Letters

Identification of significant factors for air pollution levels using a neural network based knowledge discovery system

Kit Yan Chan\(^a\), Le Jian\(^b\)\(^,*\)

\(^a\) Department of Electrical and Computer Engineering, Curtin University, Perth, Australia
\(^b\) School of Public Health, WHO Collaborating Center for Environmental Health Impact Assessment, Curtin Health Innovation Research Institute, Faculty of Health Sciences, Curtin University, Kent Street, GPO Box U1987, Perth, Western Australia 6845, Australia

ARTICLE INFO

Article history:
Received 4 May 2012
Received in revised form 15 June 2012
Accepted 20 June 2012
Communicated by V. Palade
Available online 1 July 2012

Keywords:
Artificial neural network
Main effect analysis
Air pollution
Air monitoring
Meteorological factors
Particulate matter

ABSTRACT

Artificial neural network (ANN) is a commonly used approach to estimate or forecast air pollution levels, which are usually assessed by the concentrations of air contaminants such as nitrogen dioxide, sulfur dioxide, carbon monoxide, ozone, and suspended particulate matters (PMs) in the atmosphere of the concerned areas. Even though ANN can accurately estimate air pollution levels they are numerical enigmas and unable to provide explicit knowledge of air pollution levels by air pollution factors (e.g. traffic and meteorological factors). This paper proposed a neural network based knowledge discovery system aimed at overcoming this limitation in ANN. The system consists of two units: (a) an ANN unit, which is used to estimate the air pollution levels based on relevant air pollution factors; (b) a knowledge discovery unit, which is used to extract explicit knowledge from the ANN unit. To demonstrate the practicability of this neural network based knowledge discovery system, numerical data on mass concentrations of PM2.5 and PM1.0, meteorological and traffic data measured near a busy traffic road in Hangzhou city were applied to investigate the air pollution levels and the potential air pollution factors that may impact on the concentrations of these PMs. Results suggest that the proposed neural network based knowledge discovery system can accurately estimate air pollution levels and identify significant factors that have impact on air pollution levels.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Air pollution is a major environmental risk to health in many developed and developing cities of the world. The air pollution levels are usually determined by concentrations of air pollutants such as nitrogen dioxide, sulfur dioxide, carbon monoxide, ozone and suspended particulate matters (PMs). PMs are defined by the U.S. Environmental Protection Agency as “very small pieces of solid or liquid matter, such as particles of soot, dust, fumes, mists, or aerosols” [10]. They are usually produced by air pollution factors such as energy production from power plants, burning of fossil fuels in factories, power plants, industrial processes, residential heating, combustion of gasoline, diesel and hydrocarbon fuels in vehicles, etc. [5,21]. Meanwhile, unfavorable meteorological factors may also affect the formation and growth of new air pollutants and the ability of the atmosphere to disperse air pollutants [3,25,38,40]. Severe air pollution levels can be life threatening, can cause breathing difficulty, headache, dizziness, and result in heart attack [20]. Long term exposure to air pollutants can result in chronic respiratory and cardiovascular diseases including cancers [9,17,24]. Therefore, it is essential to monitor criteria air pollutants in the atmosphere by developing accurate models which can indicate the relationship between air pollution levels and air pollution factors. However, it is difficult to develop such models using traditional statistical methods, as they are unable to model complex nonlinear relationships between air pollution factors [1]. More recently, a universal estimator, namely artificial neural network (ANN), has been demonstrated its capability to model non-linear relationships between input and output variables to estimate, evaluate and forecast air pollution levels [12,22,41,42]. ANN is an unsupervised learning technique whereby collected air pollution data are trained in order to create a black-box model, which maps between two domains, namely (i) the domain for air pollution factors and (ii) the domain for air pollution levels. In the literature, ANN has been applied to estimate air pollutant levels such as concentrations of sulfur dioxide [6], carbon monoxide [27,29], PMs [13,33,36,43], and ozone [48]. However, ANN has a black-box structure which yields no explicit knowledge [2]. Because of this limitation, traditional ANNs may not be appropriate to use in estimating air pollution levels, even though they may achieve more accurate estimates than other explicit modeling methods such as statistical methods [30].
In this paper, a neural network based knowledge discovery system is proposed not only to estimate air pollution levels, but also to generate relative significance of air pollution factors to air pollution levels. This knowledge is important because it provides researchers with a better insight into the extent to which a particular air pollution factor impacts on air pollution levels. As road traffic is a major source of air pollution, this knowledge is also useful for the transportation and infrastructure sectors to estimate pollution levels before planning or developing new transportation infrastructures.

