Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 86 (2016) 277 – 280

Procedia Computer Science

2016 International Electrical Engineering Congress, iEECON2016, 2-4 March 2016, Chiang Mai, Thailand

# Improvement on PM-10 forecast by using hybrid ARIMAX and Neural Networks Model for the summer season in Chiang Mai

Rati Wongsathan[a]*, Supawat Chankham[b]

[a] *Electrical Engineering Department, North-Chiangmai University HangDong Chiang Mai Thailand 50230*
[b] *Electrical Engineering Department, North-Chiangmai University HangDong Chiang Mai Thailand 50230*

**Abstract**

Since the air monitoring stations do not provide the relation between other toxic gas and meteorological parameters with the particulate matter up to 10 micrometer or PM-10.The influence of meteorological as well as correlation with other toxic gas is investigated and used them to forecast PM-10 in the case of Chiang Mai province of Thailand. In this paper an attempt to develop hybrid models of an Autoregressive Integrated Moving Average (ARIMA) model with other exogenous variables (ARIMAX) and Neural Networks (NNs), the two hybrid models, i.e. hybrid ARIMAX-NNs model and hybrid NNs-ARIMAX model were implemented to forecast PM-10 for highly season during January-April of Chiang Mai Province. Simulation results of hybrid model are compared with the results of ARIMA, ARIMAX and NNs model. The experimental results demonstrated that the hybrid NNs-ARIMAX model outperformed best over the hybrid ARIMAX-NNs model, ARIMAX model, NNs model, and ARIMA model respectively. In this case study and maybe other cases, it has proved that the NNs model should be priori captured and filtered the non-stationary non-linear component while the fully linearly stationary residuals were accurately predicted by ARIMAX model later.


Peer-review under responsibility of the Organizing Committee of iEECON2016

*Keywords:* PM-10; Hybrid ARIMAX and Neural Networks.

## 1. Introduction

Chiang Mai, the largest city in the north of Thailand, severely experienced with the pollution related to PM-10 for a decade. PM-10 start climbing to the dangerous level especially between January to April, dry-season aridity and rising temperatures coincide with forest fire, wood and agricultural burning. By comparison, PM-10 in Bankok

* Corresponding author. Tel.: +66(81)2893400; fax: +66(53)819998.
*E-mail address:*rati1003@gmail.com


Peer-review under responsibility of the Organizing Committee of iEECON2016
doi:10.1016/j.procs.2016.05.062

has only 40-50 during this same period. PM-10's sources go beyond farming, forest and grass burning, e.g only Mae Chaem one of the district of Chiang Mai alone produces and burns over 37,000 tons of corncob waste every year [1]. For our assumption that PM-10 is nonlinear and complex model, the capture of advantage both ARIMA and NNs model is alternate selected and implemented which gives the forecast result better than any single model. However, the residual usually severe occurs in the high season period. To solve this problem, other exogenous variables which relate to PM-10 are considered to include in an ARIMA model and is referred to an ARIMAX model.

## 2. Methodology and Methods

In this work, special emphasis is focused for high season period with highly disturbance by various factors and implemented the hybrid model on PM-10 forecast. The data are collected from year 2011-2013 for both PM-10 and exogenous data which are 4 toxic gas variables i.e. *CO*, $O_3$, $NO_2$ and $SO_2$ and 4 meteorological variables i.e. gust wind (*GW*), temperature (*T*), pressure (*P*), and relative humidity (*H*) [2]-[3]. The data in year 2011-2012 was used for model training and the data in year 2012-2013 was used to test the performance of the model.

### *2.1. ARIMAX model*

It assumes that input sequences are expressed by $\{X_{1t}\},\ldots,\{X_{Kt}\}$, dependence sequence is represented by $\{Y_t\}$, and ARIMA($p,d,q$)X($K$) model can be described as following,

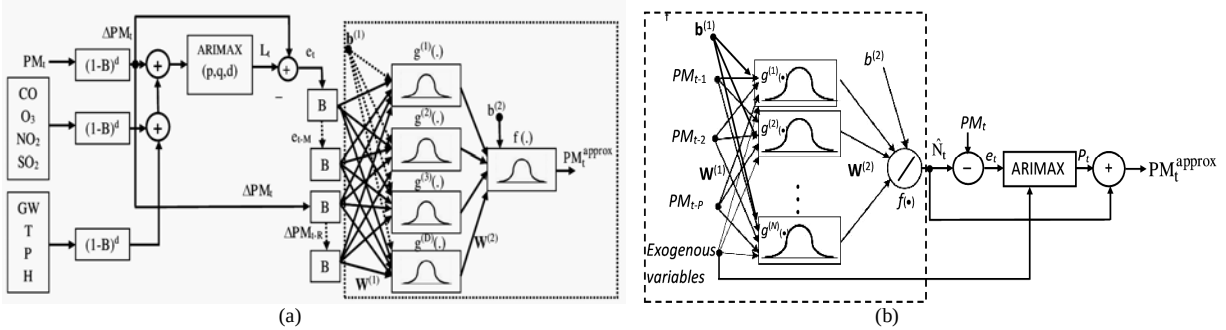
$$\Delta^d Y_t = \delta + \sum_{i=1}^{K} \mu_i X_{t-i} + \sum_{i=1}^{P} \varphi_i Y_{t-i} + \sum_{i=0}^{Q} \theta_i \varepsilon_{t-i} \quad . \tag{1}$$

