KKU ENGINEERING JOURNAL 2016;43(S2):203-206 Research Article

KKU Engineering Journal

https://www.tci-thaijo.org/index.php/kkuenj/index

Hybrid forecast models for PM-10 prediction: A case study of Chiang Mai city of Thailand during high season

Rati Wongsathan*, Issaravuth Seedadan and Supawat Wanasri

Department of Electrical Engineering, Faculty of Engineering, North-Chiang Mai University, Chiang Mai 50230, Thailand.

Received April 2016
Accepted June 2016

Abstract

In this study, the forecast models including to ARIMA, NNs, hybrid of ARIMA-NNs and hybrid of NNs-ARIMA model are employed to predict the PM-10 level during the high season in Chiang Mai, in the province of northern Thailand. The k-folds cross validation technique is used in the experimental design in order to prevent the over-fitting which is generally existed with strong impact on the forecast model. The historical PM-10 data are taken as the input of the forecast model. The statistics test and parameter designed experiments are used to optimize the linear models while the back-propagation (BP) learning algorithm is used to train the neural networks. The average of root mean squared error (RMSE) and mean absolute error (MAE) are used to indicate their performances. The results indicate that the hybrid NNs-ARIMA model is highly able to predict the PM-10 over the rest.

Keywords: PM-10, ARIMA, Neural networks, Hybrid ARIMA-NNs, Hybrid NNs-ARIMA

1. Introduction

People in Chiang Mai city in northern Thailand has annually facing and suffering from severe pollution related to particulate matter up to 10 micrometer or PM-10 for a decade. PM-10 starts climbing to the dangerous level especially between January to April, dry-season aridity and rising temperatures which resulted from forest fire, wood and agricultural burning. In 2014, the PM-10 level exceeded the threshold level (120 µg/m^3) that literally thousands of people were admitted to the hospital with various respiratory illness. Therefore, the PM-10 forecasting will necessarily prevent the illness of the people by preparing themselves in advance. Traditionally, most of the forecast model frequently uses the historical values of PM-10 to estimate the current PM-10 value e.g. in ARIMA model [1-2] or in Neural Networks (NNs) [3-4]. In our previous work [3], we had already predicted the PM-10 in the Chiang Mai city in moat area in all seasons by using various NNs model. The accuracy resulted well, however the complexity of the model was the main problem. In our present research [5], the experimental results demonstrated that the hybrid ARIMA-NNs model outperformed best over NNs and ARIMA respectively for all season forecasting. However, the number of the hybrid model parameter leads to the overfiting. Furthermore, in our previous work [6] the priority processing between linear and nonlinear in hybrid model was the significant issue which will be considered.

2. Materials and methods

In this work, the PM-10 data were collected from Jan-April between 2011-2015. The data set from 2011-2014 was used for training and validation followed by 4-folds validation technique while the data set in year 2015 was used to test the performance of the model.

2.1 The ARIMA model

The ARIMA model typically consists of three parts i.e. auto regression AR (order p), moving average MA (order q) and differencing in order to strip off the integration of the series (order d) and then form with ARIMA (p, d, q). This linear model is mathematically expressed through the following,

$$\Delta^d Y_t = \delta + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \quad (1)$$

where $\Delta(Y_t)=Y_t-Y_{t-1}$, Y_t is the observation data and ε_t is the error at time t, δ is the constant, φ_i is the autoregressive parameter and $\sim N(0, \sigma^2)$, and θ_i is the moving average parameters.

A practical approach to building ARIMA model includes three iterative steps as following. In identification step, non-stationary PM-10 data uses preliminary for investigation by ACF (auto correlation function) and PACF (partial ACF), and the unit root is tested by augmented Dickey-Fuller (ADF) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. The first differencing data transformation (d =1) is used to make PM-10 stationary. In PM-10 data, the ACF and

*Corresponding author. Tel.: +6681 289 3400
Email address: rati1003@gmail.com; rati@northcm.ac.th
doi: 10.14456/kkuenj.2016.89

