ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20180204162632
Original scientific paper

# A Deep Belief Network Based Model for Urban Haze Prediction

Huimin LU, Jingjing SONG, Tianyi DI, Jamshid MORADI KURDESTANY, Hongzhi WANG

**Abstract:** In order to improve the accuracy of urban haze prediction, a novel deep belief network (DBN)-based model was proposed. Firstly, data pertaining to both air quality and the environment (e.g. meteorology) data was monitored and collected. The primary haze influencing elements were discovered by analyzing the correlations between each of the meteorological factors and haze. Secondly, a DBN combined with multilayer restricted Boltzmann machines and a single-layer back propagation network was applied. Thirdly, the meteorological data predictions were carried out by using a competitive adaptive-reweighed method. A stable model was established by big-data training and its accuracy was verified by experiments. Results demonstrate that the pollution haze occurs in accordance with regular laws, and is greatly affected by wind direction, atmospheric pressure, and seasons. The correlation coefficient (CC) between the actual haze value and the prediction of the proposed model is 0.8, and the mean absolute error (MAE) is 26 μg/m$^3$. Compared with the traditional prediction algorithms, the CC is improved by 18 % on average, while the MAE is reduced by 15.7 μg/m$^3$. The proposed method has a good prospect to predict haze and investigate the main causes of it. This study provides data support for urban haze prevention and governance.

**Keywords:** big data; deep belief network; deep learning; haze prediction; model

## 1 INTRODUCTION

Hazy days are mainly caused by suspended particulate matter (*PM*), especially the influence of fine *PM* (e.g. $PM_{2.5}$). Haze causes low atmospheric visibility, which not only disturbs the 'outdoor activity' of citizens, and thus influences their lives, but also results in serious damage to their health. Air pollution caused by $PM_{2.5}$ undergoes strong dynamic changes and leads to regional differences. The important factors affecting haze pollution therefore involve emission and regional transmission of pollution, secondary aerosol pollution, and weather conditions. However, these factors are subjected to uncertainty, resulting in a high degree of randomness in the spatial and temporal distributions of the haze [1]. Predicting urban haze pollution levels is a complex project. Therefore, how to improve the accuracy of haze prediction and how to provide support for preventing and controlling haze pollution becomes a key study issue in the field of environmental science.

The present existing methods for predicting urban air quality can be classified into three categories: potential, numerical, and statistical predictions. Potential predictions involving secondary predictions are based on weather forecast, and show monotonic features and poor prediction accuracy. Based on aerodynamic theory, numerical predictions use mathematical models for the diluted and diffused air pollutants, and also have a few defects such as being computationally complex and providing a poor real-time prediction performance. In contrast, statistical predictions have attracted numerous scholars' attention due to their speed and simplicity. Statistical predictions dissect the statistical rules governing the input-output of resources related to the air pollution, rather than depending on the physical, chemical, and biological aspects of the pollutants themselves.

Saeed, Hussain, Awan, & Idris made a comparative analysis of models for forecasting air quality using different statistical theories [2]. Nonetheless, the traditional statistical methods of predicating air quality suffer from low prediction accuracy as they simplify the factors of influence and make too many hypotheses. The prediction algorithms using artificial neural networks (ANNs) are hard to implement because of the complexity (especially in the determination of the network structure) and time-consuming training although the prediction accuracy can be ameliorated. Hence, how to build a rapid and precise haze prediction model and how to identify the crucial elements influencing the development of haze as well as the correlation between various factors and haze are urgently needed to be solved.

Based on the above discussion, this study aims to establish a deep belief network (DBN) based model for urban haze prediction by using deep learning (DL) theory to accurately predict the tendency of haze development. Moreover, by thoroughly analyzing the key factors affecting haze formation and the correlations between these factors and haze, the timeliness and accuracy of the haze predictions can be more enhanced.

## 2 STATE OF THE ART

Traditional models for air pollution prediction mainly involve linear and nonlinear regression, and other techniques such as auto-regressive and moving average (ARMA), classification, regression tree (CART), and ANN methods. Singh, Gupta, Kumar, & Shukla evaluated the dependency relationships between air quality and factors of influence using linear and nonlinear modelling methods to forecast urban air quality [3]. Utilizing an ARMA model, Li, Peng, Shao, Cui, & Tian studied the stability, autocorrelation, and partial correlation of a data series to further predict by treating air quality as a time series [4]. Established on a CART, Chen selected an optimal penalty parameter by a generalized cross validation to produce an optimal regressive tree model and thus realized air quality predictions [5]. Singh, Gupta, & Rai made predictions of the urban air quality in London, UK, by adopting principal component analysis (PCA) and ensemble learning methods [6].