Where $\mu_i$, $\varphi_i$, and $\theta_i$ denote the coefficient parameters, *K*, *P*, and *Q* denote the maximum time lag related to the input sequences, dependence sequence and residuals respectively, and $\delta$ is a constant.

According to the [4]-[5], the basically procedure step uses to set up ARIMAX model is similar to ARIMA model. Stationary test on both $\{X_{it}\}$ and $\{Y_t\}$ is preliminary examined by ACF and PACF plot or unit root test by augmented Dickey-Fuller (ADF) test. The first differencing is applied to the time series data for non-stationary case. After series are identified as the stationary, the best fit parameters of an ARIMA model were estimated according to its order *p* and *q* by PACF and ACF plot considering. The MLR is fitted the model and the insignificance variable is eliminated by notation of P-value statistics. Diagnostic checking is used to examine at the last step by the several statistics assumption of the residuals such as Chi-Square test or the correlation of the residual plot.

### *2.2 Hybrid ARIMAX-NNs model*

It may be assume to consider PM-10 times series to be composed of a linear autocorrelation structure ($L_t$) and a nonlinear component ($N_t$) as, $Y_t = L_t + N_t + e_t$, where $e_t$ is the residuals at time *t*. In the proposed model, there are mainly two stages which has illustrated in the diagram of Fig. 1(a). In the first stage ARIMAX operates to forecast by using the *K* historical data, *Q* past error values, and the exogenous variables. The residual is then generated and provided to NNs which used this error altogether with historical of $\Delta PM$ data for the final PM-10 forecast. In the design experiment of generalized NNs, several factors including number of input node (in this case is the time lag length) and number hidden node are properly selected for accuracy and rapidly convergence of solution.




(a) (b)

Fig. 1. The struture of (a) hybrid ARIMAX-NNs model, and (b) hybrid of NNs-ARIMAX modle

### *2.3 Hybrid NNs-ARIMAX model*

Model of the hybrid NNs-ARIMAX model is shown in Fig. 1(b), the designed structure NNs with *P* input node, *N* hidden node and one output node of NNs model was adopted to forecast solution $N_t$ in the first stage. The

residuals at time $t$ will be treated as the linear model ($L_t$) in the second stage which is statistically investigated and modeled by an ARIMAX model. Once, the linear correlation will be removed from the residuals for sufficient condition of ARIMAX model. Combining of $L_t$ and $N_t$, then the hybrid NNs-ARIMAX model will be hold.

## 3. Experimental Results and Discussion

From ACF analysis, all variables are non-stationary since the ACF slowly died off. The differencing is then applied to all data series which results stationary for all variables. The order $p$ and $q$ determined from the time lag of ACF and PACF plot are both 9 which is not suitable. By MLR analysis, $\Delta PM_t$ at current time $t$ is the function of 9 historical of $\Delta PM$, 9 lag error ($\varepsilon$) and 8 exogenous variables. By the test, $\Delta PM_{t\text{-}4}$-$\Delta PM_{t\text{-}9}$, $\varepsilon_t$-4-$\varepsilon_t$-9, $\Delta P$, $\Delta T$, and $\Delta H$ have the standard error is more twice than the coefficient of parameters value and all P-values is more than 0.01 which were got rid off from the model. The residual from ARIMA(3,1,3)X(5) model will be tested the correlation by using Ljung-Box test. The Q-stat (12.75) with maximum time lag 25 at $df$=25-12 and 95% confident interval is less than the critical value (22.36), then the residual has no correlate. ARIMAX(3,1,3)X(5) is expressed by,

$$\begin{aligned}\Delta PM_t &= 2.379\Delta PM_{t-1} - 2.129\Delta PM_{t-2} + 0.712\Delta PM_{t-3} - 2.673\varepsilon_{t-1} + 2.594\varepsilon_{t-2} - 0.915\varepsilon_{t-3} \\ &\quad + 21.863(\Delta CO)_t + 1.027(\Delta O_3)_t + 3.626(\Delta NO_2)_t + 10.164(\Delta SO_2)_t + 0.466(\Delta GW)_t\end{aligned} \quad (2)$$

The forecast performance of the model also compared with the ARIMA(4,1,4) which was designed in the same manner is shown in Fig. 1. The ARIMAX model clearly performs better than ARIMA model but the error need to adjust with NNs and the hybrid model later.

