A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile

- Atmospheric Environment 42, 2008 -

Luis A. Diaz-Robles, Juan C. Ortega, Joshua S. Fu, Gregory D. Reed, Judith C. Chow, John G. Watson, Juan A. Moncada-Herrera

2012.05.21 (Mon.)
Kwangbok, Jeong
CONTENTS

1. Air quality forecasting models

2. Methodology
   2.1 Data set
   2.2 Variable selection and models construction
   2.3 Linear & Non-linear approach model
   2.4 Measures of accuracy applied in the models performance

3. Results and discussions

4. Conclusions
1. Air quality forecasting model

- **PM10 (Particulate Matter Less than 10μm)**
  - 입자의 크기가 10μm 이하인 먼지.
  - 인체의 폐포까지 침투하여 각종 호흡기 질환의 직접적인 원인이 되며, 인체의 면역기능을 약화시킴.
  - 반면, 입자의 크기가 10μm 이상인 경우에는 인체 건강에 미치는 영향이 적음
  - 따라서, PM$_{10}$을 기준으로 각국은 환경기준을 정하고 있음.

- **Episode level for air quality index (ICAP), Chile**

<table>
<thead>
<tr>
<th>Air quality levels for PM$_{10}$(μgm$^{-3}$)</th>
<th>ICAP</th>
<th>Ventilation condition</th>
<th>Episode type</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 ~ 194</td>
<td>0 ~ 200</td>
<td>Good to regular</td>
<td>None</td>
</tr>
<tr>
<td>195 ~ 239</td>
<td>201 ~ 300</td>
<td>Bad</td>
<td>Alert</td>
</tr>
<tr>
<td>240 ~ 329</td>
<td>301 ~ 500</td>
<td>Critical</td>
<td>Pre-emergency</td>
</tr>
<tr>
<td>&gt; 330</td>
<td>&gt; 501</td>
<td>Dangerous</td>
<td>Emergency</td>
</tr>
</tbody>
</table>
1. Air quality forecasting model

Air quality forecasting models

- Autoregressive Integrated Moving Average (ARIMA) and Multilinear regression (MLR) models have been widely used for air quality forecasting in urban areas. (Goyal et al., 2006)
- Artificial neural networks (ANN) have been developed as a non-linear tool for pollution forecasting. (Perez and Reyes, 2002; Perez et al., 2004; Perez and Reyes, 2006; Schlink et al., 2006; Slini et al., 2006; Sofuoglu et al., 2006; Sousa et al., 2006; Thomas and Jacko, 2007)

방법론 장·단점

<table>
<thead>
<tr>
<th>ARIMA</th>
<th>ANN</th>
<th>MLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Better to capture the linear pattern of a time series</td>
<td>Better to capture the non-linear patterns of a time series</td>
<td>Better to capture the linear pattern of a time series</td>
</tr>
<tr>
<td>Great versatility</td>
<td>Great versatility</td>
<td>Few Versatility</td>
</tr>
<tr>
<td>Better for seasonal patterns</td>
<td>Better to capture noise and extreme values (episodes)</td>
<td>Needs correction factors to capture extreme values</td>
</tr>
<tr>
<td>Needs historical data continuity</td>
<td>Does not need historical data continuity</td>
<td>Does not need historical data continuity</td>
</tr>
</tbody>
</table>

Time series forecasting using a hybrid ARIMA and neural network model
Zhang, G, P, 2003
1. Air quality forecasting model

- Former studies using hybrid ARIMA and ANN

  - Forecasting daily maximum ozone ($O_3$) concentration at Houston (Prybutok et al. 2000)
    - ANN model was more accurate than either the ARIMA or MLR models
    - Because of their non-linear patterns

  - Goyal et al. 2006
    - pointed out those linear models such as MLR and ARIMA fail to predict extreme concentrations (episodes)

    - He tested a hybrid ARIMA and ANN model over three kinds of time series.
    - It was concluded that the linear and non-linear time series patterns in the combined model improved forecasting more than either of the models used independently.
2. Methodology

Study and available data

- Hourly and daily time series of PM$_{10}$ and meteorological data (2000 ~ 2006)
- Temuco, the capital of the Araucania region of Chile
- Population: About 300,000
- Rainy climate with Mediterranean influence
- Temuco has had a fast urban expansion; great economic growth; increased woodstove and industrial source emissions; and increasing vehicle exhaust.
- Air quality monitoring station (hourly temperature, relative humidity, wind direction, and wind speed)
2. Methodology

