ARIMA/ANN(2-1-1) (Fig. 3.12) respectively.

Stage IV. final forecasting: In the same way as the previous data set, the result from stage I, I, and III are combined together by the weights from the DMSFE method. The performance of the proposed model in the three measures are 0.0135 of MSE, 0.0885 of MAE, and 3% of MAPE. The forecasting values of the lynx are presented in Fig. 3.13.

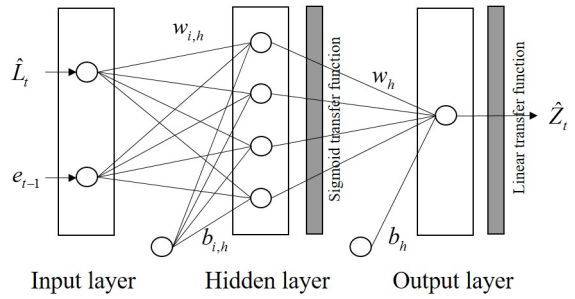

Figure 3.11: k-means/ARIMA/ANN(2-4-1) model for the Canadian lynx cluster 1

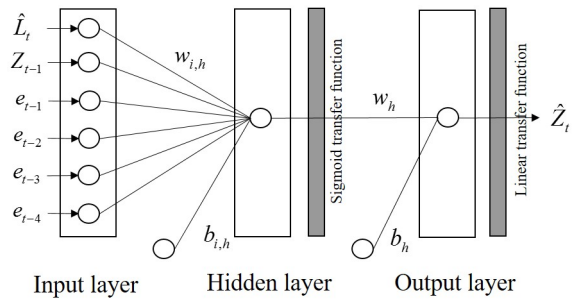

Figure 3.12: k-means/ARIMA/ANN(6-1-1) model for the Canadian lynx cluster 2

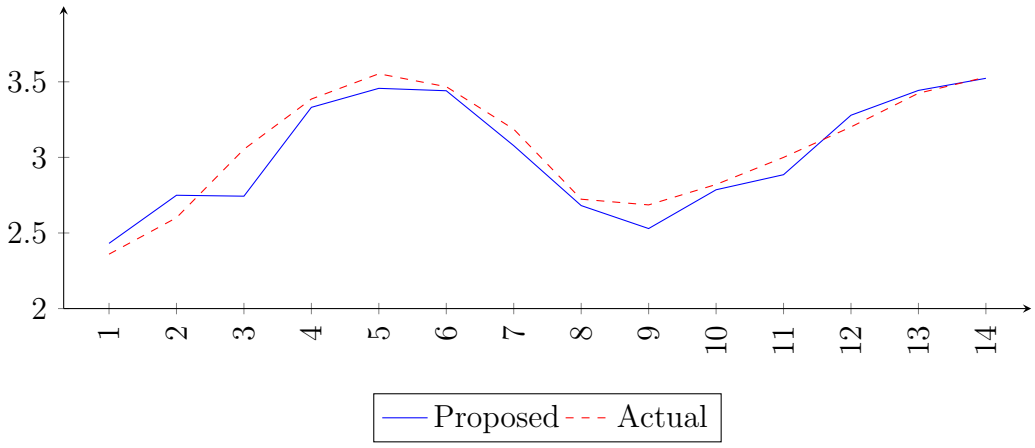

Figure 3.13: Forecasted values of the proposed model for the Canadian lynx time series

3.3.3 Exchange rate forecasting

The weekly exchange rate of British pound to US dollar in 1980-1993 (731 observations) (Fig. 4.4) is involved in the experiment. The natural logarithm is used for data transformation as in [14, 16, 68]. The first 679 records belongs to training set, and the remaining 52 records is for model testing.

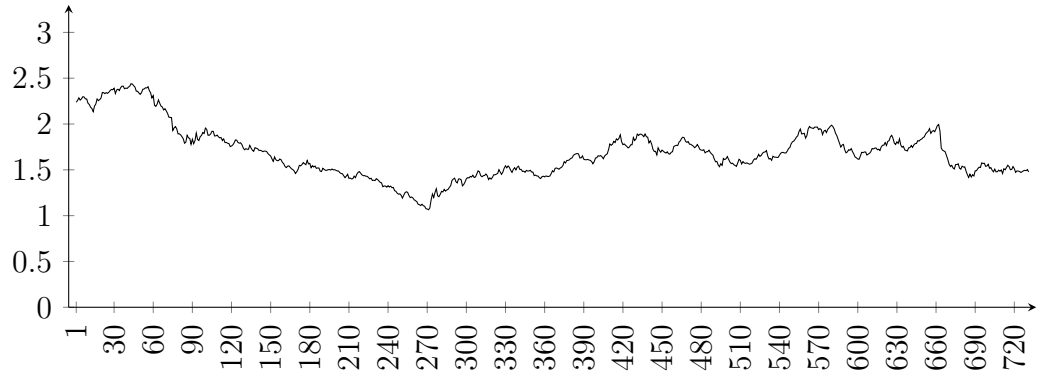

Figure 3.14: Exchange rate time series (1980-1993)

Stage I. linear modeling: The best fitted model is ARIMA(0,1,0) which is a random walk model. This model has been selected by Zhang [14] and Khashei and Bijari [16] as well. Additionally, several researchers in exchange rate forecasting research area have also recommended to use the random walk model for the linear modeling.

Stage II. linear-nonlinear modeling: The best fitted model is ARIMA/ANN(20-9-1) consisting of nine hidden nodes and the twenty inputs: the linear component from the stage I, seven periods lagged values, lagged values of ARIMA(0,1,0) residuals from one to nine periods. The structure of the model is shown in Fig. 3.15.

Stage III. linear-nonlinear modeling with clustering: The suitable number of the cluster of the training set of the exchange rate is three clusters. Nevertheless, the predicted clusters of all periods in test set belong to cluster 2. Therefore, only k-means/ARIMA/ANN for the cluster 2 is considered. The best fitted model of the cluster 2 is k-means/ARIMA/ANN(8-9-1) composed of the linear component from the stage I, one periods lagged value, and six periods lagged values of residuals of ARIMA(0,1,0) as presented in (Fig. 3.16).

Stage IV. final forecasting: The results from the stage I, II, and III are aggregated by the weights from the DMSFE method. The proposed model has prediction performance in

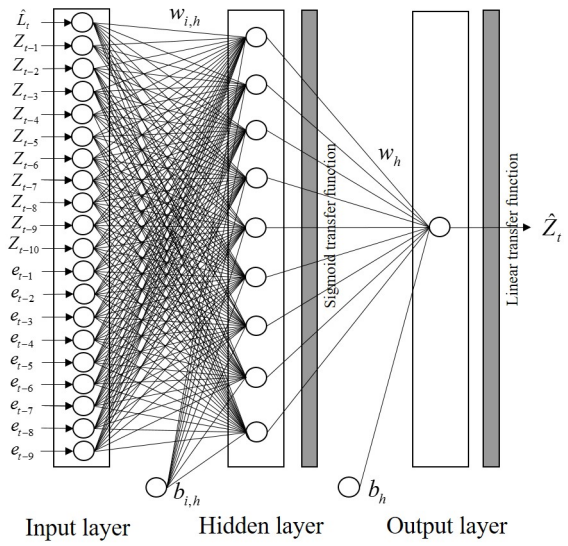

Figure 3.15: k-means/ARIMA/ANN(20-9-1) model for the exchange rate time series

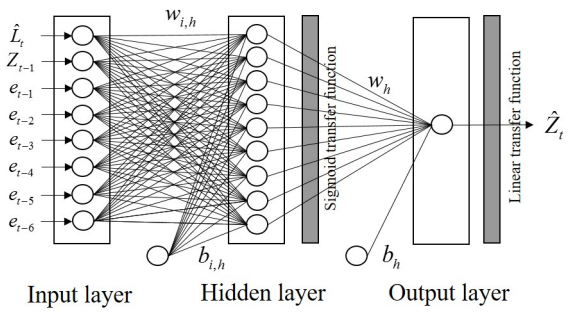

Figure 3.16: k-means/ARIMA/ANN(8-9-1) model for the exchange rate cluster 2

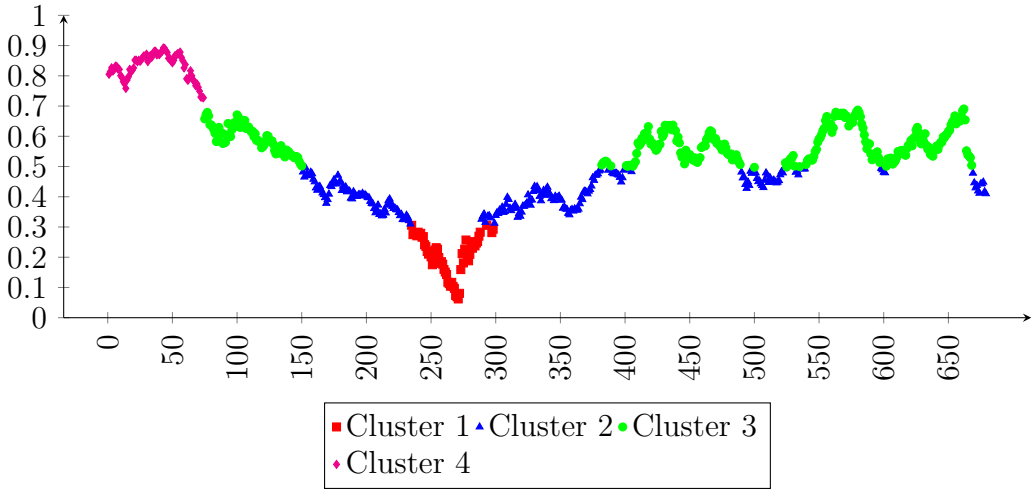

Figure 3.17: Clusters of the exchange rate training set

Figure 3.18: Forecasted values of the proposed model for the exchange rate time series

the three measures as 27.67 of MSE, 0.01347 of MAE, and 3.35% of MAPE. The predicted values of the exchange rate are shown in Fig. 3.18.

3.3.4 Comparison with the other forecasting models

In this section, the prediction performance of the proposed model is compared with the ARIMA, the ANN, the Khashei and Bijari's model (ARIMA/ANN model), and the k-means/ARIMA/ANN model. In the evaluation of the prediction performance, the forecasting models are applied to the three well-known time series which are Wolf's sunspot, Canadian lynx, and British pound per US dollar exchange rate. The prediction accuracy is measured by mean square error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

According to the experimental results, although the best fitted model of the ARIMA, the Khashei and Bijari's model, and the proposed model have been already investigated but there is no experiment for the ANN. Hence, additional experiments are performed to find the best structure of the ANN for each time series. Based on the results of the additional experiments, the suitable ANN structures for the sunspot, the lynx and the exchange rate time series are ANN(7-4-1) (Fig. 3.19), ANN(7-5-1) (Fig. 3.20), and ANN(7-6-1) (Fig. 3.21) respectively.

In case of the sunspot time series, the performance of the best fitted models is presented in Table 3.1. The proposed model can outperform the other models. Similarly, for the lynx time series, the proposed model also gives the best prediction results as shown in

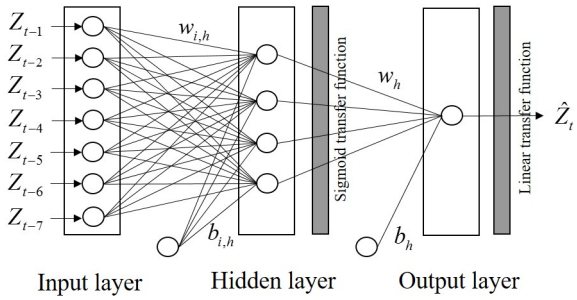

Figure 3.19: ANN(7-4-1) model for the sunspot time series

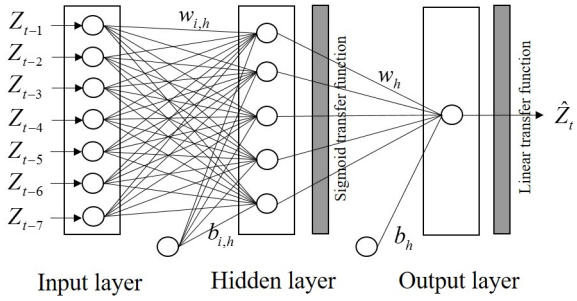

Figure 3.20: ANN(7-5-1) model for the Canadian lynx time series

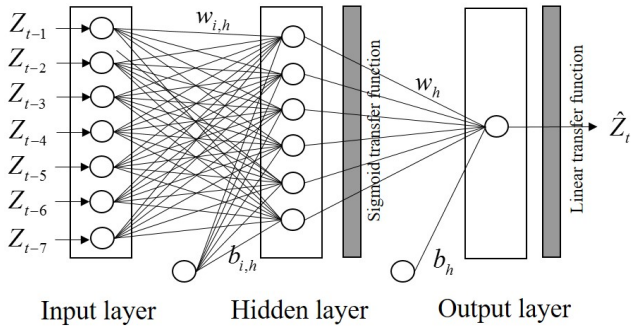

Figure 3.21: ANN(7-6-1) model for the exchange rate time series

Table 3.1: Sunspot forecasting performance comparison

Model	MSE	MAE	MAPE
ARIMA	276.35	12.55	30.11%
ANN	350.09	13.54	26.64%
Khashei and Bijari	273.15	12.14	25.31%
k-means/ARIMA/ANN	371.74	13.45	26.43%
Proposed	**244.84**	**11.60**	**24.47%**

Table 3.2: Lynx forecasting performance comparison

Model	MSE	MAE	MAPE
ARIMA	0.0229	0.1120	3.71%
ANN	0.0188	0.1042	3.57%
Khashei and Bijari	0.0160	0.0980	3.24%
k-means/ARIMA/ANN	0.0190	0.0984	3.38%
Proposed	**0.0135**	**0.0885**	**3.00%**

Table 3.2. The exchange rate forecasting performance is shown in Table 3.3. Likewise the previous data sets, the proposed model dominates the other models. In summary, the proposed model can give the better prediction performance for all three data sets and in all three measures. Moreover, the percentage of performance improvement is presented in Table 3.4. The lynx time series seems to have the most improvement following by the sunspot and the exchange rate time series.

Furthermore, in order to demonstrate the prediction performance in more detail, absolute percentage error (APE) in testing period of the sunspot, the lynx, and the exchange rate time series is displayed on Fig. 3.22, 3.23, and 3.24 respectively. These figures show that in some prediction periods, even the model that has the best performance in the average (e.g. MAPE) can have the worst performance.

For instance, according to the performance of the lynx time series in Table 3.2, excluding the proposed model, the best model is the Khashei and Bijari's model. However, at periods 7 and 12 in Fig. 3.23, the Khashei and Bijari's model performs worse than the other models, as well as at some prediction periods of the sunspot and the exchange rate time series in Fig. 3.22 and 3.24. These empirical results support the benefit of the proposed model in combining different models in term of types and complexity.