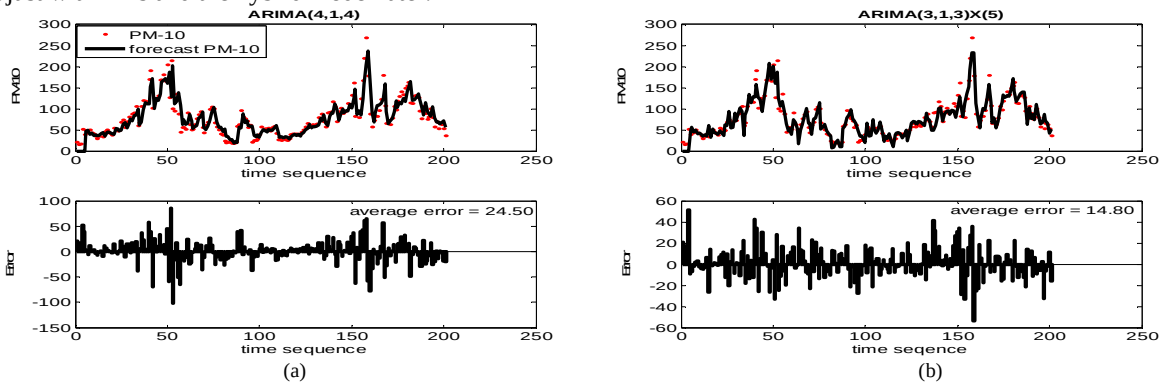



Fig. 2. The forecast results during Jan.-Apr. in 2012-2013 of (a) ARIMA(4,1,4), and (b) ARIMA(3,1,3)X(5).

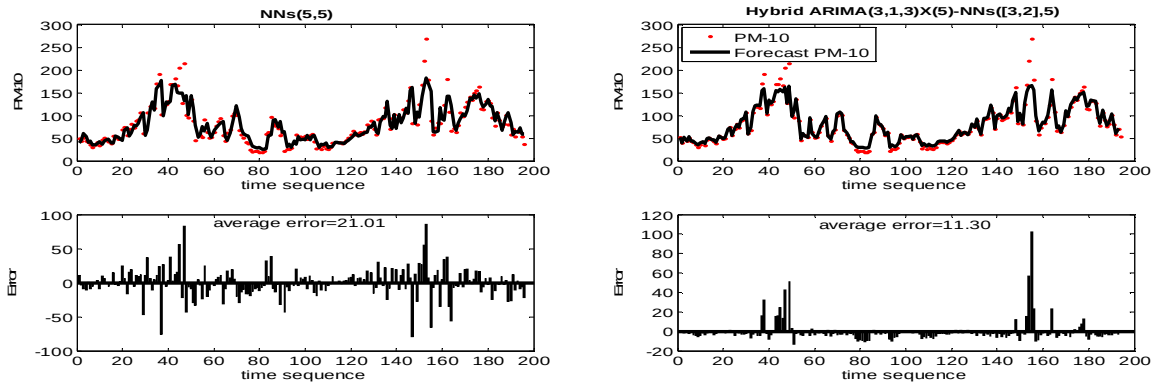
The number of input and hidden node of NNs represented by MLP are designed while minimized the MSE. Result from ARIMAX design, NNs(5,5) and hARIMA(3,1,3)X(5)-NNs([2,2],5) are generated. The forecast of both models are illustrated in Fig.2(a) and (b) respectively, the hybrid ARIMAX-NNs model performed the forecast better than ARIMAX, NNs and ARIMA respectively. The mathematics expression is explicit made in (3),

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

The forecast result from hARIMAX-NNs model is still not good, the other hybrid model is observed with operation in the inverse direction. The designed NNs(5,5) was adopted to forecast solution $N_t$ in the first step. The residuals will be fed into ARIMAX model in the second step. The first difference made the residual resulted stationary. The ACF and PACF determined the order of $p$ equal to 2 and $q$ equal to 3. The MLR analysis eliminated all the exogenous variables except for $O_3$ variable which has P-value less than 0.01. The residual of the ARIMA(2,1,3)X(1) will be tested the serial correlation by using Ljung-Box test. The Q-stat (10.02) with maximum time lag 25 at $df$ = 25-6 and 95% confident interval is less than the critical value (10.12), then the residuals has no correlate. ARIMAX(3,1,3)X(5) is then finally selected and expressed by (4) with $N_t$ in (5),

