Accuracy enhancements of the ARX model by introducing LSW theory in ozone peak prediction

J. S. Park\textsuperscript{1}, J.-Y. Lee\textsuperscript{1}, H. Lee\textsuperscript{2} & J.-J. Lee\textsuperscript{3}

\textsuperscript{1}Department of Environmental Engineering, University of Seoul, South Korea
\textsuperscript{2}Department of Civil and Environmental Engineering, University of Incheon
\textsuperscript{3}Department of Environmental Health and Science, Yongin University, South Korea

Abstract

The ARX (Auto Regressive eXternal) model is one of the commonest modelling approaches for ozone prediction. In this model, the history of the atmospheric ozone level determines the ozone concentration in the following sequence. The ARX model, however, is known to be improper for predicting the peak levels of ozone variation, as the long-term history is normally applied. Even then there exists a certain amount of time lag in predicted concentrations. In this study, the ARX model was modified and examined by introducing LSW (Local Stationary Wavelet) theory to enhance the accuracy of the conventional ARX model in ozone prediction. The recent updates will be reported in this paper.

Keywords: ARX, LSW, time series, ozone peak prediction.

1 Introduction

Photochemical smog, as is well-known, is a complicate phenomenon with the existence of reactive organic compounds (ROCs), nitrous oxides (NOx), and the ultra-violet waves in sunlight. Especially when meteorological conditions in an urban area are stable, the photochemical smog is formed. It in turn generates a significant amount of ozone which has the potential to cause health risks to the public. There are also many other forms of facts concerning ozone formation, which implies a strong requirement for tools for ozone level prediction.
Currently there are various modelling tools available. A modelling technique is to build a reaction model by considering the chemical reaction mechanisms. Solomon et al. (2000) conducted a study using the monitored data from the U.S., Canada, Mexico, etc [13]. This approach, however, requires a large amount of cost and time to obtain various information including ozone mechanisms (generation, accumulation). The concentration and emission data of precursors, NOx, and VOCs in the field are typically necessary for validation purposes. On the other hand, prediction techniques employing statistical approaches such as the empirical black box model, grey model, etc., are frequently introduced owing to less limitations compared to the modelling approach previously described. This statistical approach, in brief, is to build prediction models by analyzing the statistical characteristics of atmospheric conditions and precursors in the past and present.

In this study, the long-term linear trend in ozone concentrations using the wavelet technique was decomposed. The analyzed ozone concentration with a linear shape is then applied to the auto-regressive external (ANX) model. The non-linear components are used by a local stationary wavelet (LSW) model. The disadvantages with existing ARX models in predicting non-linear variations are compensated by combining the LSW model. Finally, a new modelling approach is proposed to predict the peaks of ozone variation with reasonable accuracy.

2 Ozone prediction model

AR (Autoregressive time series) models can describe the persistence and even periodic behaviour owing to their constant characteristics; however we expect them to be of limited use for describing the result of the complex processes of photochemical formation and atmospheric transport of ozone. In order to ease this problem, we use the ARX model, a kind of AR model, to describe seasonal trends. LSW is used to forecast random work processes [10]. Both ARX and LSW models are conjoined to take the prediction advantages such as ARX trends and LSW of random work processes. The modified model will be tested to forecast ozone peak efficiently in this study.

2.1 Mathematical review of modified model (ARX + LSW)

A time series of ozone data is composed of long term components and short term components. We initially decompose components using a wavelet decomposition method in this study. We used the Coifet [5] function as a basic function for wavelet decomposition, suggested by Daubechies [5]. This function has 2N moments and 2N-1 scale function and is effective for rapid fluctuation series, for example wind speed, stock value etc. [8].

In this study, we try to decompose long and short term components using a wavelet filter. Decomposed components are predicted using ARX and LSW separately. Then the decomposed components are convolved at the merging step. Finally the modified model is obtained as shown in Fig. 1.
This procedure of Eqs. (1)–(5) represents the procedure that the raw ozone data is composed of long/short and white noise components

\[ O_3(t) \rightarrow \text{wavelet} \rightarrow e(t) + w(t) + n(t) \]  

where \( O_3 \) is the ozone time series, \( e(t) \) the long term component, \( n(t) \) the white noise. It is assumed that the long term components have a periodic cycle over 24 hour and the short components have shorter cycle than 24 hour with all decomposed data series. Long components are then modelled using ARX and others using the LSW model. This can be written as Eq. (2),

\[ O_3(t) = ARX(t) + LSW(t) + \varepsilon(t) \]  

We can express the model through Eqs. (3)–(5)

\[ e(t) \rightarrow ARX \rightarrow \sum_{j=1}^{p} \alpha_j X_{t-j} + \varepsilon; \approx \hat{e}(t) \]  

\[ w(t) \rightarrow LSW \rightarrow \sum_{j=-1}^{\infty} \sum_{k \in Z} S_j \left(\frac{k}{T}\right) X_{t-j} \psi_{j,k-(t-1+h)} \psi_{j,k-m} + \varepsilon; \approx \hat{w}(t) \]  

\[ \hat{e}(t) + \hat{w}(t) \rightarrow \text{convolution} \rightarrow \sum_{j=1}^{p} \alpha_j X_{t-j} + \sum_{j=-1}^{\infty} \sum_{k \in Z} S_j \left(\frac{k}{T}\right) X_{t-j} \psi_{j,k-(t-1+h)} \psi_{j,k-m} + E(t) = \hat{O}_3(t) + E(t) \]
When each model structure was decided properly, the developed model produces a minimized prediction error \((E(t))\).