Unfortunately, the aforementioned methods have unsatisfactory forecast precisions, which are due to their simplification of numerous affecting factors and having excessive hypotheses. ANNs have been widely used in data analysis and mining by scholars, and have been applied in air quality prediction. For example, Russo, Raischel, & Lind optimized an ANN by utilizing random variables to predict air quality [7]. Also according to ANN, Nejadkoorki, & Baroutian and Gennaro et al. built models to predict PM10 pollution levels [8, 9]. Voukantsis et al. also evaluated air pollution ($PM_{10}$ and $PM_{2.5}$) using an ANN based on PCA [10]. Corani made a comparison of various prediction methods (utilizing feed-forward neural networks, pruned neural networks, and lazy learning) and highlighted their individual characteristics and applications [11]. However, using an ANN to predict air pollution refers to numerous influencing factors. The process used to determine a reasonable structure for the network is complex, and network training also takes up a great deal of time.

Although many scholars have carried out plenty of contributions on the prediction of air pollution, there are still some problems that need to be addressed: (1) there are only a few relevant studies about the diffusion trends of the haze, and the correlation between haze and meteorological factors is still not well known; (2) the accuracy of traditional models needs to be more improved (due to the reasons stated above); (3) massive, and multi-source data accumulated in big-data time is not utilized completely. Furthermore, a weak automatic-learning ability leads to low prediction efficiencies. Therefore, in-depth study of DL should provide new ideas and methods for haze prediction.

As data mining (DM) technology is becoming very popular in these years, DL has been widely used in various fields of artificial intelligence. DL also relates to ANN structures [12]. By setting up a multi-implicit layer machine learning (ML) model, the training using massive data was conducted, and useful characteristics and correlations can be derived to enhance classification and prediction accuracies. DL differs from ANNs mainly in the training method employed, that is, layer-by-layer initialization was put into use to greatly reduce training time and improve the real-time prediction performance. There are four reasons for utilizing DL to predict and analyse haze: (1) the rapid development in computer hardware and software provides technological support for realizing DL. The amelioration in the training method thus achieved, promotes real-time performance; (2) the scholars from many different countries in the world have collected and accumulated a large body of data monitored in real-time, e.g. air quality and meteorology data. The age of big data provides a massive and multi-source set of training data for DL and guarantees the accuracy of the predictions made for haze based on DL; (3) a DL based model for predicting haze can mine the correlations and internal relationships between various input factors. Thus, the modes relating to high-level changes and laws governing the haze on a semantic level can be found, and some effective analysis and predictions of polluted weather caused by haze can be accomplished; (4) a DL prediction model belongs to the category of a statistical prediction with strong extendibility. By reasonably setting the input factors, other prediction algorithms can be combined with the model to overcome the defects and uncertainties of a single model and further boost the precision of the haze predictions.

In recent years, there has been an increasing amount of achievements on the application of DM, big-data analysis, and DL in predicting air quality. According to DM and ML technologies, Peng carried out a predictive analysis of air quality to improve the accuracy of prediction [13], whereas, the efficiency of the method needs to be further enhanced. Li, Peng, Hu, Shao & Chi proposed an approach for predicting air quality based on spatiotemporal DL and essentially examined the correlation between the spatial and temporal data [14]. Ahn, Shin, Kim, & Yang took the advantage of sensor data to investigate changes in indoor air quality by adopting a DL method [15]. Li et al. automatically dissected the air pollution process and developed a predictive platform for air quality that came from big-data analysis and recognition based on multidimensional historical pollution processes and the weather situation [16]. By utilizing a DL method, Yin et al. forecasted air pollution levels under a big-data background and transformed the characteristics of the spatial data into semantic characteristics [17]. The performance and accuracy of their air pollution predictions could be built up using automatic learning to acquire hierarchical characteristics. However, the correlation between haze and meteorological factors was not fully taken into account and, therefore, the predictions were relatively simple.

Overall, the existing models primarily predict the developmental trends in the air quality index. Nevertheless, few studies focus on predicting the developmental trends in haze by analyzing the correlations between the haze and meteorological factors under the background of big data. Thus, it is necessary to investigate air quality forecasts using a novel DL technology. Considering the deficiencies in existing prediction models, and the advantages and characteristics of DL, the correlations between the haze and meteorological factors were acquired after evaluating the crucial meteorological factors affecting haze diffusion. On this basis, the prediction precision can be increased by establishing a haze prediction model after studying the trends in the diffusion of the haze. The self-learning and training efficiency of the proposed model is guaranteed further by fully utilizing massive and multi-source big data.

The remainder of this study is organized as follows. Section 3 states the DBN model for predicting haze pollution based on the DL method. Section 4 discusses data processing and expounds some practical cases. Section 5 summarizes our conclusions.

## 3 METHODOLOGY

### 3.1 Overview of the DBN Model for Urban Haze Prediction

DL can realize the approach of a complex function based only on a simple network structure, and show the capacity to intensively learn the essential characteristics of a dataset from plenty of unlabelled sample sets. DL can favourably represent the characteristics of the data and is able to represent large-scale data as models exhibit a deep level and strong expression ability.

A DBN is a deep neural network (DNN) composed of multiple layers of latent variables (implicit units), which is widely applied in various fields such as image classification and speech recognition, and has a strong classification and prediction capacity. In this study, a new DBN based model (DBN-UHP or, more briefly, DBN-H) was put forward to predict urban haze ($PM_{2.5}$), whose structure is illustrated in Fig. 1.