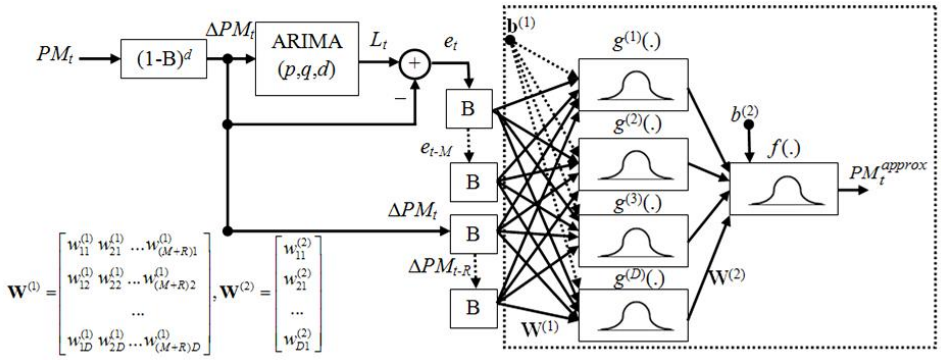

Figure 1 The structure of hybrid ARIMA(*p*,*d*,*q*)-NNs([*M*,*R*],*D*) model

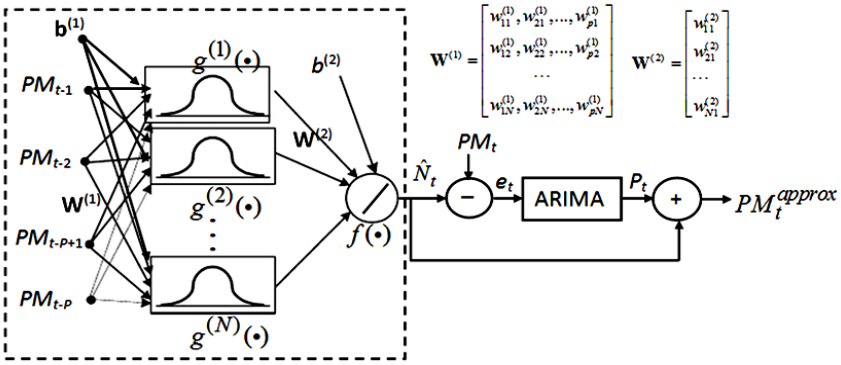

Figure 2 The structure of hybrid NNs(*P*,*N*)-ARIMA(*p*,*d*,*q*) model

PACF of Δ*PM* identified the order *q* and *p* to 3 and 4 respectively, which yields ARIMA (4, 1, 3) model. The set of ϕ and θ parameters are estimated in the estimation step. In the last step, diagnostic checking of the residuals by Box-Pierce Chi-Square test verified that ARIMA (4, 1, 3) is sufficient since it reveals no correlation of the residuals.

2.2 A Neural networks model (NNs)

In this work, a multi-layer perceptron (MLP) was selected to use as the forecast model and is shown in dash line box of Figure 1 The NNs output is referred as predicting PM-10 at current time *t* can be expressed in matrix-vector form as

$$PM_t^{approx} = f\left(\mathbf{W}^{(2)} \times g\left(\mathbf{W}^{(1)} \times \mathbf{PM} + \mathbf{b}^{(1)}\right) + b^{(2)}\right) \quad (2)$$

where W(1) and b(1) are weight matrix and bias vector between input and hidden layer respectively, $\mathbf{W}^{(2)}$ and $\mathrm{b}^{(2)}$ are the weight vector and bias value between hidden and output layer respectively, and **PM** is the input matrix of the historical PM-10. In the test, the hyperbolic tangent was selected as an activation function of *g* and *f* for the linear transfer function. The parameter of NNs i.e. weights and biases are searched by the well-known BP algorithm.

2.3 The hybrid ARIMA-NNs model

Integrating the residuals from ARIMA model and the historical PM-10 data can be modeled by the NNs to discover nonlinear pattern which is composed of hybrid ARIMA-NNs in Figure 1 The PM-10 series data is assumed to be a composition between a linear autocorrelation (L_t) structure and nonlinear component (N_t) as $Y_t = L_t + N_t + e_t$, where and e_t is the residuals at time *t*. This hybrid model can be expressed as

$$PM_t^{approx} = h[(\Delta PM_{t-1}, \Delta PM_{t-2}, ..., \Delta PM_{t-R}), (e_{t-1}, e_{t-2}, ..., e_{t-M})], \quad (3)$$

where ΔPM_t = (1-B)(PM_t) and *h* is a nonlinear function determined by NNs.