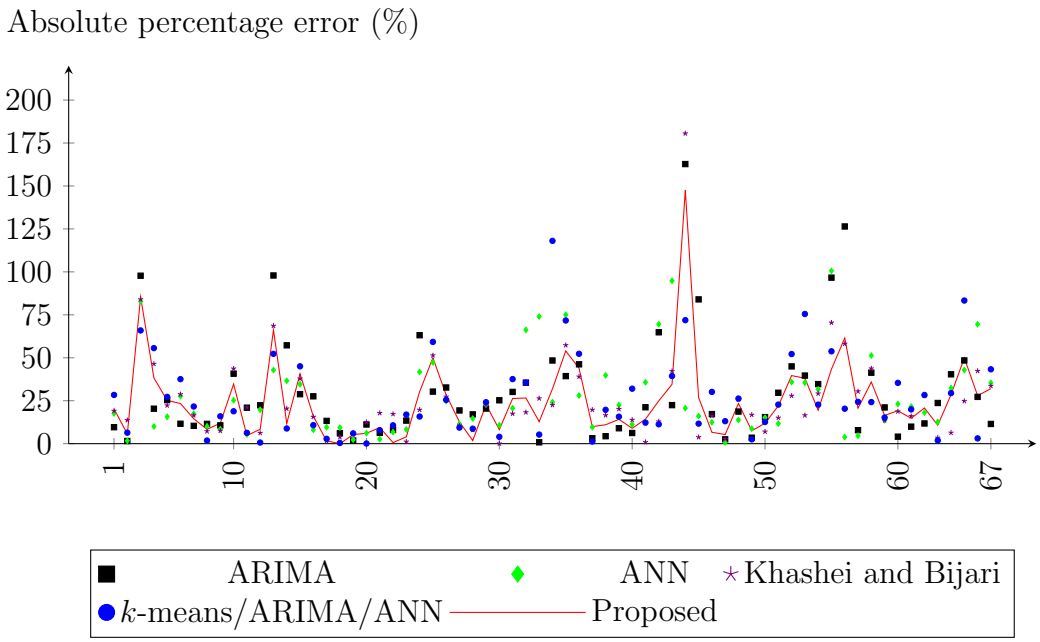

Figure 3.22: APE of the proposed model for the sunspot time series

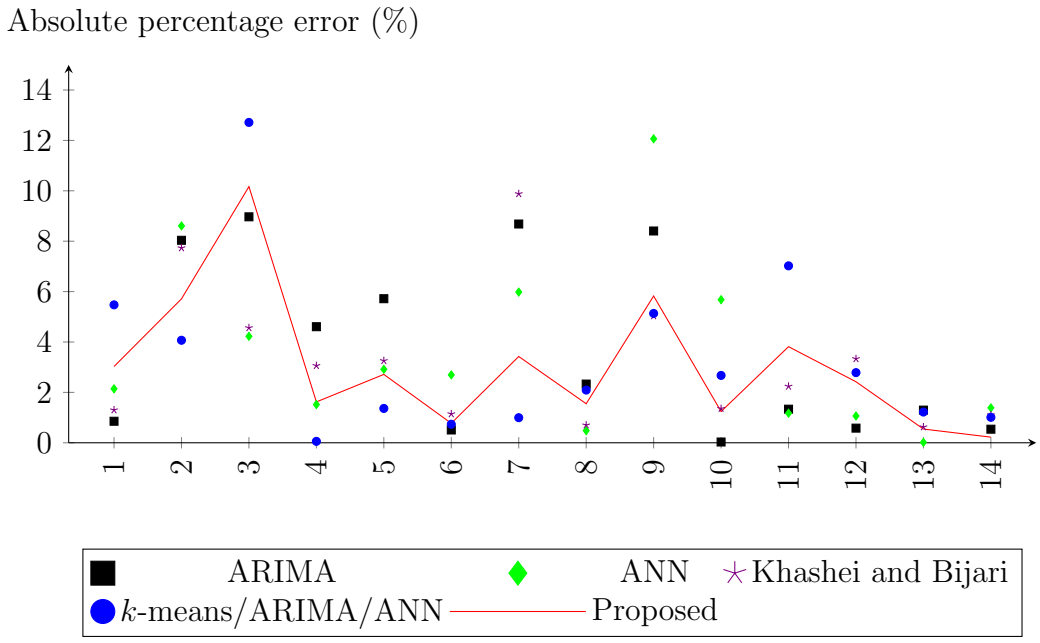

Figure 3.23: APE of the proposed model for the Canadian lynx time series

Table 3.3: Exchange rate forecasting performance comparison

Model	MSE	MAE	MAPE
ARIMA	28.7808	0.01375	3.429%
ANN	29.5617	0.01372	3.425%
Khashei and Bijari	28.3242	0.01351	3.360%
k-means/ARIMA/ANN	28.0819	0.01387	3.447%
Proposed model	**27.6667**	**0.01347**	**3.350%**

All MSE is multipled by 10^{-5}.

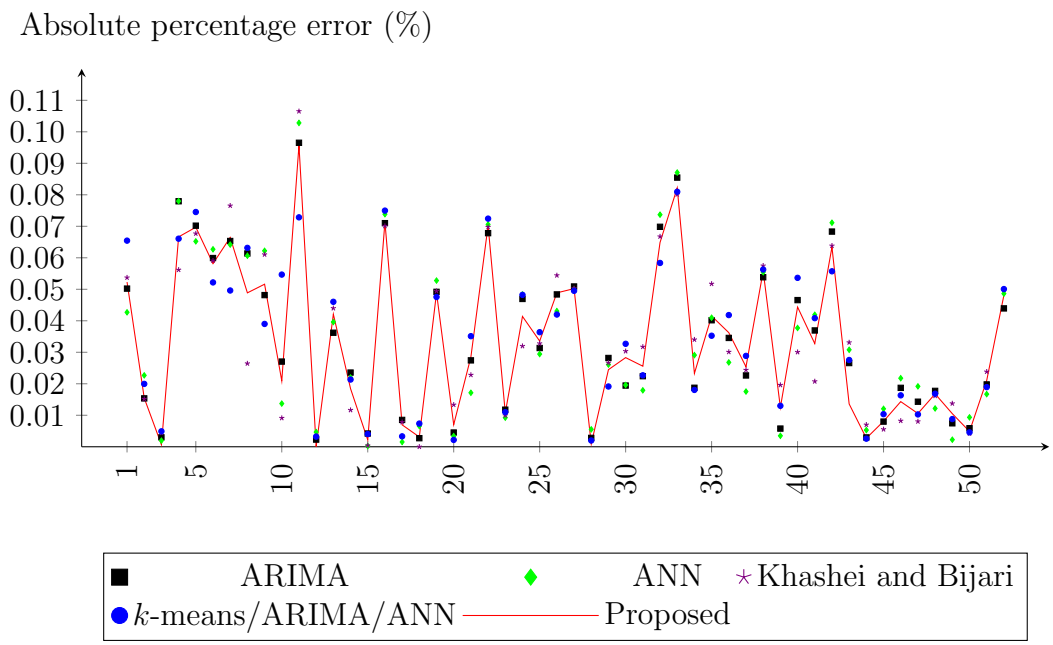

Figure 3.24: APE of the proposed model for the exchange rate time series

Table 3.4: Percentage improvement of the proposed model

Model	Sunspot			Lynx			Exchange rate		
	MSE	MAE	MAPE	MSE	MAE	MAPE	MSE	MAE	MAPE
ARIMA	11.40	7.63	18.74	39.07	19.05	17.02	3.87	2.10	2.28
ANN	27.23	13.76	20.61	25.78	12.94	3.843	6.41	1.83	2.19
Khashei and Bijari	10.36	4.47	3.35	12.60	7.45	5.03	2.32	0.34	0.29
k-means/ARIMA/ANN	34.14	13.78	7.42	26.50	7.87	9.05	1.48	2.89	2.81

3.4 Conclusion

Improving time series forecasting is an important yet often difficult task. Although, the numerous time series forecasting models have been developed, but the research for developing new forecasting models to improve the prediction accuracy has never stopped. Recently, hybrid models, combinations of single models which can be same or difference model types, became popular time series forecasting approach especially hybridizations of linear and nonlinear forecasting models. Usually, these hybrid models promise a better prediction accuracy in average than forecasting by a single forecasting model. However in some prediction periods, the single model still give the better accuracy. Therefore, there is a potential to improve the forecasting accuracy with combination of single and hybrid models.

The hybrid model of the ARIMA, the Khashei and Bijari's model (ARIMA/ANN model) and the k-means/ARIMA/ANN model are proposed to obtain unique advantages among the different model types and complexity levels in time series forecasting. The prediction capability of the proposed model is tested with well-known data sets: the Wolf's sunspot, the Canadian lynx, and the British pound per US dollar exchange rate. From the empirical results, the proposed model gives the best performance in MSE, MAE, and MAPE for all three data sets.

In conclusion, the proposed hybridization between linear, linear-nonlinear, and linear-nonlinear with clustering technique has shown its forecasting capability over traditional single and hybrid models. Therefore, it can be used as an alternative model for time series prediction. However, the proposed model includes only one model per each model type (i.e. the ARIMA as the linear model, the ANN as the nonlinear model, and the k-means as the clustering technique). Actually, there are many other linear models, nonlinear models, and clustering techniques that can be involved the future study. Furthermore, the drawback of the proposed model is that in the stage III, the cluster prediction relies on the forecasted values from the ARIMA/ANN model; this can cause the wrong predicted cluster especially when the forecasted values are very close to the boundary between the clusters.

Chapter 4

Hybrid model of ARIMA and ANN with discrete wavelet transform (DWT) for time series forecasting

4.1 Introduction

Time series forecasting is an active research area that plays important role in planning and decision making in several practical applications [12]. The main task of this research area is to improve the prediction accuracy. For decades, the autoregressive integrated moving average (ARIMA) and the artificial neural network (ANN) are widely used for time series prediction. The ARIMA is popular due to capability in dealing with various types of data such as stationary and non-stationary data. However, the linear relationship between historical and predicted time series is pre-assumed. Such assumption is very difficult to be completely satisfied in practical situations.

On the other hand, the ANN can predict the future time series without any prior assumption. Nevertheless, there is no forecasting model that can be the best for all time series. For instance, the ARIMA works properly for linear time series but for nonlinear time series, the ANN can model nonlinear time series while the ARIMA is not appropriate. Therefore, using the single model is insufficient for dealing with the time series of real-world applications which normally contain both linear and nonlinear relationship [14].

The wavelet transform is traditionally applied for decomposing signal into low fre-

quency (approximation) and high frequency (detail) components. Recently, the discrete wavelet transform (DWT) has been adapted to decompose time series. Prediction performances of both ARIMA and ANN have been improved through the wavelet transform in several applications: electrical price [69, 70]; short term electrical load [71]; monthly river discharge [72, 73, 74]; groundwater level [75]; rainfall and runoff [76, 77]; hourly flood forecasting [78]. Furthermore, even though, the wavelet transform has been combined with the ARIMA and ANN models [79], this approach assumed that the approximation contains only nonlinear component. Such assumption is not practical since the DWT is not a method to transform time series into linear or nonlinear components.

This chapter proposes a novel forecasting model capturing both linear and nonlinear components of the detail and the approximation in stead of original time series. Firstly, the DWT is used to decompose the time series. Then, the hybrid model of ARIMA and ANN are constructed for the approximation and the detail to extract their linear and nonlinear components. Eventually, the final prediction is the combination of the linear and nonlinear components.

4.2 Proposed forecasting model

The main idea of the proposed model is using the unique strength of the ARIMA and the ANN in capturing linear and nonlinear components from time series while not assuming the approximation and the detail from the DWT as either linear or nonlinear. The proposed model consists of thee main steps: time series decomposition, extracting linear and nonlinear components, and final forecasting (Fig. 4.1).

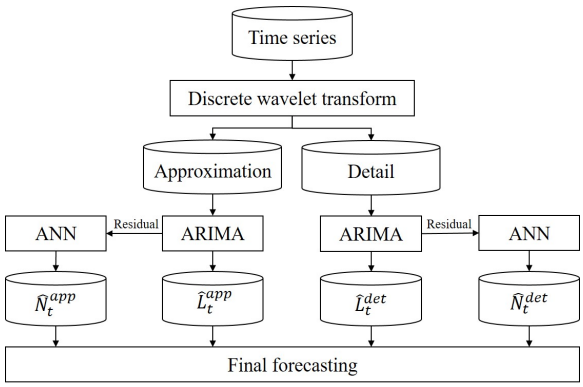

Figure 4.1: The proposed forecasting model

In the first step, the DWT decomposes actual time series (Z_t) into the approximation (Z_t^{app}) and the detail (Z_t^{det}) by using Daubechies wavelet basis function [80]. In this study, only the first level of the decomposition is considered because we want to test whether the most simplest DWT can enhance the prediction capability. Currently, to our best knowledge, there is no theoretical method for finding the optimal number for the decomposition level of the DWT, and the researchers still do trial and error to investigate the suitable decomposition level with the smallest prediction error.

In the second step, the Zhang's hybrid model of ARIMA and ANN [14] is applied to the approximation and the detail. Normally, the Zhang's model predicts the future value at period t ($\hat{Z}_t$) from a combination of linear ($\hat{L}_t$) and nonlinear ($\hat{N}_t$) components as:

$$\hat{Z}_t = \hat{L}_t + \hat{N}_t \tag{4.1}$$

The linear component ($\hat{L}_t$) can be obtained from the result of applying the ARIMA to the actual time series (Z_t). Then the ARIMA residuals (e_t) are computed as:

$$e_t = Z_t - \hat{L}_t \tag{4.2}$$

For the nonlinear component ($\hat{N}_t$), it is the result of the ANN using the ARIMA residuals (e_t) as the inputs as:

$$\hat{N}_t = f(e_{t-1}, e_{t-2}, \ldots, e_{t-n}) \tag{4.3}$$

where f is a function defined by the ANN, n is total lagged periods.

In case of the proposed model, different Zhang's models are dedicatedly constructed for the approximation and the detail as:

$$\hat{Z}_t^{app} = \hat{L}_t^{app} + \hat{N}_t^{app} \tag{4.4}$$

$$\hat{Z}_t^{det} = \hat{L}_t^{det} + \hat{N}_t^{det} \tag{4.5}$$

where $\hat{Z}_t^{app}$ and $\hat{Z}_t^{det}$ are the predicted approximation and detail respectively, at period t; $\hat{L}_t^{app}$ and $\hat{L}_t^{det}$ are linear components of the approximation and the detail respectively, at

period t; $\hat{N}_t^{app}$ and $\hat{N}_t^{det}$ are nonlinear components of the approximation and the detail respectively, at period t.