$$\Delta(e_t) = 1.144\Delta(PM-N)_{t-1} - 0.978\Delta(PM-N)_{t-2} + 2.114\varepsilon_{t-1} + 2.10\varepsilon_{t-2} - 0.976\varepsilon_{t-3} - 0.16(\Delta O_3)_t, \quad (4)$$

$$N_t = f\left(b^{(2)} + \sum_{j=1}^{5} w_{j1}^{(2)} \cdot g\left(\sum_{i=1}^{5}(w_{ij}^{(1)}(\Delta PM)_{t-i} + b_i^{(1)})\right)\right). \quad (5)$$




(a) (b)

Fig. 3. The forecast results of Jan.-Apr. in 2012-2013 of (a) NNs(5,5), and (b) hARIMA(3,1,3)X(5)-NNs([2,2],5).

The forecast result by hNNs(5,5)-ARIMA(2,1,2)X(1) is illustrated in Fig. 4. The forecast performance quite good while the average error is less than hARIMA(3,1,3)X(5)-NNs([2,2],5), ARIMA(3,1,3)X(5), NNs(5,5), ARIMA(4,1,4) by 52%, 63%, 74%, and 77% respectively.

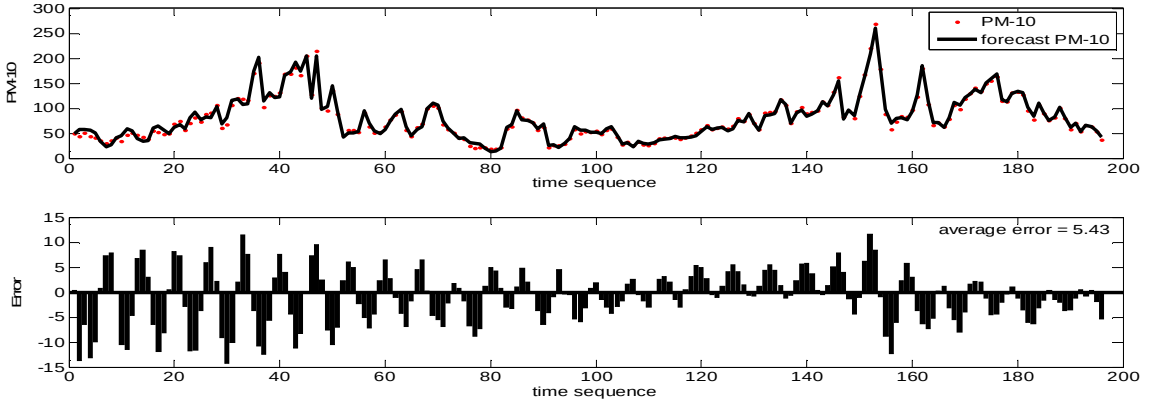



Fig. 4. The fore cast result by the hybrid NNs(5,5)-ARIMA(2,1,3)X(1) during Jan.-Apr. in 2012-2013.

## 4. Summary

The forecast on highly and variance of PM-10 for summer season in the case of Chiang Mai city has already done in the computer simulation by the hybrid of ARIMAX and NNs model with using the related exogenous variable from 4 toxic gas and 4 meteorological variables. A hybrid linear and nonlinear forecast model i.e. hybrid ARIMAX-NNs and hybrid NNs-ARIMA can clearly give more accuracy than a single linear or nonlinear model i.e. ARIMAX, NNs, ARIMA. However the priority processing between linear and nonlinear is the significant issue which will be considered. The first priority processing should be filtered the nonlinear before it propagates to the last step which can occurs more complex and managed the remaining by other linear model. In general case, there is no any theoretical guarantee which hybrid model is better but depends on the nature of the problem.

## References

[1] Information online available at http://asiafoundation.org/in-asia/2014/03/26/transboundary-pollution-in-northern-thailand-causes-dangerous-levels-of-smog/

[2] A. Russo, P.G. Lind, F. Raischel, R. Trigo and M. Mendes, “Daily pollution forecast using optimal meteorological data at synoptic and local scales,” Atmos. and Oceanic Phys., (2014)11.

[3] S. Hormann, B. Pfeiler, and E. Stadlober, “Analysis and prediction of particulate matter PM10 for the winter season in Graz,” Austrian J. of Statistics, 34(2005)307-326.

[4] C.S. Bos, P.H. Franses, and M.Ooms, “Inflation, forecast intervals and long memory regression models,” J. of Forecasting, 18(2002)11-12.

[5] Y. Wand, and Y. Xu,“The application of ARIMAX model,” Statistics and Decision,(2007)9:1-1.