2. The neural network based knowledge discovery system

A schematic representation of the proposed neural network based knowledge discovery system is depicted in Fig. 1. The system consists of two main units: (a) the artificial neural network unit, namely ANN unit, and (b) the knowledge discovery unit, namely KD unit. The ANN unit is developed to estimate the specified indicators of air pollution level, y (e.g., nitrogen monoxide or PMs), based on the n air pollution factors, $x_1, x_2, \ldots, x_n$, which consist of meteorological and traffic flow factors. The indicators of air pollution level can be the levels of PMs or gaseous pollutants such as ozone, nitrogen monoxide and nitrogen dioxide. The meteorological factors can be temperature, relative humidity, barometric atmosphere pressure and wind speed. When data on both air pollution levels and air pollution factors are collected, an ANN unit can be developed. The KD unit is developed to extract informative knowledge from the ANN unit, which traditionally is a black-box or implicit in nature, and no explicit information can be indicated clearly between $x_1, x_2, \ldots, x_n$ and y. Based on the new KD unit, the relative significance of the concerned air pollution factors, that mainly affect the air pollution level, can be indicated. If a slight change in an air pollution factor results in a great change in the air pollution level, this air pollution factor is considered as a critical contributor to the air pollution level. Therefore, it is essential and important to consider those significant factors in analyzing and controlling air pollution level.

2.1. The artificial neural network unit

The ANN unit consists of three types of neural nodes: (a) the input nodes, which feed the air pollution factors $x_1, x_2, \ldots, x_n$ into the ANN unit; (b) the output node, which estimates the air pollution level, y; and (c) the hidden nodes, which link input nodes of the air pollution factors and the output nodes of the air pollution level. The ANN unit is suitable for complex and nonlinear interactions between air pollution factors in order to estimate the air pollution level based on the input-output functional relationship, $f$, which is denoted by the Eq. (1):

$$y = f(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{n} w_j y_j (\sum_{i=1}^{n} (v_i x_i - b_j)) - b \quad (1)$$

where $n_h$ denotes the number of hidden nodes of the ANN unit, $w_j$ denotes the weight of the link between the j-th hidden node and the output node for the air pollution level, y with $j = 1, 2, \ldots, n_h$, $v_i, b_j, b$ denote the biases for the j-th hidden nodes and the output node respectively, and $\Psi$ is the transfer function of the hidden set in which the sigmoid function is used.

The ANN weights are determined based on $N_d$ pieces of collected air pollution data in the form of

$$d(k) = [y(k), \phi(k)] \quad \text{with } k = 1, 2, \ldots, N_d \quad (2)$$

where $y(k)$ and $\phi(k) = [x_1(k), x_2(k), \ldots, x_n(k)]$ are the k-th air pollution data with respect to the air pollution level and the n air pollution factors, respectively. The ANN unit is evaluated based on the mean absolute relative error, $\epsilon_{\text{MARE}}$, formulated in Eq. (3), where both small and large errors have the same weights. $\epsilon_{\text{MARE}}$ indicates the differences between the actual observations and the estimates of the ANN unit:

$$\epsilon_{\text{MARE}} = \frac{1}{N_d} \sum_{k=1}^{N_d} \left| \frac{y(k) - \hat{y}(k)}{y(k)} \right| \quad (3)$$

where $\hat{y}(k)$ is the estimate based on Eq. (1) with respect to $\phi(k)$ and $y(k) \neq 0$. The Levenberg–Marquardt algorithm is then used to train the ANN unit by minimizing $\epsilon_{\text{MARE}}$ [15]. It starts by randomly generating the first two initial ANN weights, $w(0)$ and $\omega(1)$, at the 0-th and the 1-st iterations, where

$$w(0) = [w_{10}(0), w_{20}(0), \ldots, w_{n0}(0), v_{11}(0), v_{12}(0), \ldots, v_{n1}(0), v_{21}(0), v_{22}(0), \ldots, v_{n2}(0), v_{31}(0), v_{32}(0), \ldots, v_{n3}(0)]$$