The linear components ($\hat{L}_t^{app}$ and $\hat{L}_t^{det}$) are the results of applying the ARIMA to Z_t^{app} and Z_t^{det} respectively. Then, the ARIMA residuals of the approximation (e_t^{app}) and the detail (e_t^{det}) can be computed as:

$$e_t^{app} = Z_t^{app} - \hat{L}_t^{app} \tag{4.6}$$

$$e_t^{det} = Z_t^{det} - \hat{L}_t^{det} \tag{4.7}$$

For the nonlinear components ($\hat{N}_t^{app}$ and $\hat{N}_t^{det}$), they are obtained from the ANN as:

$$\hat{N}_t^{app} = f^{app}(e_{t-1}^{app}, e_{t-2}^{app}, \ldots, e_{t-n_1}^{app}) \tag{4.8}$$

$$\hat{N}_t^{det} = f^{det}(e_{t-1}^{det}, e_{t-2}^{det}, \ldots, e_{t-n_2}^{det}) \tag{4.9}$$

where f^{app} and f^{det} are functions fitted by the ANN, n_1 and n_2 are total lagged periods which are determined by trial and error in the experiments.

At this step, we obtain two linear ($\hat{L}_t^{app}$ and $\hat{L}_t^{det}$) and two nonlinear components ($\hat{N}_t^{app}$ and $\hat{N}_t^{det}$). Finally, the final forecasting can be done by aggregation of the linear and nonlinear components as:

$$\begin{aligned} \hat{Z}_t &= \hat{y}_t^{app} + \hat{y}_t^{det} \\ &= \hat{L}_t^{app} + \hat{N}_t^{app} + \hat{L}_t^{det} + \hat{N}_t^{det} \end{aligned} \tag{4.10}$$

In summary, instead of applying the Zhang's model directly to the time series, the DWT is used for filtering the time series into the approximation and the detail, then, the Zhang's model is applied to each of them separately. Therefore, with the same characteristics of the data, the ARIMA and the ANN would have more potential to capture the pattern in the data effectively.

4.3 Experiments and results

In order to evaluate the prediction capability of the proposed model, the proposed model is applied to three well-known time series (Table 4.1): Wolf's sunspot (Fig. 4.2), Canadian lynx (Fig. 4.3) and British pound/US dollar exchange rate (Fig. 4.4). The prediction performance measures involved in this study consist of three measures: mean square error (MSE), mean absolute error (MAE) and mean absolute error (MAPE). The performance of the proposed model is compared to the ARIMA, the ANN and the Zhang's hybrid model.

For the sunspot time series, there are totally 288 annual records (1700-1987). The first 221 records (1700-1920) are used as training set. The remaining 67 records (1921-1987) are test set. Firstly, the sunspot time series is passed through the DWT to generate the approximation and the detail. Secondly, the ARIMA is fitted to both the approximation and the detail. ARIMA(0,0,6) and ARIMA(0,0,3) are the most fitted models. Thirdly, the residuals of the ARIMAs are computed and analyzed by the ANN. The best ANNs for the residuals of the approximation and the detail are ANN(10-10-1) and ANN(2-2-1). After the final forecasting, the performance measures of short term (35 years) and long term (67 years) horizontal predictions are evaluated (Table 4.2). From the comparison, the proposed model gives the lowest error in all three measures. In the short term prediction, the MSE, MAE and MAPE are 121.12, 6.16 and 19.45% respectively. For the long term prediction, the MSE, MAE and MAPE are 199.24, 7.77 and 22.56% which are higher than the short term prediction because the long term prediction includes the highest peak at period 37 (see Fig. 4.5a) that causes shift up in variability of the time series. However, the proposed model is still the best model because the measures of the other models are increased as well. Therefore, the proposed model is the best model in short and long term prediction for the sunspot time series.

The Canadian lynx time series contains 114 observations (1821-1934). The size of the training and the test set are 100 observations (1821-1920) and 14 observations (1921-1934) respectively. After obtaining the approximation and the detail from the DWT, their best fitted ARIMAs are ARIMA(0,0,5) and ARIMA(2,0,0) respectively. The most suitable ANNs for predicting the residuals of these two ARIMAs are ANN(3-8-1) and ANN(4-4-1) respectively. The performance comparison is shown in Table 4.3. The proposed model has

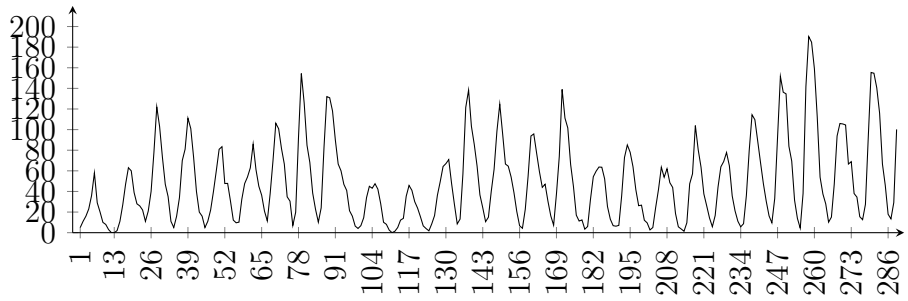

Figure 4.2: Sunspot time series (1700-1987)

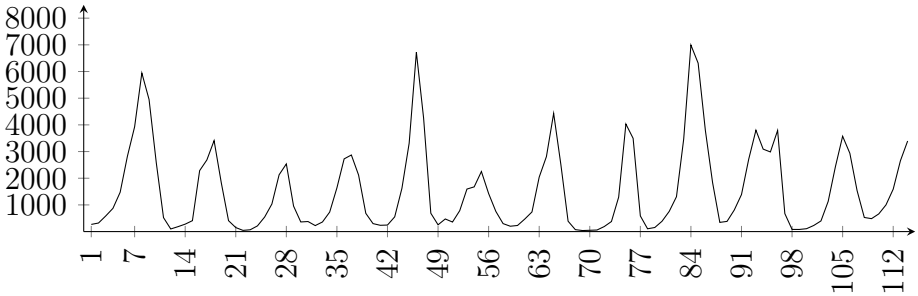

Figure 4.3: Canadian lynx time series (1821-1934)

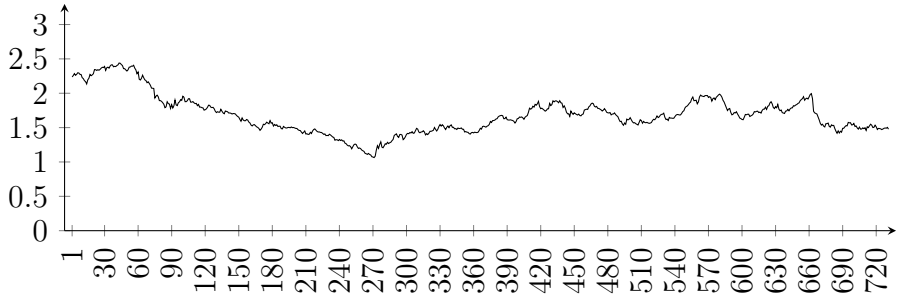

Figure 4.4: Exchange rate time series (1821-1934)

the best performance in MSE, MAE and MAPE which are 0.0154, 0.1014 and 3.4575% respectively. With the proposed model, the MSE is significantly improved. Based on the mathematical formula of the MSE, it is more sensitive to a big error; the lower MSE has more chance to promise lower maximum of error. In this case, the proposed model has the lowest maximum error which is at period 1 (see Fig. 4.5b)

In case of the exchange rate, the data set includes 731 weekly exchange rates (1980-1993). The data is partitioned into 679 and 52 observations as training and test set. The ARIMAs fitted to the approximation and the detail are ARIMA(0,1,0) and ARIMA(0,0,3) respectively. The ANNs of the ARIMAs residuals are ANN(2-3-1) and ANN(7-9-1). In evaluation of the prediction performance, three hirozontal predictions are considered: 1

Table 4.1: Detail of time series and experiment

Time series	Size (total, training, test)
Sunspot (1700-1987)	(288, 221, 67)
Canadian lynx (1821-1934)	(114, 100, 14)
Exchange rate (1980-1993)	(731, 679, 52)

Table 4.2: Sunspot forecasting result

Model	35 year ahead			67 year ahead		
	MSE	MAE	MAPE	MSE	MAE	MAPE
ARIMA	197.87	10.52	29.17%	323.48	13.25	32.86%
ANN	164.08	9.51	31.76%	413.90	14.19	33.34%
Zhang's model	156.76	9.63	30.22%	300.88	12.74	32.08%
Proposed model	**121.12**	**6.16**	**19.45%**	**199.24**	**7.77**	**22.56%**

Table 4.3: Lynx forecasting result

Model	MSE	MAE	MAPE
ARIMA	0.0229	0.1120	3.7062%
ANN	0.0201	0.1165	4.0156%
Zhang's model	0.0247	0.1083	3.5504%
Proposed model	**0.0154**	**0.1014**	**3.4575%**

Table 4.4: Exchange rate forecasting result

Model	1 month			6 months			12 months		
	MSE	MAE	MAPE	MSE	MAE	MAPE	MSE	MAE	MAPE
ARIMA	36.6514	0.01508	3.6628%	35.5387	0.0155	3.8883%	29.0517	0.01375	3.4286%
ANN	33.7797	0.0147	3.5767%	37.6072	0.0157	3.9298%	30.4533	0.01404	3.5065%
Zhang's model	36.1826	0.0151	3.6605%	35.6239	0.0156	3.8994%	28.8004	0.01379	3.4335%
Proposed model	**31.0259**	**0.0102**	**2.5541%**	**10.9023**	**0.0073**	**1.8240%**	**8.1812**	**0.00601**	**1.4959%**

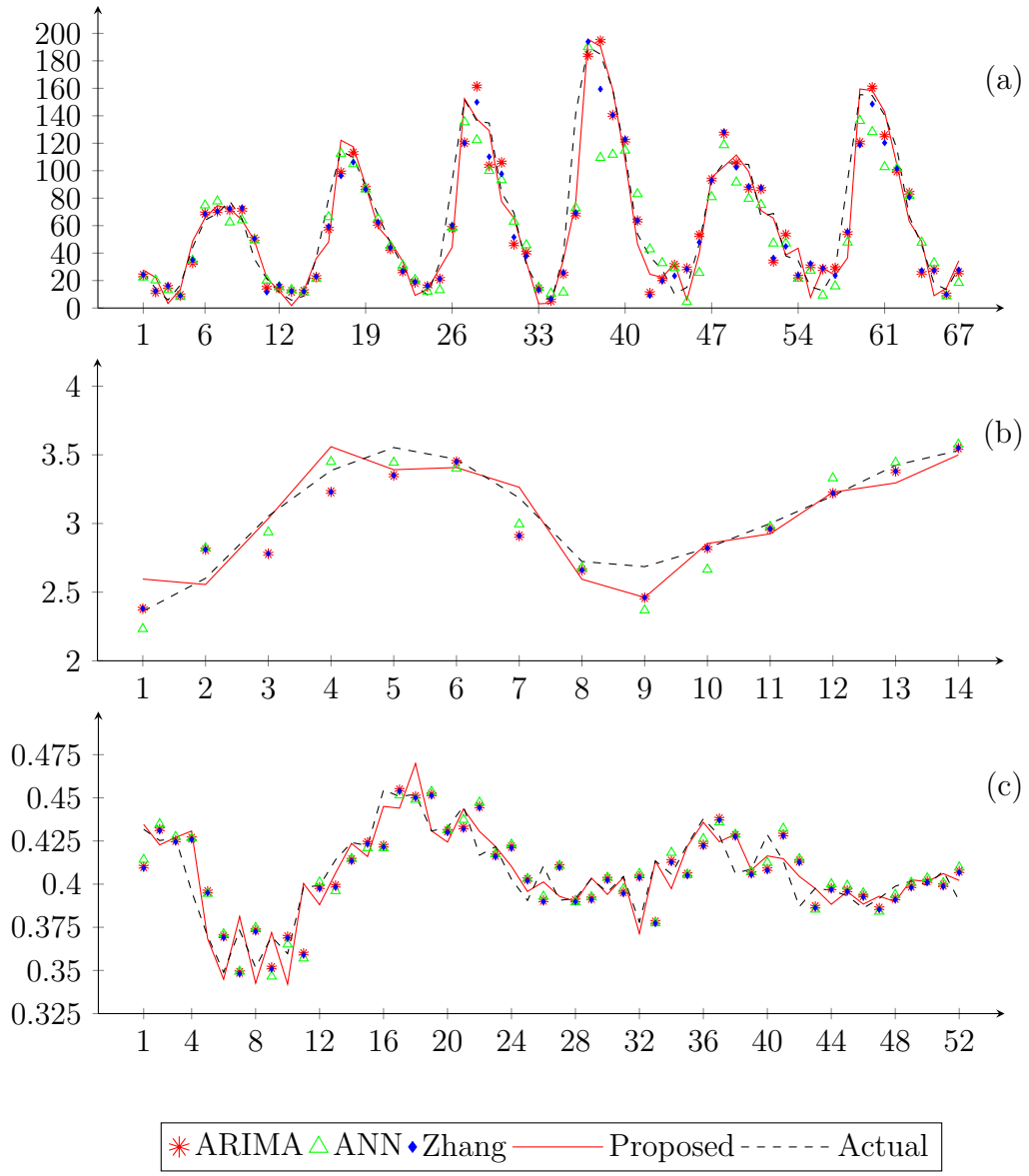

Figure 4.5: Forecasted values: (a) Sunspot, (b) Canadian lynx, (c) Exchange rate

month, 6 months and 12 months. From Table 4.4, the proposed model can outperform the other models in all cases. Surprisingly, the longest prediction period has the highest accuracy; MSE, MAE and MAPE are 8.1812, 0.00601 and 1.4959% respectively. Even though, the prediction period (12 months) is the longest but the average variability is the lowest because it includes the end of the year (see Fig. 4.5c) which has more stable exchange rates than the beginning and middle of the year.

4.4 Conclusion

In purpose of improving prediction accuracy in time series forecasting, the novel hybrid model of ARIMA and ANN with discrete wavelet transform (DWT) has been developed. Its capability is tested with the sunspot, the Canadian lynx and the exchange rate time series. The experimental results imply that the proposed model can give the best performance in all three data sets and measures (i.e. MSE, MAE and MAPE).

The better forecasting accuracy indicates the advantage of combining the DWT, the ARIMA and the ANN in extracting linear and nonlinear components of the approximation and the detail without linear or nonlinear assumption. The limitation of this study is that the decomposition level of the DWT is only one. In the future works, the effect of multiple decomposition levels the will be concerned. The statistical testing of the performance measures will be conducted to confirm the significance of the improvement in prediction accuracy.