2.2 Decomposition process

We use the wavelet decomposition to extract the trend components in ozone data, and use the extracted data as the input variables or the basic variables for the model. For example, extracted data will be analyzed for each correlation coefficient. In this study, we use the Coeflet [5] the function as the basic wavelet function suggested by Daubechies [5]. The wavelet is a modified FFT (Fast Fourier Transformation) where, \(\psi(t)\) is the wavelet basic function, \(x(t)\) the ozone time series, \(c\) is the transformed wavelet coefficient, which indicates the scale and time in FFT for the wavelet function. We extracted long term trends and short fluctuations through the inverse transformation of wavelet coefficients.

2.3 Data set

The data set used in the study was produced by the air pollution monitoring system at Jeongdong station [TMX: 197.868, TMY: 501.606] from Ministry of Environment and Meteorological data of Songweoldong station from Korea Meteorological data [KMA]. All station is located near the city hall of Seoul in Korea. The measuring station is influenced by the local urban traffic flow especially during rush hour. Nitrogen Dioxide (NO\(_2\)) especially is a sensitive air pollutant. The fundamental air pollution data is the hourly averaged concentration produced by the measured data for a 5 minute, time period used in this study from 2000.1.1 to 2003.12.31. The acquisition rate of data was 98.1% level during the period. Missing data were interpolated by the close station data. Ozone and nitrogen dioxide items are the air pollutant analyzed in this study. We selected input variables through principal component analysis for ozone peak modeling. This analysis was performed by S-plus software package.

2.4 Specification of ARX and LSW model

For the selection of ARX and LSW model structure, we considered that AIC, MDL, Fit index etc. are used to determine the order of the model.

\[
AIC = \log \left(1 + \frac{2n}{N}\right) \times V
\]  
(6)

\[
FPE = \frac{1 + n/N}{1 - n/N} \times V
\]  
(7)

\[
MDL = Dim \theta \cdot \frac{\log N}{N}
\]  
(8)

\[
Fit = \frac{\hat{X}_t - X_t}{\sqrt{\text{length}(X_t)}}
\]  
(9)
Here, \( n \) is the number of estimated parameter, \( N \) is the length of the data record, \( V \) is the loss function. In the case of the ARX model, the prediction efficiency is increased as the model order increased. Unknown variances are converged to 0.5 when the model order increased to a large value. We determined ARX structure \([45, 25, 10]\) based on the information.

We also used the MSPE (Mean Squared Prediction Error) \([3]\) value for LSW model structure selection. MSPE is defined in Eq. (10) as follows:

\[
MSPE = E(\hat{X}_{t+h-1,T} - X_{t+h-1,T})^2
\]

(10)

In this study, we determined a rough order considering the ACF (Auto Correlation Function) value for LSW structure in the initial stage. The program performed an iterative calculation for more accurate prediction values. As the final stage, the program determined a proper LSW order using MSPE minimization. LSW\([p,q]=\([23,27]\) was accurate in general and the fit value is over 0.8.

![Figure 2: Zoomed ozone trend components using Coiflet 5 filter during 2003. 6.1-2003.6.14 (a: raw series, b: daily cyclic component, c: short term component less than 24 hours).](image)

### 3 Results and consideration

#### 3.1 Wavelet decomposition of ozone time series

We performed the decomposition processes using hourly measured ozone data in the periods of 2000. 1.1-2003. 12. 31 at Jeongdong station in Seoul. Seasonal
cyclic components are far from the station process. We obtain a better stationary process when long term components are extracted from the raw time series. The stationary characteristic is improved by performing iterative filtering using a wavelet filter.

In this study, we can get similar components to temperature and solar radiation flux at a $2^2$ scale through the number of wave peaks. In Figure 2, the first subplot shows the raw time series, and the second plot is the seasonal components by temperature. The third plot is the solar radiation flux with sine cycle. It is impossible to analyze physical properties about wavelet decomposed series. So we performed partially selected component analysis affected by the temperature and other meteorological data. First we extracted 3 (solar cycle) and then the extracted components will be used as input variables for the prediction model.

In Figure 2, the components are zoomed in for the period of 2003.6.1-2003.6.14. (a) the raw ozone series, (b) the daily cycle, (c) the short term cycles under 24 hour.