Fig. 1. Framework of neural network based knowledge discovery system.
The numerator in Eq. (7) represents the total change with respect to \( f \), when \( x_i \) changes from a level to another. When the total change is small, \( x_i \) is not significant to \( f \). Hence, the outcome of \( f \) changes slightly, even though \( x_i \) changes largely. In another extreme, if the total change is large, \( x_i \) is significant to \( f \). Hence, the outcome of \( f \) changes largely, even \( x_i \) only change slightly. We can consider a simple case with only two levels, 'high' and 'low' levels, where \( N_i = 2 \).

The significance of \( x_i \) can be evaluated by the difference between the main effect in 'low' level and the main effect in 'high' level. When the difference is small, the outcome of \( f \) changes slightly whenever \( x_i \) is in 'low' or 'high' level. Therefore, \( x_i \) is not too significant. Otherwise, when the difference is large, the significance of \( x_i \) is large.

The relative significance of \( x_i \), is given by

\[
d_i = \frac{d_i}{\sum_{j=1}^{N} d_j} \times 100\%
\]

where \( d_i \) indicates the relative significance of the \( i \)-th air pollution factor, \( x_i \), with respect to the air pollution level, \( y \), and \( d_i \) is relative to the total significance of all air pollution factors. The analysis of relative significances can be organized more efficiently, as \( d_i \) represents the rate of the \( i \)-th air pollution factor to the total significance of all air pollution factors.

3. A case study of ambient particulate matter concentrations

In order to illustrate the operation of the neural network based knowledge discovery system, data from a case study of ambient air monitoring was undertaken to estimate the air pollution level with respect to the ambient PM concentrations, which are important indicators of ambient air quality because of their detrimental effects on health and visibility impairment [33,40]. Here, the concentrations of PM2.5 (with aerodynamic diameter of PM \( \leq 2.5 \mu m \)) and PM10 (with aerodynamic diameter of PM \( \leq 10.0 \mu m \)) were studied because they are the indicators of fine and ultrafine particles that can enter the thorax and lower respiratory tract. Although a number of studies in the last decade have quantified and characterized PM2.5 in China [7,11,14,16,43,44], there is relatively scarce research on traffic related PM10 in China [23,35].

This case study was conducted near a busy road (Zhong He Viaduct) in the city center of Hangzhou in 2010. The total number of newly registered on-road cars, buses and trucks in 2008 was 3.6 times greater than the numbers in 2000; the number of gasoline fueled vehicles increased 4.4 times and diesel fueled vehicles increased 1.2 times [18]. Zhong He Viaduct is a two-way vehicle-only viaduct with two lanes in each direction. The length of the viaduct is about 20 km from north to south with 10 exits on each side. In order to develop the models for estimating the concentrations of PMs at the roadside, the data on PM2.5 and PM1.0, as well as meteorological variables and the traffic flow were collected. The concentrations of PM2.5 and PM1.0 were measured by TSI DustTrak DRX Aerosol Monitor 8533, which was calibrated by the manufacturer to the respirable fraction of the standard ISO 12103, A1 Arizona road dust. The DustTrak DRX can simultaneously measure multiple size segregated mass fractions of the sampled aerosol, including PM2.5 and PM1.0 [39]. Two models namely \( f_1 \) and \( f_2 \), which estimate PM2.5 and PM1.0 at time \( t \) namely \( y_1(t) \) and \( y_2(t) \), were developed based on Eqs. (10) and (11) respectively:

\[
\dot{y}_1(t) = f_1(x_1(t), x_2(t), x_3(t), x_4(t), x_5(t), y_1(t-T_i))
\]

and

\[
\dot{y}_2(t) = f_2(x_1(t), x_2(t), x_3(t), x_4(t), x_5(t), y_2(t-T_i))
\]