Chapter 5

Case study: Thailand's cassava export forecasting

5.1 Introduction

Cassava (*Manihot esculenta Crantz*) is a main source of calories, after rice and maize, for the world's population particularly in developing countries. The report of Food and Agriculture Organization (FAO) of the United Nations in 2012 has indicated that the cassava is the ninth rank in term of production quantity. In addition, animal foods and ethanol production for alternative energy industries use the cassava as a raw material [13]. It is worth emphasizing that most available cassava in the international market are exported from Thailand such that the accurate prediction of Thailand's future cassava export quantity can support decision makers involved in the cassava supply chain to improve production planning, policies making for helping cassava farmers, profit of cassava trading in futures market, etc. Although the future cassava export quantity from Thailand is important in the cassava international trading, to our best knowledge, so far, the researches of the cassava export forecasting are limited to using the ARIMA [24, 25]. and in Pannakkong et al. [26], we have applied the ANN for cassava export forecasting and compared its performances to the ARIMA. This chapter is a further attempt in hybridizing the ARIMA model and ANN in forecasting cassava export.

In order to determine the appropriate structure of the cassava export forecasting models, the comprehensively experimental comparisons are conducted, and then the perfor-

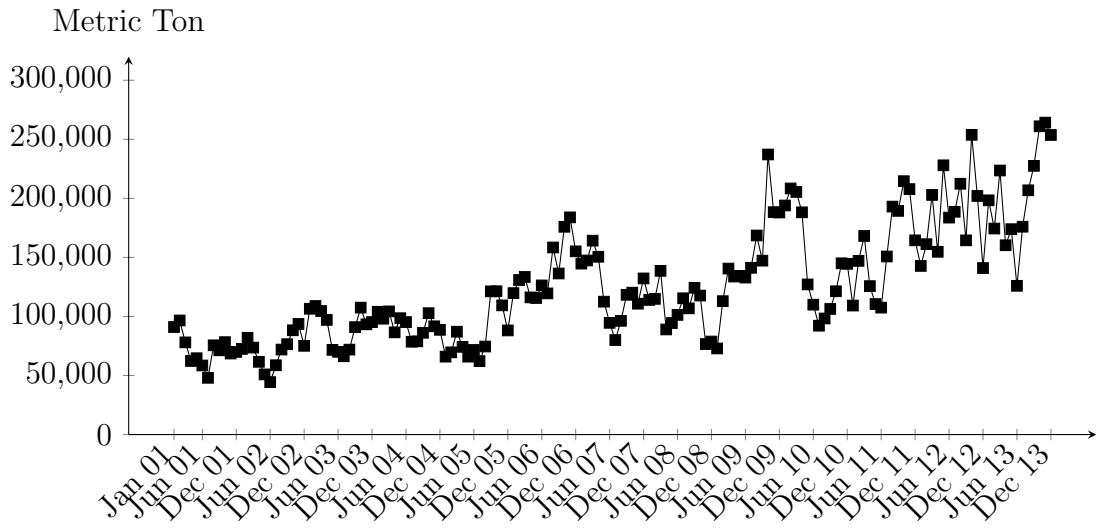

Figure 5.1: Native starch time series

mances of the forecasting models are evaluated and compared.

5.2 Cassava export time series

The cassava starch export time series contains monthly historical records of three types of the cassava export (i.e. native starch, modified starch and sago) for 13 years (2001-2013). In term of average export quantity, the native starch is the first rank followed by the modified starch and the sago. In addition, these time series are non-stationary because their mean and variance are not constant as implied in Fig. 5.1 - 5.3.

There is uncertainty in the cassava export time series. In long term, the cassava export quantity is increased due to world's population growth and need of alternative energy resources. In short term, within a year, there is somewhat annual seasonality in the cassava time series due to the relevant events which are usually repeated annually such as seasonal weather, environmental issues (e.g., drought and pests), annual cassava pawn policy of the government, selling whole pawned cassava stock of the government in every September, and farmers' behaviors in harvesting cycle.

Data partitioning is required to split the time series into training set and test set. The training set contains 144 monthly data during 2001-2012. The test set involves 12 monthly data in 2013.

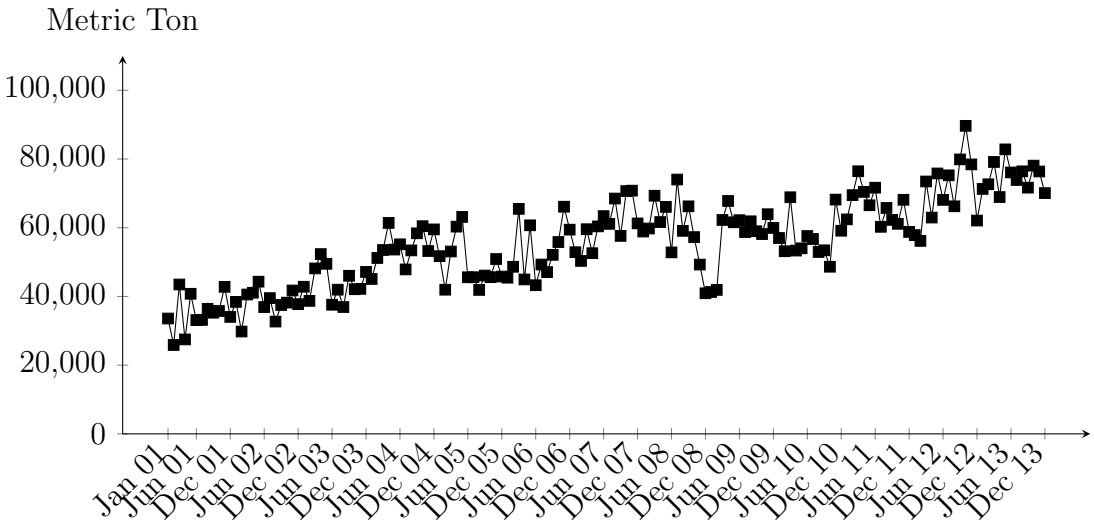

Figure 5.2: Modified starch time series

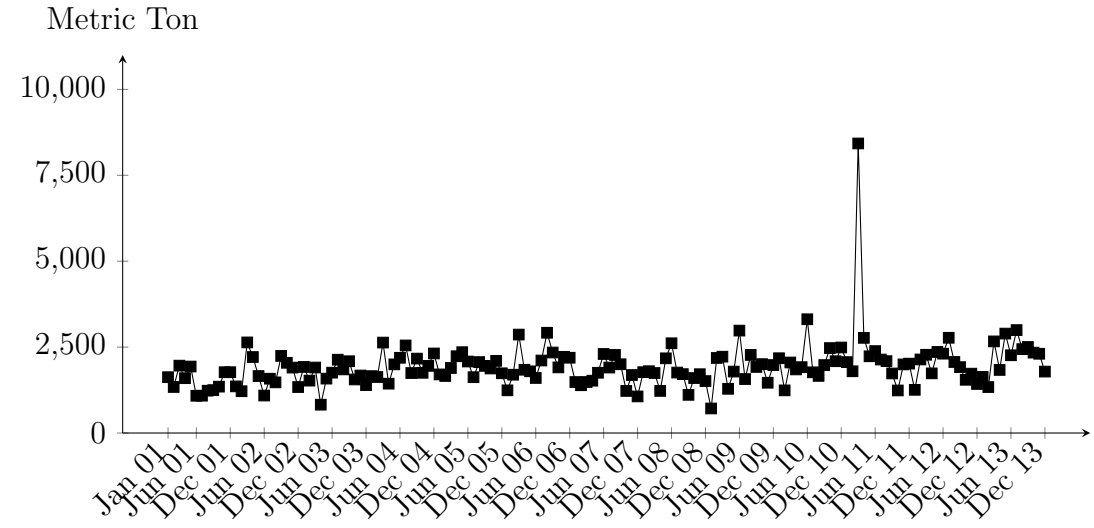

Figure 5.3: Sago time series

5.3 Hybrid model of k-mean clustering, ARIMA and ANN for Thailand's cassava export forecasting

In this section, Thailand's cassava export is predicted by the hybrid model of k-mean clustering, ARIMA and ANN model (ARIMA+ARIMA/ANN+k-means/ARIMA/ANN) from Chapter 3. Totally, there are four stages including linear-nonlinear modeling, linear-nonlinear modeling, linear-nonlinear modeling with k-mean clustering, and final forecasting.

Stage I. linear modeling: The cassava time series are analyzed by the ARIMA to obtain the linear component. ARIMA(1,1,0)$(0,1,1)_{12}$ is the best fitted model for native starch and modified starch. ARIMA(1,0,1)$(1,0,0)_{12}$ is the best fitted for the sago.

Stage II. linear-nonlinear modeling: The Khashei and Bijari's model is applied to deal with linear and nonlinear components of the cassava time series. For the native starch, the best fitted model includes the ARIMA results, seven lagged values of time series, six lagged values of ARIMA residuals, and nine hidden nodes. In case of the modified starch, the best fitted model composes of the ARIMA results, six lagged values of time series, six lagged values of ARIMA residuals, and nine hidden nodes. For the sago, the best fitted model composes of the ARIMA results, five lagged values of time series, eight lagged values of ARIMA residuals, and ten hidden nodes.

Stage III. linear-nonlinear modeling with k-mean clustering: Firstly, the training set of the cassava time series are grouped into the clusters by the k-mean clustering algorithm. According to the highest average of Silhouette (s_i), two clusters is suitable for all three kinds of cassava export as illustrated in Fig. 5.4 - 5.6.

Hence, for the native starch, the best fitted model for cluster 1 consists of the ARIMA results, two lagged values of time series, seven lagged values of the ARIMA residual, and seven hidden nodes. The best fitted model for cluster 2 includes only the ARIMA results. In case of the modified starch and the sago, the best fitted models are the same as in the State II because the predicted values in test period from the Khashei and Bijari's model shows that only the native starch has two clusters while the modified starch and the sago have only one cluster.

Stage IV. final forecasting stage: The forecasted values from stage I, II and III are

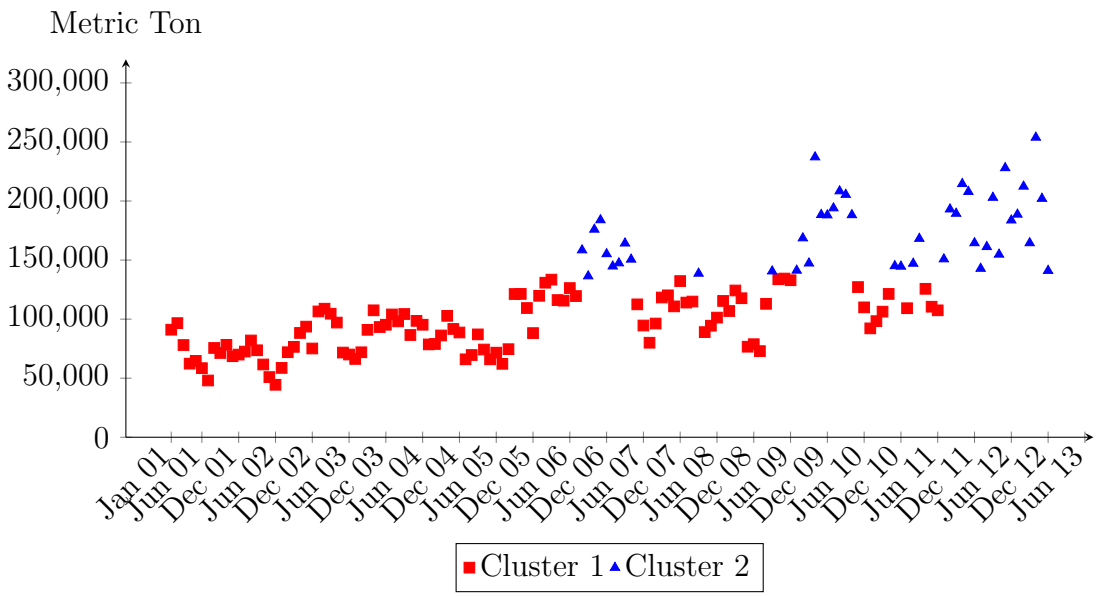

Figure 5.4: Clusters of the native starch training set

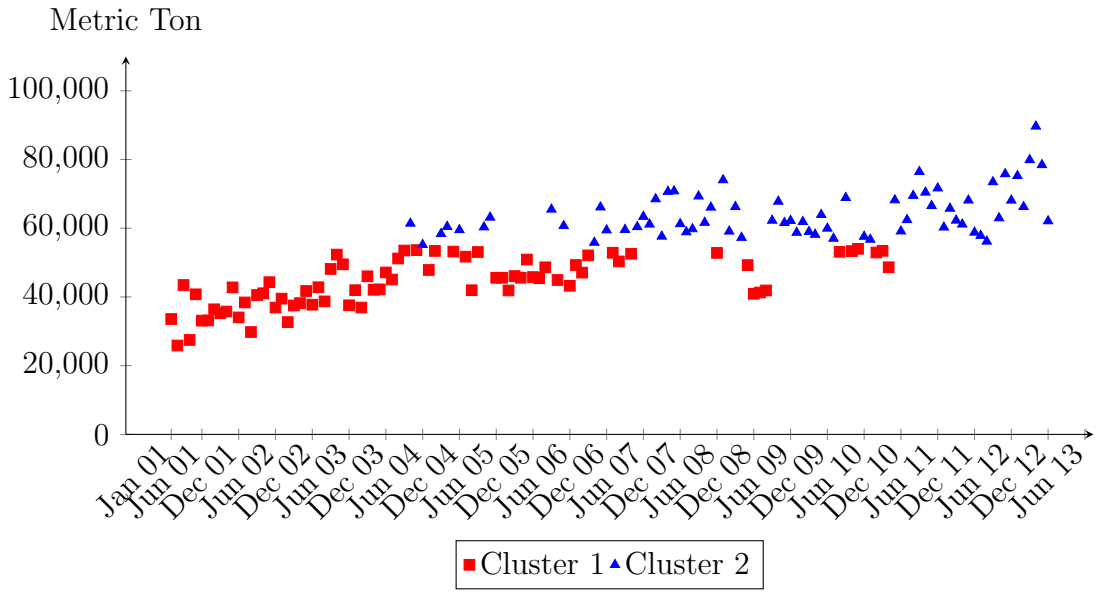

Figure 5.5: Clusters of the modified starch training set

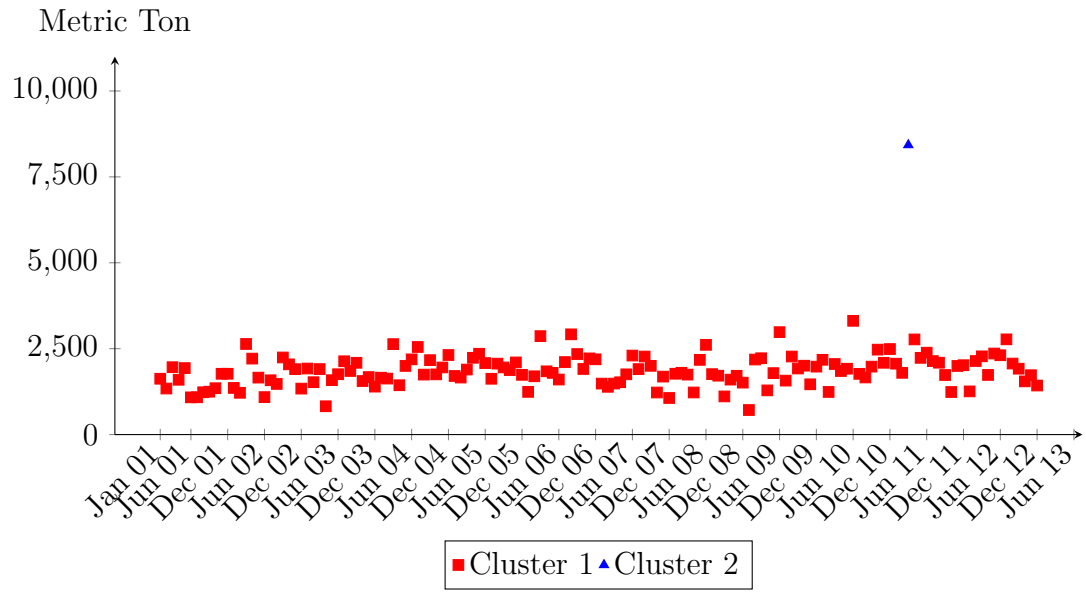

Figure 5.6: Clusters of the sago training set

combined by using the weights from the discount mean square forecast error (DMSFE) method. The performance in term of MSE, MAE, and MAPE are shown in Table 5.1.