Training and validation data sets

<table>
<thead>
<tr>
<th>Data</th>
<th>Number of observations</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>2,080</td>
<td>2000.07.21</td>
<td>2006.03.31</td>
</tr>
<tr>
<td>Validation</td>
<td>183</td>
<td>2006.04.01</td>
<td>2006.09.30</td>
</tr>
</tbody>
</table>

Maximum 24-h PM10 moving average by beta attenuation monitor time series at the Las Encinas site in Temuco, Chile (2000.07.21~2006.09.30)
2. Methodology

- **Linear approach: Multiple Linear Regression (MLR)**
  - Training data: 92%, Validation data: 8%
  - Predictor variables (statistically significant)
    - Previous day maximum hourly PM$_{10}$ (L1PM10), WS, $T_{\text{min}}$, and $T_{\text{max}}$
    - $Y = b_0 + b_1x_1 + \cdots + b_qx_q$

- **Linear approach: Box-Jenkins ARIMA model**
  - 시계열 자료: 시간흐름에 따라 관측된 (발생하는) 자료
    - (예) 주가, 매출액, 관광객수, 교통량, 이자율, 범죄율 ...

![Graph showing time series data](image)
2. Methodology

- **ARIMA model**
  - 단기예측에 적합: 먼 과거의 자료보다 최근 시점에 가까운 자료에 더 많은 비중을 두기 때문.
  - 계절적 변동(즉, 주기적 변동)이 존재하는 시계열 자료를 예측할 때 유용.
  - 최소 50개 이상의 관측값 필요 (Box-jenkins, 1994)
  - 자기상관함수: 현재의 상태가 과거와 미래의 상태와 밀접한 상관관계를 가지는 것을 함수로 표현

- **ARIMA(p,0,0) = AR(p)**
  - \( Y_t = \theta_0 + \varnothing_1 Y_{t-1} + \varnothing_1 Y_{t-2} + \cdots + \varnothing_p Y_{t-p} + e_t \)

- **ARIMA(0,0,q) = MA(q)**
  - \( Y_t = \mu - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} + e_t \)

- **ARIMA(p,0,q)**
  - \( Y_t = \theta_0 + \varnothing_1 Y_{t-1} + \varnothing_1 Y_{t-2} + \cdots + \varnothing_p Y_{t-p} + e_t + \mu - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} + e_t \)
2. Methodology

- **Non-Linear approach: Artificial neural network (ANN)**
  - Training data: 92%, Validation data: 8%
  - Predictor variables (*statistically significant*)
    - Previous day maximum hourly PM$_{10}$ (L1PM10), WS, $T_{\text{min}}$, and $T_{\text{max}}$
  - Levenberg-Marquardt algorithm & MLP
    - Iterative technique that finds a local minimum of a function that is expressed as the sum of squares of nonlinear function
    - It can be thought of as a combination of steepest descent and the Gauss-Newton method.
    - Steepest descent: current solution is far from the correct one
    - Gauss-Newton: current solution is close to the correct one

[* www.ics.forth.gr*]
2. Methodology

Result and discussion (1/2)

- **MLR model**
  \[
  \text{MaxPM}_{10\text{ma}} = 0.80845(L1\text{PM}_{10}) - 0.13283(\text{WS}) - 0.15332(\text{T}_{\text{min}}) + 0.05605(\text{T}_{\text{max}})
  \]

- **ARIMA**
  \[
  Y_t = 0.98122Y_{t-1} + \mu - 0.87057e_{t-1} + e_t
  \]

- **ARIMAX**
  \[
  \text{MaxPM}_{10\text{ma}} = \text{ARIMA}(1,0,1) + 0.60363(L1\text{PM}_{10}) - 0.15999(\text{WS}) - 0.18140(\text{T}_{\text{min}}) + 0.09197(\text{T}_{\text{max}})
  \]

- The wind direction was not a significant variable, suggesting that the sources of PM were more local rather than transported from other regions.