where \( T_i \) is the sampling time; \( x_1(t), x_2(t), x_3(t) \) and \( x_4(t) \) are denoted as the meteorological variables temperature (°C), relative
humidity RH (%), wind speed \((\text{m s}^{-1})\) and barometric pressure (hPa), respectively, at time \(t\). The meteorological data were collected by using TSI 9555 A Advanced Anemometer; \(x_i(t)\) is denoted as the traffic flow at time \(t\) and was measured by a real-time traffic surveillance system automatically counting the number of vehicles passing the surveillance point. This traffic flow data was provided by the Hangzhou City Traffic Control and Administration Center; \(y_1(t-T_1)\) and \(y_2(t-T_2)\) are PM2.5 and PM1.0 measured at time \(t\), respectively. Previous studies \([18,19,37]\) indicated that traffic flow was a significant predictor of particles from vehicle emissions, and those meteorological factors were also significant estimators in forecasting roadside atmospheric concentrations of submicron particles \([45]\). Hence, those air pollution factors were selected in this study. All those data were collected from 14th May to 16th May from 7:30 am to 15:30 pm with a sampling interval time of 1 min (i.e. \(T_i=1\)). Hence, 24 h of data or a total of 1440 pieces of data were collected.

Apart from using ANN unit, linear regression \([34]\) was used to develop models for \(f_1\) and \(f_2\), which are formulated in Eqs. (10) and (11), respectively. To evaluate the performance of all these models, cross validation, namely repeated random sub-sampling validation, was carried out using the same data mentioned above. All those data were collected from 14th May to 16th May from 7:30 am to 15:30 pm with a sampling interval time of 1 min (i.e. \(T_i=1\)). Therefore, 24 h of data or a total of 1440 pieces of data were collected.

### Table 1

<table>
<thead>
<tr>
<th>Validation number</th>
<th>PM2.5 - Linear Regression</th>
<th>ANN</th>
<th>PM1.0 - Linear Regression</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.64</td>
<td>7.74</td>
<td>7.88</td>
<td>5.92</td>
</tr>
<tr>
<td>2</td>
<td>6.02</td>
<td>5.16</td>
<td>14.96</td>
<td>13.45</td>
</tr>
<tr>
<td>3</td>
<td>10.85</td>
<td>9.91</td>
<td>9.74</td>
<td>6.71</td>
</tr>
<tr>
<td>4</td>
<td>8.71</td>
<td>6.47</td>
<td>16.00</td>
<td>13.39</td>
</tr>
<tr>
<td>5</td>
<td>7.13</td>
<td>4.26</td>
<td>10.50</td>
<td>8.27</td>
</tr>
<tr>
<td>6</td>
<td>11.41</td>
<td>7.81</td>
<td>11.29</td>
<td>8.91</td>
</tr>
<tr>
<td>7</td>
<td>10.62</td>
<td>9.04</td>
<td>8.87</td>
<td>4.86</td>
</tr>
<tr>
<td>8</td>
<td>7.89</td>
<td>6.19</td>
<td>7.86</td>
<td>6.37</td>
</tr>
<tr>
<td>9</td>
<td>11.34</td>
<td>10.34</td>
<td>7.71</td>
<td>4.02</td>
</tr>
<tr>
<td>10</td>
<td>7.67</td>
<td>5.78</td>
<td>13.9</td>
<td>11.34</td>
</tr>
<tr>
<td>11</td>
<td>10.02</td>
<td>7.78</td>
<td>7.93</td>
<td>6.78</td>
</tr>
<tr>
<td>12</td>
<td>8.34</td>
<td>6.41</td>
<td>8.72</td>
<td>4.18</td>
</tr>
<tr>
<td>13</td>
<td>9.70</td>
<td>8.02</td>
<td>16.25</td>
<td>11.76</td>
</tr>
<tr>
<td>14</td>
<td>19.34</td>
<td>18.20</td>
<td>7.95</td>
<td>6.70</td>
</tr>
<tr>
<td>15</td>
<td>7.69</td>
<td>5.84</td>
<td>6.02</td>
<td>5.04</td>
</tr>
<tr>
<td>Mean ± SD</td>
<td>9.69 ± 3.12</td>
<td>7.93 ± 3.32</td>
<td>10.24 ± 3.44</td>
<td>7.85 ± 3.22</td>
</tr>
<tr>
<td>t value</td>
<td>8.83</td>
<td>8.18</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Fig. 2