Table 5.1: Performance of k-means/ARIMA/ANN in Thailand's cassava export forecasting

Cassava export	MSE	MAE	MAPE
Native starch	649,608,755.56	20,280.04	9.92%
Modified starch	22,650,310.02	3,904.25	5.17%
Sago	195,325.28	314.47	13.19%

5.4 Hybrid model of ARIMA and ANN with discrete wavelet transform for Thailand's cassava export forecasting

The hybrid model of ARIMA and ANN with discrete wavelet transform (DWT/ARIMA /ANN) is applied to the cassava time series. In the first step, the DWT decomposes time cassava time series (Z_t) into the approximation (Z_t^{app}) and the detail (Z_t^{det}) by using Daubechies wavelet basis function. The approximation and the detail of the cassava time series are shown in Fig. 5.7 - 5.12.

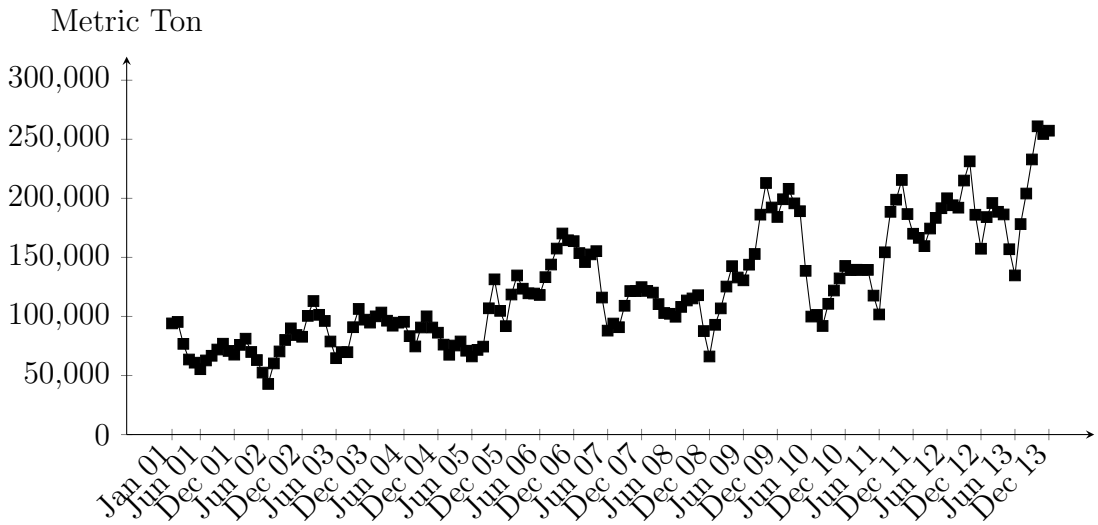

Figure 5.7: Approximation of the native starch time series

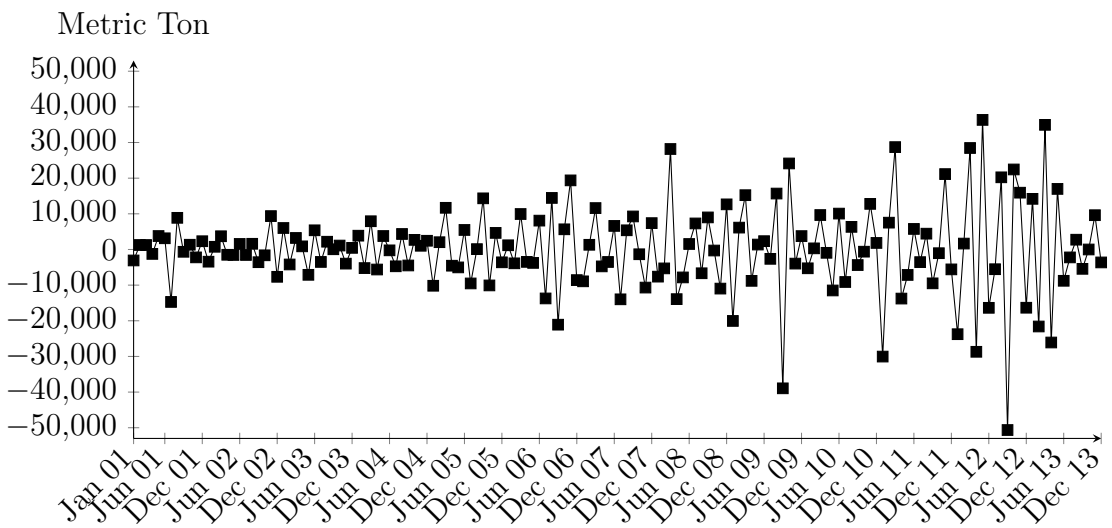

Figure 5.8: Detail of the native starch time series

In the second step, different Zhang's models are dedicatedly constructed for the approximation and the detail for each cassava export as:

$$\hat{Z}_t^{app} = \hat{L}_t^{app} + \hat{N}_t^{app} \tag{5.1}$$

$$\hat{Z}_t^{det} = \hat{L}_t^{det} + \hat{N}_t^{det} \tag{5.2}$$

where $\hat{Z}_t^{app}$ and $\hat{Z}_t^{det}$ are the predicted approximation and detail respectively, at period t; $\hat{L}_t^{app}$ and $\hat{L}_t^{det}$ are linear components of the approximation and the detail respectively, at

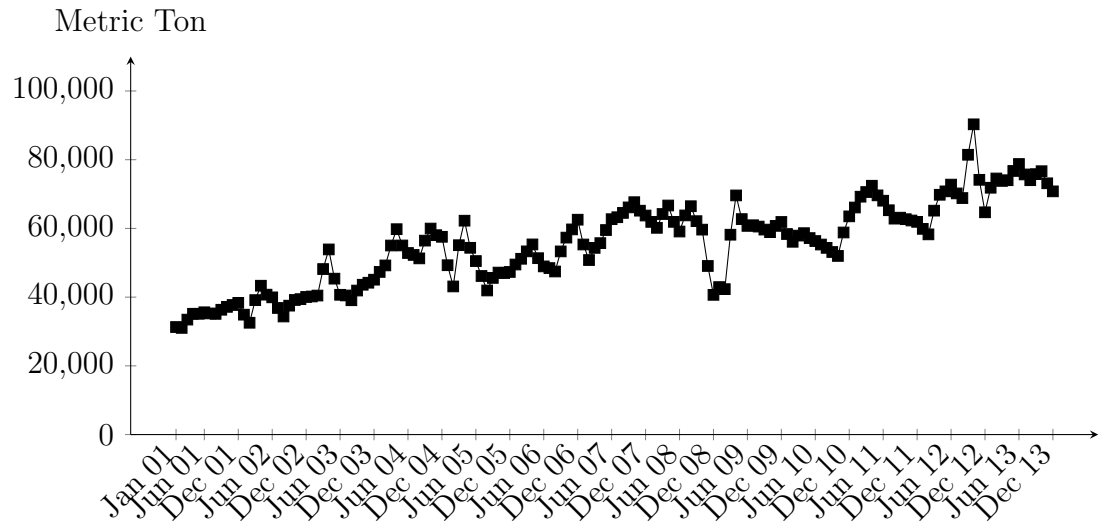

Figure 5.9: Approximation of the modified starch time series

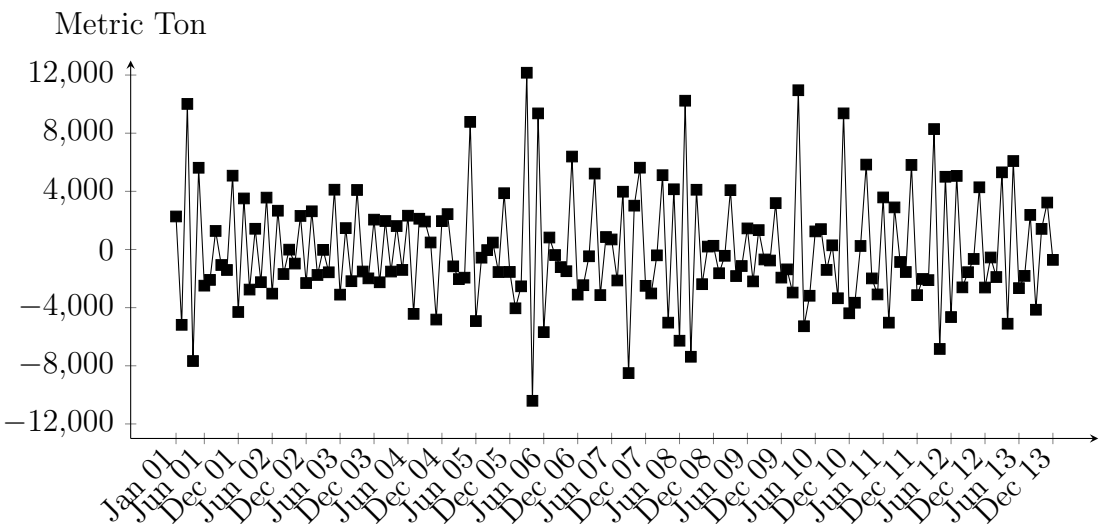

Figure 5.10: Detail of the modified starch time series

period t; $\hat{N}_t^{app}$ and $\hat{N}_t^{det}$ are nonlinear components of of the approximation and the detail respectively, at period t.

The linear components ($\hat{L}_t^{app}$ and $\hat{L}_t^{det}$) are the results of applying the ARIMA to Z_t^{app} and Z_t^{det} respectively. Then, the ARIMA residuals of the approximation (e_t^{app}) and the detail (e_t^{det}) can be computed as:

$$e_t^{app} = Z_t^{app} - \hat{L}_t^{app} \tag{5.3}$$

$$e_t^{det} = Z_t^{det} - \hat{L}_t^{det} \tag{5.4}$$

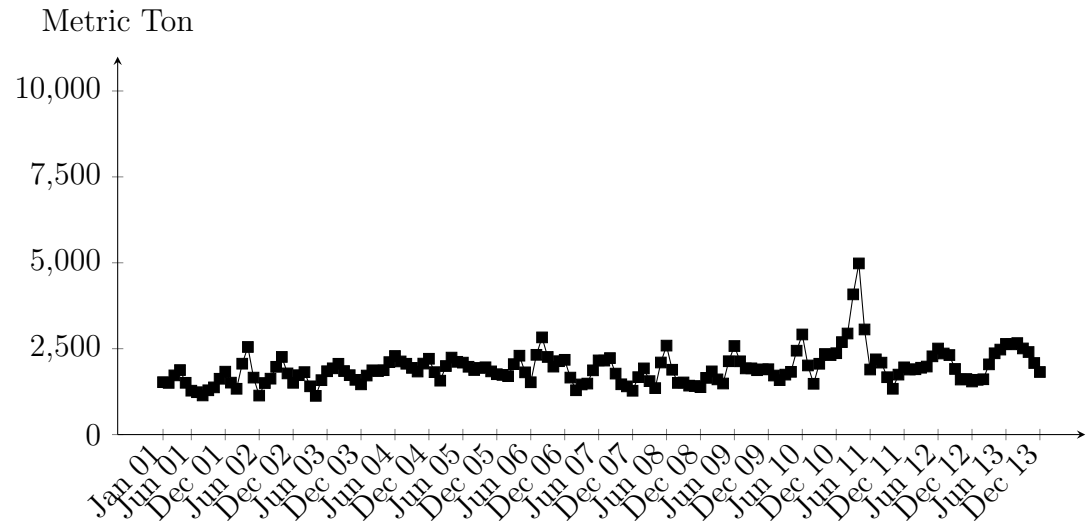

Figure 5.11: Approximation of the sago time series

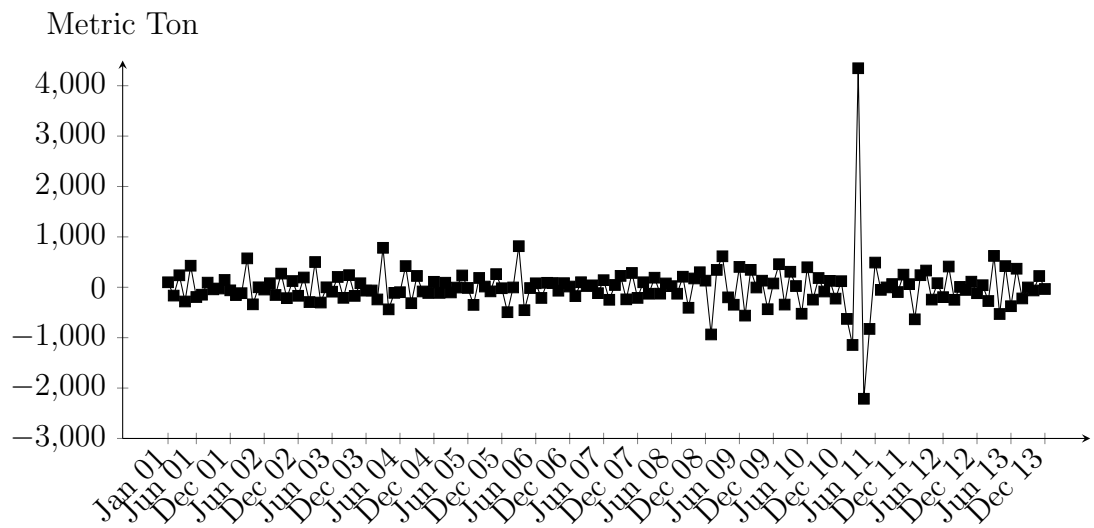

Figure 5.12: Detail of the sago time series

For the nonlinear components ($\hat{N}_t^{app}$ and $\hat{N}_t^{det}$), they are obtained from the ANN as:

$$\hat{N}_t^{app} = f^{app}(e_{t-1}^{app}, e_{t-2}^{app}, \ldots, e_{t-n_1}^{app}) \tag{5.5}$$

$$\hat{N}_t^{det} = f^{det}(e_{t-1}^{det}, e_{t-2}^{det}, \ldots, e_{t-n_2}^{det}) \tag{5.6}$$

where f^{app} and f^{det} are functions fitted by the ANN, n_1 and n_2 are total lagged periods which are determined by trial and error in the experiments.