![Figure 2: Components of ozone series](image)

Figure 3: The analysis of correlation between prediction value and real value of trend components (a) time series plot during 2003.6.1-6.14 b) correlation between prediction and measurements the same periods).

### 3.2 Prediction of long term components

To build a prediction model, we used the data set from Jeongdong station in the period of 2000.1.1-2002.12.31. For model validation, we used ozone data of the
2003 measured data. These components are forecasted by ARX in Fig. 3. extracted from the raw ozone series.

To select input variables, PCA (principle components analysis) were performed and it is based on the existing dependency among temperature, humidity, solar radiation flux. The relative humidity especially has an anti-correlation with the temperature and solar flux. So we select the input variable with wind speed, solar flux considering independence. The optimum order was of ARX [42,25,10] counting AIC, MDL. The increment of order, however, does not improve the prediction efficiency remarkably, by 3%. So we determined the proper ARX order as ARX [39,21,19].

Figure 3 plots the prediction results for evaluation during 2003.6.1-2003.6.15. The upper plot is the time series plot extracted daily cyclic components and the lower plot is the analysis of correlation between the measured and the predicted values. As a result of the plot, we can predict daily ozone cycle components only using solar flux, wind speed. Prediction efficiency was 98% level in confidence interval of 95%. The correlation coefficient was $R^2 = 0.83$.

3.3 Prediction of short term variation less than 24 hour cyclic component

For the short term cyclic components, we convolved all decomposed series except for series extracted at $2^3$ scale. As mentioned previously, this series have non-linear variation, and partly have non-stationary process (the average 4.5 ppb less than 22.5), the variance decreased 70% level compared with raw data in amplitude. To perform LSW process, we needed to determine [p] the optimum window size, additionally, [q] the moving average parts. This condition is available using ACF (auto correlation function), and PACF (partial auto correlation function). ACF value is converged at zero after 6th in confidence interval 95%. We can get p=6 at the initial stage. So, we can get q=72 using PACF values.

![Figure 4: The scatter plots of prediction and measuring values by ARX (a) and ARX+LSW (b).](image-url)
3.4 Validation of prediction model

Finally, we get the predicted values by merging the trend components (ARX) and the short term components (LSW). Validation period was 3 month from 2003.6.1 to 2003.8.31. Figure 4 is the scatter plot of observed values and predicted value using ARX and ARX + LSW model modified in this study. The regression equation is \( y_p = 0.48 \times x_{obs} + 6.5 \) in classical ARX with the least square method. ARX+LSW is \( y_p = 0.67 \times x_{obs} + 5.7 \). This reveals that ARX prediction values were higher than the modified method in points of correlation coefficients.

In particular, we get the high quality prediction values at over 60 ppb with ARX+LSW, this concentration range is mainly interested in points of alarm or warning system. Accuracy of ARX+LSW were 75% level, this value is 7% higher than classical ARX model. The accuracy of high concentration range only is enhanced by 15%.

3.5 Summary for ozone peak prediction by statistical performance evaluation

In this paper, we used statistical indices to evaluate the results. IOA (index of agreement), Fit, RMSE (Root Mean Square Error) indices were used for evaluation. Under, IOA were defined as follows.

\[
IOA = 1 - \frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} (|P_i - O_{mean}| + |O_i - O_{mean}|)^2}
\]

(11)

Table 1 shows the summary of indices during 2003.6.1-2003.8.31 for the daily peak ozone concentration.

Table 1: Comparison model performance efficient between ARX and ARX+LSW.

<table>
<thead>
<tr>
<th>Item</th>
<th>ARX model</th>
<th>ARX merged LSW</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>0.67</td>
<td>0.63</td>
</tr>
<tr>
<td>Fit</td>
<td>89.5</td>
<td>95.2</td>
</tr>
<tr>
<td>RMSE</td>
<td>15.9</td>
<td>16.9</td>
</tr>
<tr>
<td>IOA</td>
<td>0.66</td>
<td>0.92</td>
</tr>
</tbody>
</table>
As demonstrated in Table 1, it is proper to use the ARX+LSW model to efficiently predict with similar correlation coefficients. Especially, the IOA is improved by 30% when LSW+ARX process is applied.

4 Conclusion

This paper describes the new methodology, based on the wavelet technique and the AR model, to predict ozone peak concentration. The prediction model is constructed by introducing LSW theory to the conventional ARX. The conventional ARX model has disadvantages in describing peak concentration but effectively describes variation of daily trend. This characteristic of the ARX model is also limited to apply alarm or warning system, because the peak concentration is more important for ozone which could be seriously harmful to humans by the strong oxidation potential.

In this study, we developed a new methodology for an ozone peak prediction model. This method was modified ARX with LSW. As the result of modification, the ARX+LSW model is effective in predicting the peak ozone levels. We confirm the relatively small input requirement compared to the conventional ARX model by combining LSW theory. Consequently, the model of ARX+LSW in this study has a high potential to apply to the ozone warning/alarm system. It is expected that such a modified model would be superior to forecasting than a multiple time series AR model.

References