Relative significance of air pollution factors contributing to PM2.5 (a) and PM1.0 (b).
(10.24% vs. 7.85%, t value = 8.18, P  0.05). Therefore, the generalization capability of the models developed by the ANN unit in predicting both PM2.5 and PM1.0 levels is significantly better than that of the statistical regression model. Based on the KD unit, the main effect of each variable in f1 (i.e. PM2.5) and f2 (i.e. PM1.0) can be determined and are shown in Fig. 2(a) and (b), respectively. The results indicate the relative significance of each air pollution factor with respect to the concentrations of PM2.5 and PM1.0. Compared with meteorological factors (wind speed, x3(t), temperature, x4(t), relative humidity, x5(t) and barometric pressure, x6(t)), traffic flow, x7 (t), provides more contribution to the estimated concentrations of PM2.5 and PM1.0. This is unsurprising, as the increase of the total number of vehicles passing through the area is associated with increased emission of PM2.5 and PM1.0. In addition, the four measured meteorological factors also have influence on the concentrations of PM2.5 and PM1.0, but at a less significant extent compared with that from the traffic flow. Results from Fig. 2 also indicate that wind speed plays more important role than other meteorological factors on PM2.5 concentration, but barometric pressure, exceed wind speed, impacts more on PM1.0 concentration. In summary, this KD unit overcomes the limitation of current neural network approaches whereby no explicit information can be indicated within the neural networks.

In addition, a hypothesis test was used to evaluate the significance of the air pollution factors based on the result of the linear regression model. The t-values are used to indicate whether the air pollution factors are significant or not. When the t-value of the corresponding air pollution factor is less than 2.09, this air pollution factor is considered insignificant to the air pollution level with 98% confidence interval. Otherwise, this air pollution factor is considered significant. The t-values of the air pollution factors with respect to emissions of PM2.5 and PM1.0 are shown in Table 2. The t-values are bolded, when they are large than 2.09. For PM2.5, the t-values for wind speed and traffic flow are large than 2.09. Hence, it indicates that wind speed and traffic flow are significant factors to PM2.5. For PM1.0, it also indicates that wind speed and traffic flow are the two significant factors. The t-value of barometric pressure is 2.00 which is near to the significant level. Hence, barometric pressure is somewhat important to PM1.0 compared with other insignificant weather factors. The results of hypothesis tests are similar to those obtained by the KD, where both wind speed and traffic flow are significant to PM2.5 and PM1.0, and barometric pressure is relatively significance to PM1.0 compared with other insignificant pollution factors.

4. Conclusions

In this paper, a neural network based knowledge discovery system has been developed to estimate air pollution levels based on a set of measured air monitoring data. Cross validation tests and the case study results have demonstrated that this new system is able to overcome the limitation of traditional ANNs and generate accurate explicit knowledge of the significant contribution of each pollution factor on air pollution levels (PM2.5 and PM1.0) within the existing ANNs for estimating air pollution levels. In another word, the system is able to estimate PM2.5 and PM1.0 concentrations based on traffic flow, meteorological conditions at a busy traffic roadside, and past measured PM concentrations. Based on this explicit knowledge, researchers can gain a better insight into the significance and influence of a particular air pollution factor on the mass concentrations of PMs. Therefore, the system has potential application value in planning future air monitoring programs to achieve cost-effective outcomes.

As for future work, we will develop a rule discovery system [31,32] which extracts symbolic rules from ANNs, in order to illustrate relations between air pollution factors and air pollution levels. We will also enhance the effectiveness of the ANNs by integrating the mechanism of hybrid approaches [28].

Acknowledgments

The authors would like to thank staff at Hangzhou Traffic Administration and Control Center, The Vehicle Licensing Center and The Meteorological Data Service Center for their kind help in providing relative data for this study.

References

K. Yan Chan, L. Jian / Neurocomputing 99 (2013) 564–569