At this step, we obtain two linear ($\hat{L}_t^{app}$ and $\hat{L}_t^{det}$) and two nonlinear components

($\hat{N}_t^{app}$ and $\hat{N}_t^{det}$). Finally, the final forecasting can be done by aggregation of the linear and nonlinear components as:

$$\begin{aligned}\hat{Z}_t &= \hat{Z}_t^{app} + \hat{Z}_t^{det} \\ &= \hat{L}_t^{app} + \hat{N}_t^{app} + \hat{L}_t^{det} + \hat{N}_t^{det}\end{aligned} \tag{5.7}$$

Finally, the final forecasting results of the cassava time series are compare with the actual cassava time series to evaluate prediction accuracy (i.e. MSE, MAE, and MAPE) as shown in Table 5.2.

Table 5.2: Performance of DWT/ARIMA/ANN in Thailand's cassava export forecasting

Cassava export	MSE	MAE	MAPE
Native starch	537,731,682.15	17,528.39	9.88%
Modified starch	21,930,118.85	3,627.60	4.74%
Sago	51,055.45	188.00	8.83%

5.5 Hybrid model of ARIMA and ANN with preprocessed variables for Thailand's cassava export forecasting

The hybrid models such as the Zhang's hybrid model [14] and the Khashei and Bijari's hybrid model [16] have performed well in time series forecasting, it is of interest to note that, however, these hybrid models consider only lagged values of the time series as their input, there may be an opportunity to improve their prediction quality by including processed variables such as moving average and annual seasonal index into the models as in [26].

In this section, we propose a new hybrid model that also combines the ARIMA and the ANN but additionally incorporates the moving average (MA) and the annual seasonal index into the model for Thailand's cassava export forecasting, as graphically depicted in Fig. 5.13. The MA are the averages of previous N period of time series. The time series included in the MA correspond to the current period (t). The mathematical formula of the MA can be expressed as:

$$\mathrm{MA}_t(N) = \sum_{n=1}^{N} Z_{t-n} \tag{5.8}$$

where $\mathrm{MA}_t(N)$ is the average time series of pervious N periods at current period t, and N is number of previous periods included in the moving average.

The seasonal index is the ratio between the time series at considered period and the average of all time series in the cycle. The number of periods for each cycle is the length of cyclic pattern that can be identifed by autocorrelation analysis of the time series. The seasonal index can be computed as below:

$$\text{Seasonal index}_t = \frac{Z_t}{\sum_{p=1}^{s} Z_{(p+((nc-1)\times s))}} \tag{5.9}$$

where Seasonal index$_t$ is the seasonal index at period t, s is the number of periods in the cycle (which is 12 months), nc is the number of current cycle, and p is the period in the cycle. For the seasonal index of future time series which is unknown, we use the average of seasonal index at the same period in all previous cycles.

In this study, the SARIMA model with 12 months seasonal time span is applied to the cassava starch export time series due to the characteristic of the cassava as an agricultural product that is influenced by the factors which have annual seasonal pattern such as seasonal weather, harvesting cycle and government policy.

Box et al. [29] described a manual approach for obtaining best-fit parameters of the ARIMA. Nonetheless, in this thesis, the best-fit parameters are automatically determined by IBM SPSS Statistics software such that it can reduce error from manual selection of the parameters. In addition, we can also construct hundred scenarios of the models and compare them at once rather than doing it manually for one model at a time.

Totally, there are nine input variables; six of them are the technical variables: three lagged values at time $t-1, t-3$ and $t-12$ (denoted by Z_{t-1}, Z_{t-3} and Z_{t-12}); two moving averages with three and twelve previous periods (denoted by MA(3) and MA(12)); and an annual seasonal index. The remaining variables are fundamental variables corresponding to three time indices: number of period (Sequence), month in year (Month), and number of quarter (Quarter).

Table 5.3: ANN inputs for native starch and modified starch

Scenario 1: All	Scenario 2: Correlated
Sequence	Sequence
Month	Z_{t-1}
Quarter	Z_{t-3}
Z_{t-1}	Z_{t-12}
Z_{t-3}	MA(3)
Z_{t-12}	MA(12)
MA(3)	Seasonal Index
MA(12)	
Seasonal Index	

Table 5.4: ANN inputs for sago

Scenario 1: All	Scenario 2: Correlated
Sequence	Sequence
Month	MA(3)
Quarter	MA(12)
Z_{t-1}	Seasonal Index
Z_{t-3}	
Z_{t-12}	
MA(3)	
MA(12)	
Seasonal Index	

The correlation analysis is performed at first to screen the variables that have statistically significant correlation with the actual cassava export. Then, the variables that can pass the correlation analysis are included into the ANN. The all inputs and the significant correlated inputs of each cassava export are shown in Tables 5.3 and 5.4.

The proposed model starts with separating the cassava time series into training and test set. Then, the linear and nonlinear components of the training set are determined as in the Khashei and Bijari's model.

Additionally, instead of taking only lagged values of time series into consideration, the proposed model also includes the correlated variables shown in the second column of Tables 5.3 and 5.4 (denoted by $C_t^1, \ldots, C_t^o$) as an additional nonlinear component N_t^3 as defined in (5.10).

$$N_t^3 = f^3(C_t^1, \ldots, C_t^o) \tag{5.10}$$

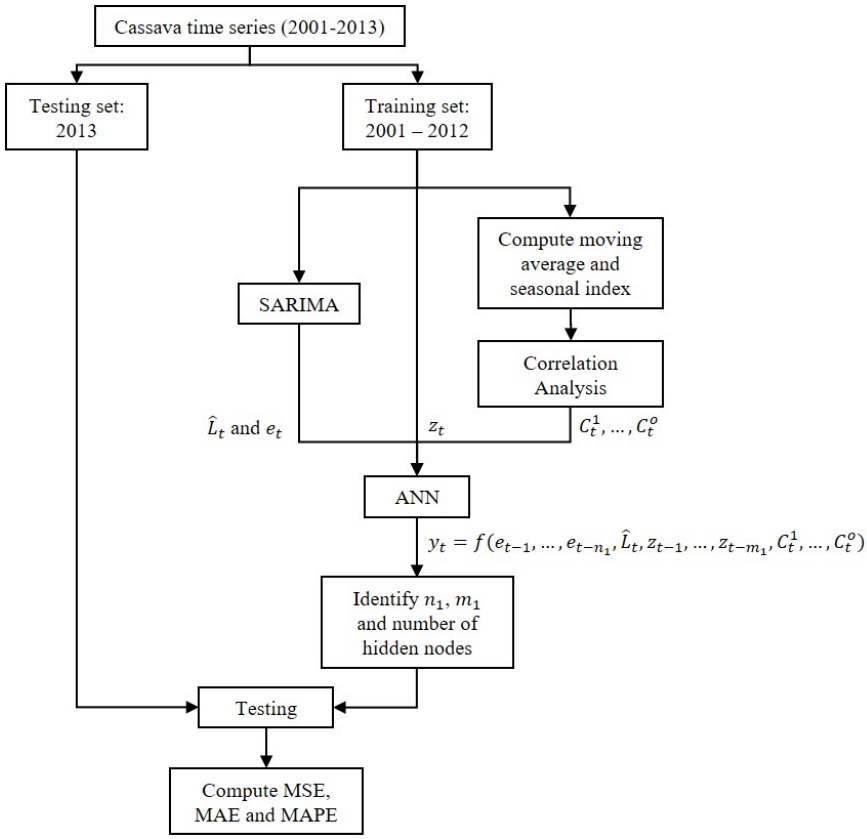

Figure 5.13: The proposed hybrid ARIMA and ANN model

where f^3 is the nonlinear function determined by the ANN. The number of statistically correlated significant variables is o. Eventually, the proposed model can be written as in (5.11) below:

$$\begin{aligned} Z_t &= f(\hat{L}_t, N_t^1, N_t^2, N_t^3) \\ &= f(\hat{L}_t, e_{t-1}, \ldots, e_{t-n_1}, Z_{t-1}, \ldots, Z_{t-m_1}, C_t^1, \ldots, C_t^o) \end{aligned} \tag{5.11}$$

where f is the nonlinear function determined by the ANN; $n_1 \leq n$ and $m_1 \leq m$ are integers determined in the design process.

In general, the number of input nodes corresponds to the number of the lagged values, which is used to discover underlying pattern in a time series and to make forecasts for future values [30]. Therefore, n_1 and m_1 are varied in the experiments from one to n and m, which are set to 12 because of intention to cover annual seasonal pattern in time series.

To determine the appropriate parameters of the proposed model, firstly, the results of

the ARIMA are generated and the residuals are computed. Secondly, the suitable n_1 and m_1 are identified by running the proposed model while varying n_1, m_1 and the number of hidden nodes from one node to 10 nodes. The forecasting results are compared with the testing set to evaluated the performance. The values of n_1 and m_1 that give the lowest MAPE are considered as the suitable parameters for the proposed model. Finally, in order to find the suitable number of hidden nodes, the proposed model is run while the number of hidden modes are varied as in the second step.

After the comprehensive experimental runs, the forecasting accuracy measures are computed and compared. The MSE, MAE, and MAPE of the best models in each cassava type are presented in Table 5.5.

Table 5.5: Performance of ARIMA/ANN with pre-processed variables in Thailand's cassava export forecasting

Cassava export	MSE	MAE	MAPE
Native starch	810,312,946.46	23,727.98	11.91%
Modified starch	8,004,634.46	2,116.81	2.83%
Sago	192,689.30	366.61	15.74%

5.6 Performance comparison

This section summarizes the forecasting performance of the three hybrid models present in this chapter (Table 5.6), and also compare to the existing forecasting models such as the ARIMA, the ANN, the ANN with pre-processed variable, the Zhang's model, and Khashei and Bijari's model. Furthermore, in order to show further detail of characteristic of the best model, its forecasted values in testing period are presented and interpreted (Fig. 5.14 - 5.16).

The best model for the native starch is the DWT/ARIMA/ANN followed by the k-means/ARIMA/ANN. These two novel hybrid models can outperform the others models in all three accuracy measures. From Table 5.14, the DWT/ARIMA/ANN model can track the direction of the cassava export for whole year except only four months (i.e. January, February, June, and September). Zhang's model performs similar to the ARIMA that can capture only rough trend of the native starch. On the other hand, although the Khashei and Bijari's model is the hybridization of the ARIMA and the ANN, it is more

effective than the Zhang's model in tracking the direction of the actual export. This result emphasizes the draw back of the additive relationship assumption of the Zhang's model.

For the modified starch, the best model is the ARIMA/ANN with pre-processed variables followed by the ANN with pre-processed variable (Table 5.6). However, the accuracy of the ARIMA/ANN with pre-processed variables is better than the ANN with pre-processed variable around two times. Based on the results shown in Table 5.15, the ARIMA/ANN with pre-processed variables is capable to track the direction of the actual values for whole year except August and September. The other traditional models are obviously underperform. The significant improvement of prediction accuracy implies the importance of the pre-processed variables (e.g. moving average and seasonal index)

In case of the sago, the best model is the DWT/ARIMA/ANN which is significantly outperform the other models in all three measures (Table 5.6). In Table 5.16, the results show that the DWT/ARIMA/ANN can follow the direction of the actual export in several months (i.e. February, March, May, July, September, and November) while the other models cannot.

Table 5.6: Performance comparison

Cassava Export	Forecasting Model	MSE	MAE	MAPE
Native Starch	ARIMA	730,862,890.29	23,886.84	12.51%
	ANN	1,800,857,623.61	38,710.34	19.19%
	ANN with pre-processed variables [26]	930,342,573.39	25,552.21	12.96%
	Zhang [14]	659,898,682.79	22,647.13	11.96%
	Khashei and Bijari [16]	935,361,300.68	22,647.41	10.75%
	ARIMA/ANN with pre-processed variables	810,312,946.46	23,727.98	11.91%
	DWT/ARIMA/ANN	**537,731,682.15**	**17,528.39**	**9.88%**
	ARIMA+ARIMA/ANN+k-means/ARIMA/ANN	649,608,755.56	20,280.04	9.92%
Modified Starch	ARIMA	26,601,773.21	4,626.13	6.21%
	ANN	23,109,442.25	4,372.43	5.79%
	ANN with pre-processed variables [26]	17,614,154.64	3,384.87	4.51%
	Zhang [14]	27,202,943.30	4,430.75	5.99%
	Khashei and Bijari [16]	25,605,091.84	4,148.98	5.47%
	ARIMA/ANN with pre-processed variables	**8,004,634.46**	**2,116.81**	**2.83%**
	DWT/ARIMA/ANN	21,930,118.85	3,627.60	4.74%
	ARIMA+ARIMA/ANN+k-means/ARIMA/ANN	22,650,310.02	3,904.25	5.17%
Sago	ARIMA	258,205.54	415.22	17.84%
	ANN	294,645.20	455.55	19.57%
	ANN with pre-processed variables [26]	193,225.72	354.90	15.16%
	Zhang [14]	252,833.13	410.72	17.81%
	Khashei and Bijari [16]	184,196.20	303.08	12.67%
	ARIMA/ANN with pre-processed variables	192,689.30	366.61	15.74%
	DWT/ARIMA/ANN	**51,055.45**	**188.00**	**8.83%**
	ARIMA+ARIMA/ANN+k-means/ARIMA/ANN	195,325.28	314.47	13.19%