2.4 The Hybrid of NN-ARIMA model

The diagram of the hybrid NNs-ARIMA model is shown in Figure 2, the NNs design in section 2.2 were applied to forecast the solution N_t in the first stage. The residuals are treated as the linear model (L_t) in the second stage which is statistically investigated and modeled by an ARIMA model. From the Figure 2, the forecast value at time *t* for *P* input node, *N* hidden node and one output node of MLPNNs can be expressed as

$$\hat{N}_t = w_0 + \sum_{j=1}^{N} w_j \cdot g\left(w_{0,j} + \sum_{i=1}^{P} w_{i,j} \cdot y_{t-i}\right). \quad (4)$$

The residual at time *t* from the nonlinear model and followed by, $e_t = PM_t^{Actual} - \hat{N}_t$, where $\hat{N}_t$ refers to the forecast value for time *t*. These residuals are modeled as $\Phi(\mathrm{B})\Delta^d e_t = \delta + \Theta(\mathrm{B})\varepsilon_t$, by the ARIMA model, where linear relationships can be discovered.

3. Forecast model setting results and discussion

For ARIMA (4, 1, 3) model, the suitable parameters are selected and proved as

Table 1 Comparison of RMSE and MAE between four forecast models

Model	ARIMA(4,1,3)		NNs(1,1)		hARIMA(4,1,3)NNs([3,1],5)		hNNs(1,1)ARIMA(1,1,00)	
	Train-valid	Test	Train-valid	Test	Train-valid	Test	Train-valid	Test
RMSE	18.75	25.12	20.73	37.55	12.10	41.73	1.59	1.67
MAE	14.35	17.07	18.23	16.28	9.83	31.73	1.77	1.873
Number of parameter	7		4		7+31		4+1	

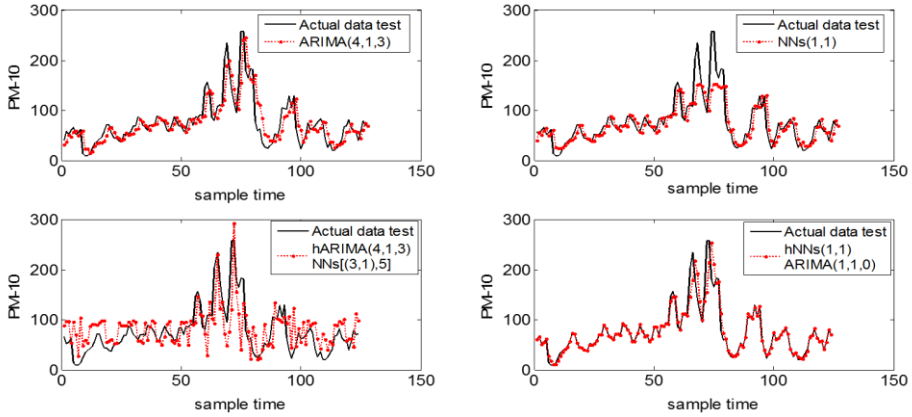

Figure 3 The prediction results of PM-10 data test set for all four forecast models

$$\Delta PM_t = (1.161)\Delta PM_{t-1} + (-1.006)\Delta PM_{t-2} + (0.345)\Delta PM_{t-3} + (-0.224)\Delta PM_{t-4} + (-1.389)\varepsilon_{t-1} + (1.066)\varepsilon_{t-2} + (-0.311)\varepsilon_{t-3} \quad (5)$$

In NNs model, the optimized hidden node corresponds with 1 input node is 1. Then the NNs model can be represented by NNs (1, 1, 1) which first, second and third of number 1 refers to number of input, hidden and output nodes respectively. The NNs (1, 1, 1) explicitly expressed as

$$PM_t = 0.5553\left(\frac{2}{1+\exp(-2\times(2.1717\times PM_{t-1}+0.8987))}-1\right)-0.4035 \quad (6)$$

In hybrid ARIMA-NNs, the residuals were generated from the ARIMA(4,1,3) which are used as the partial input altogether with ΔPM_t. The input node number (M and R) are determined by the experimental as well as the number of hidden node (D) of NNs which is denoted by hARIMA(4,1,3)NNs([M,R],D). From the test, the optimized value of M, R and D is found and represented by hARIMA(4,1,3)NNs([3,1],5) which can be expressed as,

$$PM_t^{approx} = f\left(b^{(2)} + \sum_{j=1}^{5} w_{j1}^{(2)} \cdot g\left(\sum_{i=2}^{4} (w_{ij}^{(1)}(\Delta PM - ARIMA(4,1,3))_i + b_i^{(1)}) + w_{1j}^{(1)}(\Delta PM)_{t-i} + b_i^{(1)}\right)\right) + \varepsilon_t, \quad (7)$$