- Autoregressive component of the ARIMA model is more significant than the moving average component.
Result and discussion (2/2)

- Hybrid model
  - **Input variables**: Forecasted ARIMA MaxPM10ma, the errors associated with the ARIMAX model, L1PM_{10}, WS, T_{\text{min}}, T_{\text{max}}
  - **MLP architecture**
    (1 hidden layer, 3 neurons)
  - **Square activation function**
3. Result

Model performance statistics for the validation data set

<table>
<thead>
<tr>
<th>Estimator</th>
<th>MLR</th>
<th>ARIMAX</th>
<th>ANN</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE, µg m(^{-3})</td>
<td>28.39</td>
<td>28.46</td>
<td>28.57</td>
<td>8.80</td>
</tr>
<tr>
<td>MAE, µg m(^{-3})</td>
<td>20.83</td>
<td>19.87</td>
<td>20.65</td>
<td>6.74</td>
</tr>
<tr>
<td>BIC</td>
<td>2198.9</td>
<td>2210.1</td>
<td>2216.7</td>
<td>1801.2</td>
</tr>
</tbody>
</table>

- **RMSE** (root mean square error; 평균제곱오차의 제곱근)
  
  \[ \text{RMSE} = \sqrt{\frac{1}{n-k} \sum (y_t - \hat{y}_t)^2} \]

- **MAE** (mean absolute error; 평균절대오차)
  
  \[ \text{MAE} = \frac{1}{n} \sum |y_t - \hat{y}_t| \]

- **BIC** (Baysian Information Critrion)
  
  \[ \text{BIC} = n \log(SSE) + m \log(n) \]
3. Result

- Model performance using the validation data set

![Graphs showing model performance](image-url)
### 3. Result

**Contingency table for the MLR model of the validation data set**

<table>
<thead>
<tr>
<th>Obs.</th>
<th>Forecast</th>
<th>None</th>
<th>Alert</th>
<th>Pre-emergency</th>
<th>Emergency</th>
<th>Tot.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td>169</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>171</td>
<td>99</td>
</tr>
<tr>
<td>Alert</td>
<td></td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td>Pre-emergency</td>
<td></td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Emergency</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tot.</td>
<td></td>
<td>174</td>
<td>6</td>
<td>3</td>
<td></td>
<td>183</td>
<td>94</td>
</tr>
<tr>
<td>%</td>
<td></td>
<td>97</td>
<td>33</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Contingency table for the ARIMA(1,0,1)-X model of the validation data set**

<table>
<thead>
<tr>
<th>Obs.</th>
<th>Forecast</th>
<th>None</th>
<th>Alert</th>
<th>Pre-emergency</th>
<th>Emergency</th>
<th>Tot.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td>170</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>171</td>
<td>99</td>
</tr>
<tr>
<td>Alert</td>
<td></td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td>Pre-emergency</td>
<td></td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Emergency</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tot.</td>
<td></td>
<td>176</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>183</td>
<td>94</td>
</tr>
<tr>
<td>%</td>
<td></td>
<td>97</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## 3. Result

### Contingency table for the **ANN model** of the validation data set

<table>
<thead>
<tr>
<th>Obs.</th>
<th>None</th>
<th>Alert</th>
<th>Pre-emergency</th>
<th>Emergency</th>
<th>Tot.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>170</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>171</td>
<td>99</td>
</tr>
<tr>
<td>Alert</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>7</td>
<td>100</td>
</tr>
<tr>
<td>Pre-emergency</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td>Emergency</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tot.</td>
<td>175</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>183</td>
<td>95</td>
</tr>
<tr>
<td>%</td>
<td>97</td>
<td>0</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

### Contingency table for the **hybrid model** of the validation data set

<table>
<thead>
<tr>
<th>Obs.</th>
<th>None</th>
<th>Alert</th>
<th>Pre-emergency</th>
<th>Emergency</th>
<th>Tot.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>170</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>171</td>
<td>99</td>
</tr>
<tr>
<td>Alert</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>100</td>
</tr>
<tr>
<td>Pre-emergency</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>80</td>
</tr>
<tr>
<td>Emergency</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tot.</td>
<td>170</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>183</td>
<td>99</td>
</tr>
<tr>
<td>%</td>
<td>100</td>
<td>88</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
4. Conclusion

**Conclusion**

- The hybrid models took advantage of the unique capabilities of ARIMAX and ANN in linear and non-linear modeling.
- So, Hybrid model can be an effective tool to improve the forecasting accuracy.
- But, the hybrid model has to be run every day to forecast the next day using the meteorological and air quality data.

프로젝트 적용에 고려할 사항

- 변수 (previous day maximum hourly PM$_{10}$, wind speed, T$_{min}$, T$_{max}$) 선택에 대한 이유가 불충분 함.
- 예측 하루 전날의 데이터를 이용하여 다음날을 예측하는 부분.
Thank you.