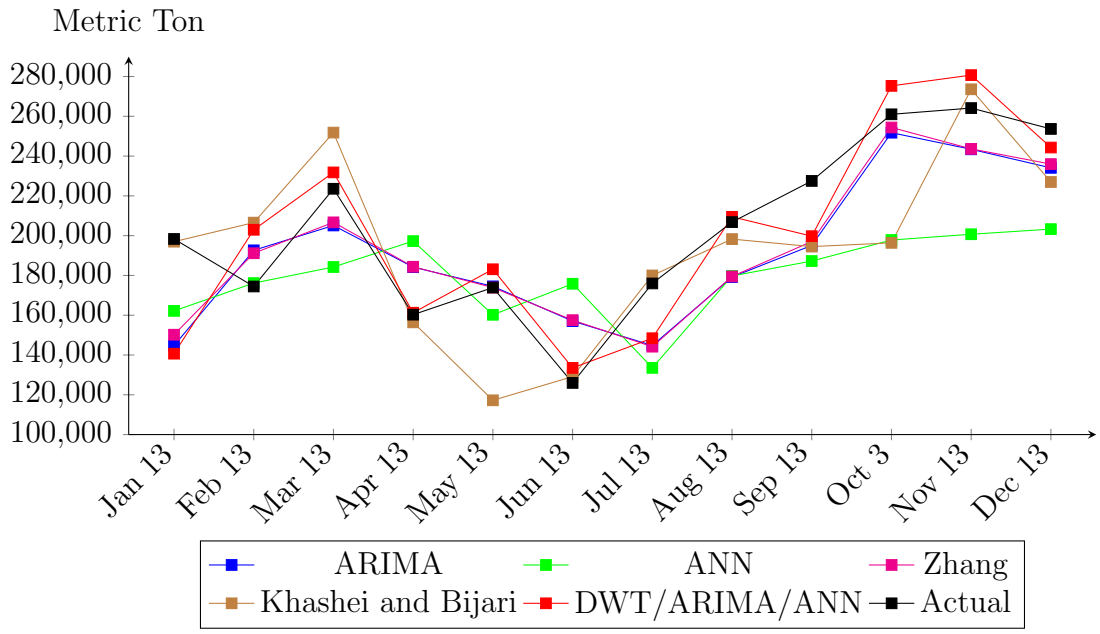

Figure 5.14: Forecasting export comparison for native starch

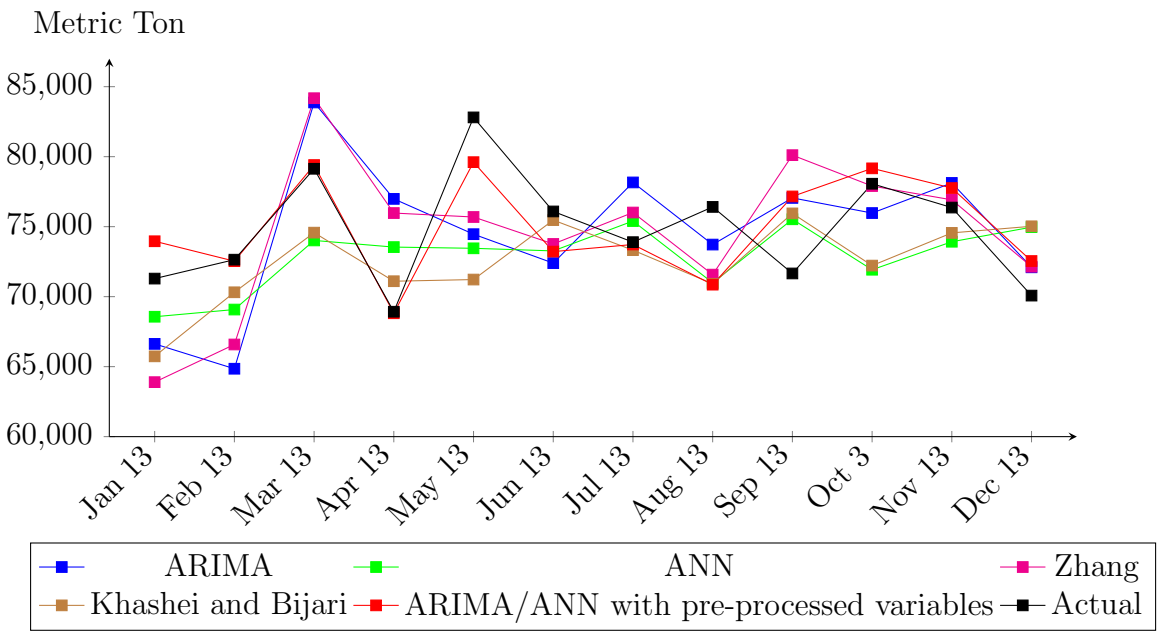

Figure 5.15: Forecasting export comparison for modified starch

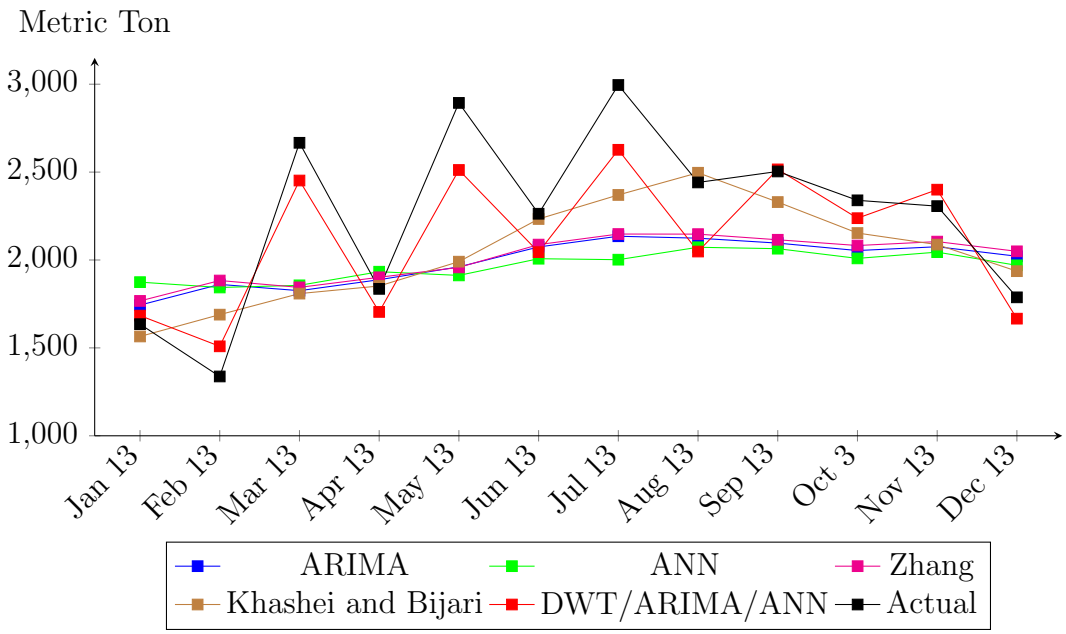

Figure 5.16: Forecasting export comparison for sago

5.7 Conclusion

The two novel hybrid models developed in this thesis (i.e. the DWT/ARIMA/ANN and the ARIMA/ANN with pre-processed variables) shows their capability in the cassava export forecasting with the lowest error in MSE, MAE and MAPE than the other models.

Regarding the capability of the proposed model, the proposed model can contribute the cassava international trading market as an alternative forecasting approaches giving better forecasting performance for Thailand's cassava export forecasting.

However, the ARIMA+ARIMA/ANN+k-means/ARIMA/ANN may not be suitable for the modified starch and the sago because the k-means clustering classifies the time series based on different in average rather than different pattern. In case of the modified starch, the average is changed in long term but in short term, the variation of the export volume is quite low. Therefore, the clusters represent the groups of the older time series which have lower average, and the newer time series which have higher average. For sago, the k-mean separates the time series from the special outlier.

Furthermore, as the cassava is a commodity product, the DWT/ARIMA/ANN and the ARIMA/ANN with pre-processed variables could be applied to other commodity products that share the similar characteristic as well.

Chapter 6

Thesis contribution

6.1 Practical implication

1. The more accurate prediction of the cassava exported from Thailand to the international market encourages more efficient planning in cassava supply chain management (e.g. balancing supply and demand, stabilizing the cassava price, preventing shortage of the cassava. These benefits would impact to a large amount of people because the cassava is one of the main food for world population.

2. The proposed hybrid models in this thesis can support decision making (e.g. buying and selling) of the stakeholders involved in the cassava supply chain.

3. Thai government can use the prediction results from the proposed hybrid models in order to make suitable policies for helping over 500,000 cassava farmer households in Thailand.

6.2 Theoretical implication

1. The first academic contribution is that a novel hybrid forecasting model between k-means, ARIMA, and ANN has been developed.

   (a) A recent work proposed to use the summation of the prediction values from every cluster [61], but in fact, it would be more logical if the future values are

produced their cluster. Therefore, a method to predict cluster of future time series has been proposed.

(b) In most studies on hybrid forecasting, the hybrid models are combinations of the same level of complexity. This study proposed a hybrid model that takes advantage of unique strength of the three approaches which have different of complexity such as linear, hybrid linear-nonlinear, and hybrid linear-nonlinear with clustering technique.

2. The second academic contribution is that a novel hybrid forecasting model of ARIMA, and ANN with DWT has been proposed.

   Although, the wavelet transform has been combined with ARIMA and ANN models [79], however, this approach assumed that the approximation contains only nonlinear component. Such assumption is not practical since the DWT is not a method to transform time series into linear or nonlinear components. Hence, we proposed the forecasting model capturing both linear and nonlinear components of the detail and the approximation in stead of original time series.

3. The third academic contribution is that hybrid models for Thailand's cassava export forecasting have been developed.

   Cassava is a main source of calories, after rice and maize, for the world's population particularly in developing countries. Most available cassava in the international market are exported from Thailand. However, to our best knowledge, there is no study of a hybrid model originated for Thailands cassava export forecasting.

6.3 Contribution to Knowledge Science

1. The results from this research enrich the knowledge in time series forecasting research area by proposing the novel hybrid models.

2. The novel hybrid models contribute knowledge discovery process as the data mining processes capturing the knowledge from the complex data such as real-world time series.

Chapter 7

Conclusion and future work

7.1 Conclusion

Improving time series forecasting is an important yet often difficult task. Although, the numerous time series forecasting models have been developed, but the research for developing new forecasting models to improve the prediction accuracy has never stopped.

In chapter 3, the hybrid model of the ARIMA, the Khashei and Bijari's model (ARIMA/ANN model) and the k-means/ARIMA/ANN model has been developed to gain unique advantages among the different model types and complexity levels in time series forecasting. The prediction capability of the proposed model is tested with well-known data sets: the Wolf's sunspot, the Canadian lynx, and the British pound per US dollar exchange rate. From the empirical results, the proposed model gives the best performance in MSE, MAE, and MAPE for all three data sets.

In chapter 4, the novel hybrid model of ARIMA and ANN with discrete wavelet transform (DWT) has been developed. Its capability is also tested with the sunspot, the Canadian lynx and the exchange rate time series. The experimental results imply that the proposed model can give the best performance in all three data sets and measures (i.e. MSE, MAE and MAPE). The improvement in forecasting accuracy implies the benefit of combining the DWT, the ARIMA and the ANN in capturing linear and nonlinear components of the approximation and the detail without linear or nonlinear assumption.

In chapter 5, the hybrid models in Chapter 3 and 4 are applied to Thailand's cassava export as the case study. In addition, another hybrid model of ARIMA and ANN

with pre-processed variables for the cassava export has been proposed as well. The DWT/ARIMA/ANN is the best model for the native starch and the sago. On the other hand, the hybrid model of ARIMA and ANN with pre-processed variables is the best model for the modified starch.

In conclusion, the proposed hybrid models have shown their forecasting capability over the traditional single and hybrid models. Therefore, they can be used as alternative models for time series prediction. The stakeholders involves in the cassava supply chain can apply the hybrid models specified for each type of the cassava export to obtain more accurate prediction results of Thailand's cassava export. Furthermore, the hybrid models for cassava forecasting can be applied to other commodity products sharing the similar characteristic with the cassava as well.

7.2 Future work

According to limitations of the experiments in this thesis, there are opportunities to conduct further research to improve the performance of the hybrid models.

1. The proposed model in Chapter 3 includes only one model per each model type (i.e. the ARIMA as the linear model, the ANN as the nonlinear model, and the k-means as the clustering technique). Actually, there are many other linear models, nonlinear models, and clustering techniques that can be involved the future study.

2. The drawback of the proposed model in Chapter 3 is that in stage III, the forecasted clusters rely on the forecasted values from the ARIMA/ANN model that can makes the error causing the wrong predicted cluster especially when the forecasted values are very close to the boundary between the clusters. In future work, fuzzy clustering can be applied to overcome such limitation by assigning the clusters of the future values as probability to reduce the effect of the wrong cluster forecasting.

3. The level of transformation of the DWT in Chapter 4 is only the first level. The approach determining an optimal level of the transformation is worth to be considered as a future work.

Bibliography

[1] J. A. Dayhoff, *Neural Network Architectures: An Introduction.* MIT press, 1995.

[2] L.-Y. Wei, "A hybrid anfis model based on empirical mode decomposition for stock time series forecasting," *Applied Soft Computing*, vol. 42, pp. 368–376, 2016.

[3] R. Adhikari and R. Agrawal, "A combination of artificial neural network and random walk models for financial time series forecasting," *Neural Computing and Applications*, vol. 24, no. 6, pp. 1441–1449, 2014.

[4] H. Ezzine, A. Bouziane, and D. Ouazar, "Seasonal comparisons of meteorological and agricultural drought indices in morocco using open short time-series data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 26, pp. 36–48, 2014.

[5] K. A. Garrett, A. Dobson, J. Kroschel, B. Natarajan, S. Orlandini, H. E. Tonnang, and C. Valdivia, "The effects of climate variability and the color of weather time series on agricultural diseases and pests, and on decisions for their management," *Agricultural and Forest Meteorology*, vol. 170, pp. 216–227, 2013.

[6] H. J. Sadaei, R. Enayatifar, A. H. Abdullah, and A. Gani, "Short-term load forecasting using a hybrid model with a refined exponentially weighted fuzzy time series and an improved harmony search," *International Journal of Electrical Power & Energy Systems*, vol. 62, pp. 118–129, 2014.

[7] S. Bahrami, R.-A. Hooshmand, and M. Parastegari, "Short term electric load forecasting by wavelet transform and grey model improved by pso (particle swarm optimization) algorithm," *Energy*, vol. 72, pp. 434–442, 2014.

[8] Y. Xiao, J. Xiao, and S. Wang, "A hybrid forecasting model for non-stationary time series: An application to container throughput prediction," *International Journal of Knowledge and Systems Science (IJKSS)*, vol. 3, no. 2, pp. 67–82, 2012.

[9] C. Zhang, L. Huang, and Z. Zhao, "Research on combination forecast of port cargo throughput based on time series and causality analysis," *Journal of Industrial Engineering and Management*, vol. 6, no. 1, p. 124, 2013.

[10] W. Deng, G. Wang, and X. Zhang, "A novel hybrid water quality time series prediction method based on cloud model and fuzzy forecasting," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 39–49, 2015.