For hybrid NNs-ARIMA model, the designed NNs(1,1) in section 2.2 were adopted to forecast solution N_t then the residuals will be fed into ARIMA model. In the diagnostic checking with the criteria mentioned in section 2.1, it is found that the residue from NNs(1,1) is modeled to ARIMA(1,1,0) and the hybrid NNs(1,1)-ARIMA(1,1,0) is expressed as

$$PM_t^{Approx} = \left[0.5553\left(\frac{2}{1+\exp(-2\times(2.1717\times PM_{t-1}^{Actual}+0.8987))}-1\right)-0.4035\right] + residue_{t-1} - 0.4558(residue_{t-1} - residue_{t-2}) \quad (8)$$

The four designed forecast models have finally tested with the data test set of PM-10 in year 2015. The performance comparison with RMSE and MAE are shown in Table 1 and illustrated in Figure 3.

In hybrid ARIMA-NNs, the residuals were generated from the ARIMA(4,1,3) which are used as the partial input altogether with ΔPM_t. The input node number (M and R) are determined by the experimental as well as the number of hidden node (D) of NNs which is denoted by

The valid and test model with a smaller average RMSE and MAE are the proposed hybrid NNs-ARIMA model following by ARIMA model, NNs model, and hybrid ARIMA-NNs model respectively. The hybrid NNs-ARIMA model, can clearly give more accuracy than the single linear or nonlinear model. While the hybrid ARIMA-NNs with many parameters were performed well for the training process but the residue is now randomly sequence which may use high complex NNs structure to capture most of the data that leads to the overfitting. Since the main pattern of PM-10 problem is nonlinear then the first priority processing should be filtered by NNs model before it propagates to the last step which may occur more complex and managed the remaining by ARIMA model. In this case, the hybrid NNs-ARIMA model is considered the best forecast model. In general case, there is no any theoretical guarantee which hybrid model is better but it depends on the nature of the problem.

4. Conclusions

PM-10 forecast models including ARIMA, NNs, hybrid ARIMA-NNs and hybrid NNs-ARIMA models are proposed to predict the PM-10 level during the high season in the case of Chiang Mai province. The results were clearly indicated that the hybrid NNs-ARIMA model performs better than the rest. The hybrid NNs-ARIMA model with the optimal structure has the capability to learn the non-linear pattern which resulted to highly linear error output in this case study. The supplement forecast by ARIMA then can keep more

accurate. Unlike the hybrid ARIMA-NNs model, an ARIMA cannot trace the non-linear pattern well which may produce the uncertain form in the second stage forecast by the more complex NNs which make the errorneous prediction. The reliable forecast model depends only in the accuracy result but it is suitable model structure.

5. References

[1] Goyal P, Chan AT, Jaiswal N. Statistical models for the prediction of respirable suspended particulate matter in urban cities. Atmospheric Environment 2006;40:2068-2077.

[2] Liu PWG. Simulation of the daily average PM10 concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis. Journal of the American Statistical Association 2009;92(439):2104-2113.

[3] Wongsathan R, Seedadan I. Prediction modeling of PM-10 in Chiangmai city moat by using artificial networks. Journal of Applied Mechanics and Materials 2015;781:628-631.

[4] Díaz-Roblesa LA, Ortegaa JC, Fub JS, Reedb GD, Chowc JC, Watsonc JG, Moncada-Herreraa JA. A hybrid arima and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. Atmospheric Environment 2008; 42(35):8331-8340.

[5] Wonsathan R, Seedadan, I. A hybrid ARIMA and neural networks model for PM-10 pollution estimation: The case of Chiang Mai city moat area. Proceeding of the 2016 International Electrical Engineering Congress; 2016 March 2-4; Chiang Mai, Thailand. Amsterdam: Elsevier; 2016.

[6] Wongsthan R, Chankham S. Improvement on PM-10 forecast by using ARIMAX and hybrid ARIMAX and neural networks model for the summer season in Chiang Mai. Proceeding of the 2016 International Electrical Engineering Congress; 2016 March 2-4; Chiang Mai, Thailand. Amsterdam: Elsevier; 2016.