[11] X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang, "Artificial neural networks forecasting of pm 2.5 pollution using air mass trajectory based geographic model and wavelet transformation," *Atmospheric Environment*, vol. 107, pp. 118–128, 2015.

[12] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," *International Journal of Forecasting*, vol. 22, no. 3, pp. 443–473, 2006.

[13] Food and Agriculture Organization of the United Nations, "Why cassava," Apr. 2015.

[14] G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[15] T. Taskaya-Temizel and M. C. Casey, "A comparative study of autoregressive neural network hybrids," *Neural Networks*, vol. 18, no. 5, pp. 781–789, 2005.

[16] M. Khashei and M. Bijari, "A novel hybridization of artificial neural networks and arima models for time series forecasting," *Applied Soft Computing*, vol. 11, no. 2, pp. 2664–2675, 2011.

[17] J.-J. Wang, J.-Z. Wang, Z.-G. Zhang, and S.-P. Guo, "Stock index forecasting based on a hybrid model," *Omega*, vol. 40, no. 6, pp. 758–766, 2012.

[18] T. Shahwan and M. Odening, "Forecasting agricultural commodity prices using hybrid neural networks," in *Computational intelligence in economics and finance*, pp. 63–74, Springer, 2007.

[19] H. Liu, D. Xie, and W. Wu, "Soil water content forecasting by ann and svm hybrid architecture," *Environmental monitoring and assessment*, vol. 143, no. 1, pp. 187–193, 2008.

[20] Y.-S. Lee and W.-Y. Liu, "Forecasting value of agricultural imports using a novel two-stage hybrid model," *Computers and Electronics in Agriculture*, vol. 104, pp. 71–83, 2014.

[21] I. Pulido-Calvo and J. C. Gutierrez-Estrada, "Improved irrigation water demand forecasting using a soft-computing hybrid model," *Biosystems Engineering*, vol. 102, no. 2, pp. 202–218, 2009.

[22] C. O. Ribeiro and S. M. Oliveira, "A hybrid commodity price-forecasting model applied to the sugar–alcohol sector," *Australian Journal of Agricultural and Resource Economics*, vol. 55, no. 2, pp. 180–198, 2011.

[23] W. Bo, W. Shouyang, and K. Lai, "A hybrid arch-m and bp neural network model for gsci futures price forecasting," *Computational Science–ICCS 2007*, pp. 917–924, 2007.

[24] C. Kongcharoen and T. Kruangpradit, "Autoregressive integrated moving average with explanatory variable (arimax) model for thailand export," in *33rd International Symposium on Forecasting, South Korea*, pp. 1–8, 2013.

[25] C. Prapasornpittaya, "Comparative study on arima, intervention and transfer function models in forecasting thailand's export value," unpublished master's thesis, Thammasat University, Pathum Thani, Thailand, 2013.

[26] W. Pannakkong, V.-N. Huynh, and S. Sriboonchitta, "Arima versus artificial neural network for thailands cassava starch export forecasting," in *Causal Inference in Econometrics*, pp. 255–277, Springer, 2016.

[27] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA., 1967.

[28] N. J. Fliege, *Multirate digital signal processing: multirate systems, filter banks, wavelets*. John Wiley & Sons, Inc., 1994.

[29] G. Box, G. Jenkins, and G. Reinsel, *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics, Wiley, 2008.

[30] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *International Journal of Forecasting*, vol. 14, no. 1, pp. 35–62, 1998.

[31] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural Networks, 1989. IJCNN., International Joint Conference on*, pp. 593–605, IEEE, 1989.

[32] I. Kaastra and M. Boyd, "Designing a neural network for forecasting financial and economic time series," *Neurocomputing*, vol. 10, no. 3, pp. 215–236, 1996.

[33] D. J. MacKay, "A practical bayesian framework for backpropagation networks," *Neural Computation*, vol. 4, no. 3, pp. 448–472, 1992.

[34] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[35] D. Ö. Faruk, "A hybrid neural network and arima model for water quality time series prediction," *Engineering Applications of Artificial Intelligence*, vol. 23, no. 4, pp. 586–594, 2010.

[36] P.-F. Pai and C.-S. Lin, "A hybrid arima and support vector machines model in stock price forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, 2005.

[37] C.-L. L. Chen-Chun Lin and J. Z. Shyu, "Hybrid multi-model forecasting system: A case study on display market," *Knowledge-Based Systems*, vol. 71, pp. 279 – 289, 2014.

[38] Y. He, Y. Zhu, and D. Duan, "Research on hybrid arima and support vector machine model in short term load forecasting," in *Sixth International Conference on Intelligent Systems Design and Applications*, vol. 1, pp. 804–809, IEEE, 2006.

[39] M. Bouzerdoum, A. Mellit, and A. M. Pavan, "A hybrid model (sarima–svm) for short-term power forecasting of a small-scale grid-connected photovoltaic plant," *Solar Energy*, vol. 98, pp. 226–235, 2013.

[40] J. Zhang, Z. Tan, and S. Yang, "Day-ahead electricity price forecasting by a new hybrid method," *Computers & Industrial Engineering*, vol. 63, no. 3, pp. 695–701, 2012.

[41] M. Shafie-Khah, M. P. Moghaddam, and M. Sheikh-El-Eslami, "Price forecasting of day-ahead electricity markets using a hybrid forecast method," *Energy Conversion and Management*, vol. 52, no. 5, pp. 2165–2169, 2011.

[42] N. Chaâbane, "A hybrid arfima and neural network model for electricity price prediction," *International Journal of Electrical Power & Energy Systems*, vol. 55, pp. 187–194, 2014.

[43] K.-Y. Chen and C.-H. Wang, "A hybrid sarima and support vector machines in forecasting the production values of the machinery industry in taiwan," *Expert Systems with Applications*, vol. 32, no. 1, pp. 254–264, 2007.

[44] K.-Y. Chen, "Combining linear and nonlinear model in forecasting tourism demand," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10368–10376, 2011.

[45] A. Aslanargun, M. Mammadov, B. Yazici, and S. Yolacan, "Comparison of arima, neural networks and hybrid models in time series: tourist arrival forecasting," *Journal of Statistical Computation and Simulation*, vol. 77, no. 1, pp. 29–53, 2007.

[46] J.-H. Lo, "A study of applying arima and svm model to software reliability prediction," in *Uncertainty Reasoning and Knowledge Engineering (URKE), 2011 International Conference on*, vol. 1, pp. 141–144, IEEE, 2011.

[47] J. Ruiz-Aguilar, I. Turias, and M. Jiménez-Come, "Hybrid approaches based on sarima and artificial neural networks for inspection time series forecasting," *Transportation Research Part E: Logistics and Transportation Review*, vol. 67, pp. 1–13, 2014.

[48] J. Shi, J. Guo, and S. Zheng, "Evaluation of hybrid forecasting approaches for wind speed and power generation time series," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 5, pp. 3471–3480, 2012.

[49] E. Cadenas and W. Rivera, "Wind speed forecasting in three different regions of mexico, using a hybrid arima–ann model," *Renewable Energy*, vol. 35, no. 12, pp. 2732–2738, 2010.

[50] L. A. Díaz-Robles, J. C. Ortega, J. S. Fu, G. D. Reed, J. C. Chow, J. G. Watson, and J. A. Moncada-Herrera, "A hybrid arima and artificial neural networks model to forecast particulate matter in urban areas: The case of temuco, chile," *Atmospheric Environment*, vol. 42, no. 35, pp. 8331–8340, 2008.

[51] B. Zhu and Y. Wei, "Carbon price forecasting with a novel hybrid arima and least squares support vector machines methodology," *Omega*, vol. 41, no. 3, pp. 517–524, 2013.

[52] S. Barak and S. S. Sadegh, "Forecasting energy consumption using ensemble arima–anfis hybrid algorithm," *International Journal of Electrical Power & Energy Systems*, vol. 82, pp. 92–104, 2016.

[53] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[54] H. Zou, G. Xia, F. Yang, and H. Wang, "An investigation and comparison of artificial neural network and time series models for chinese food grain price forecasting," *Neurocomputing*, vol. 70, no. 16, pp. 2913–2923, 2007.

[55] H. C. Co and R. Boosarawongse, "Forecasting thailands rice export: Statistical techniques vs. artificial neural networks," *Computers & industrial engineering*, vol. 53, no. 4, pp. 610–627, 2007.

[56] V. R. Prybutok, J. Yi, and D. Mitchell, "Comparison of neural network models with arima and regression models for prediction of houston's daily maximum ozone concentrations," *European Journal of Operational Research*, vol. 122, no. 1, pp. 31–40, 2000.

[57] N. Kohzadi, M. S. Boyd, B. Kermanshahi, and I. Kaastra, "A comparison of artificial neural network and time series models for forecasting commodity prices," *Neurocomputing*, vol. 10, no. 2, pp. 169–181, 1996.

[58] S. Ho, M. Xie, and T. Goh, "A comparative study of neural network and box-jenkins arima modeling in time series prediction," *Computers & Industrial Engineering*, vol. 42, no. 2, pp. 371–375, 2002.

[59] I. Alon, M. Qi, and R. J. Sadowski, "Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods," *Journal of Retailing and Consumer Services*, vol. 8, no. 3, pp. 147–156, 2001.

[60] K. Benmouiza and A. Cheknane, "Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models," *Energy Conversion and Management*, vol. 75, pp. 561–569, 2013.

[61] J. Ruiz-Aguilar, I. Turias, and M. Jiménez-Come, "A novel three-step procedure to forecast the inspection volume," *Transportation Research Part C: Emerging Technologies*, vol. 56, pp. 393–414, 2015.

[62] M. R. Amin-Naseri and E. A. Gharacheh, "A hybrid artificial intelligence approach to monthly forecasting of crude oil price time series," in *The Proceedings of the 10th International Conference on Engineering Applications of Neural Networks, CEUR-WS284*, pp. 160–167, 2007.

[63] R. L. Winkler and S. Makridakis, "The combination of forecasts," *Journal of the Royal Statistical Society. Series A (General)*, pp. 150–157, 1983.

[64] K. W. Hipel and A. I. McLeod, *Time series modelling of water resources and environmental systems*, vol. 45. Elsevier, 1994.

[65] T. Lin and M. Pourahmadi, "Nonparametric and non-linear models and data mining in time series: a case-study on the canadian lynx data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 47, no. 2, pp. 187–201, 1998.

[66] M. Campbell and A. Walker, "A survey of statistical work on the mackenzie river series of annual canadian lynx trappings for the years 1821-1934 and a new analysis," *Journal of the Royal Statistical Society. Series A (general)*, pp. 411–431, 1977.

[67] C. S. Wong and W. K. Li, "On a mixture autoregressive model," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, no. 1, pp. 95–115, 2000.

[68] R. A. Meese and K. Rogoff, "Empirical exchange rate models of the seventies: Do they fit out of sample?," *Journal of international economics*, vol. 14, no. 1-2, pp. 3–24, 1983.

[69] A. J. Conejo, M. A. Plazas, R. Espinola, and A. B. Molina, "Day-ahead electricity price forecasting using the wavelet transform and arima models," *IEEE transactions on power systems*, vol. 20, no. 2, pp. 1035–1042, 2005.

[70] Z. Tan, J. Zhang, J. Wang, and J. Xu, "Day-ahead electricity price forecasting using wavelet transform combined with arima and garch models," *Applied Energy*, vol. 87, no. 11, pp. 3606–3610, 2010.

[71] A. K. Fard and M.-R. Akbari-Zadeh, "A hybrid method based on wavelet, ann and arima model for short-term load forecasting," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 26, no. 2, pp. 167–182, 2014.

[72] H.-c. Zhou, Y. Peng, and G.-h. Liang, "The research of monthly discharge predictor-corrector model based on wavelet decomposition," *Water resources management*, vol. 22, no. 2, pp. 217–227, 2008.

[73] S. Wei, D. Zuo, and J. Song, "Improving prediction accuracy of river discharge time series using a wavelet-nar artificial neural network," *Journal of Hydroinformatics*, vol. 14, no. 4, pp. 974–991, 2012.

[74] J. Adamowski and K. Sun, "Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds," *Journal of Hydrology*, vol. 390, no. 1, pp. 85–91, 2010.

[75] J. Adamowski and H. F. Chan, "A wavelet neural network conjunction model for groundwater level forecasting," *Journal of Hydrology*, vol. 407, no. 1, pp. 28–40, 2011.

[76] V. Nourani, M. Komasi, and A. Mano, "A multivariate ann-wavelet approach for rainfall–runoff modeling," *Water resources management*, vol. 23, no. 14, pp. 2877–2894, 2009.

[77] T. Partal and O. Kisi, "Wavelet and neuro-fuzzy conjunction model for precipitation forecasting," *Journal of Hydrology*, vol. 342, no. 1, pp. 199–212, 2007.

[78] M. K. Tiwari and C. Chatterjee, "Development of an accurate and reliable hourly flood forecasting model using wavelet–bootstrap–ann (wbann) hybrid approach," *Journal of Hydrology*, vol. 394, no. 3, pp. 458–470, 2010.

[79] I. Khandelwal, R. Adhikari, and G. Verma, "Time series forecasting using hybrid arima and ann models based on dwt decomposition," *Procedia Computer Science*, vol. 48, pp. 173–179, 2015.

[80] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on pure and applied mathematics*, vol. 41, no. 7, pp. 909–996, 1988.

Publications

International journal

[1] W. Pannakkong, V. H., Pham, and V-N, Huynh, "A Novel Hybridization of ARIMA, ANNs and k-means for Time Series Forecasting", *International Journal of Knowledge and Systems Science*, IGI Global, 8(4), 2017, in press.

[2] W. Pannakkong, S. Sriboonchitta, and V-N. Huynh "A Novel Hybrid ARIMA and Artificial Neural Network Model for Cassava Export Forecasting", *Journal of Systems Science and Systems Engineering*, under review, 22 pages

Book chapter

[3] W. Pannakkong, V-N. Huynh, and S. Sriboonchitta, ARIMA Versus Artificial Neural Network for Thailands Cassava Starch Export Forecasting, in *Causal Inference in Econometrics*, Springer-Verlag, Chapter 16, 255-277, 2016

International conference

[4] W. Pannakkong, V. H. Pham, and V-N Huynh, A Hybrid Model of ARIMA, ANNs and k-Means Clustering for Time Series Forecasting, *Integrated Uncertainty in Knowledge Modelling and Decision Making* (IUKM 2016), LNCS **9978**, Springer-Verlag, 195-206

[5] W. Pannakkong and V-N. Huynh, A Hybrid Model of ARIMA and ANN with Discrete Wavelet Transform for Time Series Forecasting, *The 14th International Con-*

ference on Modeling Decisions for Artificial Intelligence (MDAI 2017), accepted, 11 pages