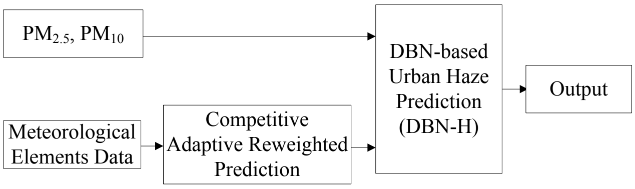

PM2.5, PM10
Meteorological Elements Data
Competitive Adaptive Reweighted Prediction
DBN-based Urban Haze Prediction (DBN-H)
Output

**Figure 1** The structure of the proposed DBN model for urban haze prediction

The features of the proposed model are: (1) data on both the air quality and environment (e.g. meteorological data) were monitored and collected. The correlation between the meteorological elements and haze was studied and, on this basis, the primary factors affecting haze were investigated to further predict the developmental trends in urban haze; (2) in order to search for the optimal network structure, haze forecasts were carried out utilizing a multilayer restricted Boltzmann machine (RBM) and monolayer back propagation (BP) network; (3) the predictions were performed by employing a competitive adaptive-reweighed method after inputting part of the meteorological data into the model to further enhance the prediction accuracy.

Considering that haze is closely correlated with several factors, e.g. historical conditions and weather, eight input characteristics of the samples were adopted. Thus, the historical haze content ($PM_{2.5}$ and $PM_{10}$) and diffusion conditions of the pollutants were comprehensively taken into account in terms of the input data of the model. The diffusion conditions of the pollutants mainly related to meteorological conditions: wind speed, wind direction, temperature, humidity, light, and atmospheric pressure spatially vertical to various layers. In addition, haze predictions were undertaken by indicating future meteorological factors and finding out the correlation between meteorological essentials and haze.

### 3.2 The Proposed Model for Prediction Based on DBN

The proposed model, DBN-H, is a DNN consisting of multiple layers of latent variables (implicit units). The components of the network are recognized to be RBMs and the DBN-H training process is conducted on a layer-by-layer basis. In each layer, a data vector is used to infer an implicit layer, which is considered the data vector for the next (higher) layer. There are connections between layers but none between units of the layers. In order to solve the problem of long training time and searching for the optimal network structure, the model is composed of multilayer RBMs and a single BP layer. An RBM consists of a visual layer and an implicit layer, and the two layers are applied for inputting characteristic data and for learning and extracting the characteristic data, respectively. The overall structure is shown in Fig. 2. The details of the structure depend on the number of nodes and number of input visual layers of the first RBM, the depth (i.e. layers of the RBM network) of the DBN-H network, and quantities of nodes in the various implicit layers. The number of input characteristics adopted for the samples determines the number of visual layers of the first RBM.

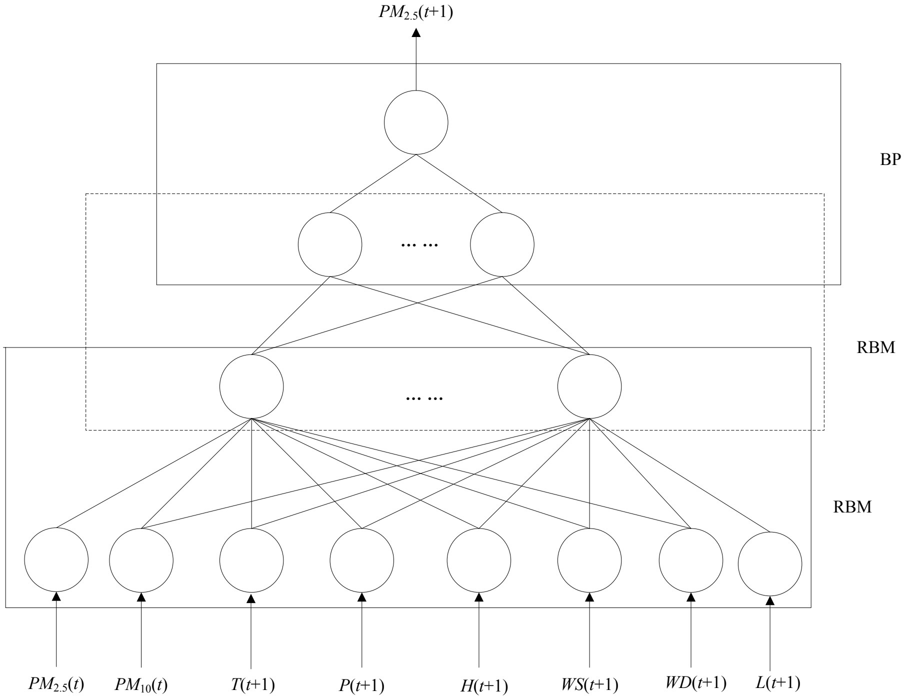


**Figure 2** The structure of the DBN-H model

In a multilayer RBM network, the output characteristics of an upper layer of the RBM network (obtained by studying the data) are taken as the input for the next layer. In this way, each layer can favourably abstract the characteristics of the upper layer to extract the data characteristics layer by layer. The characteristics extracted by the RBM network are regarded as the input to the BP network layer at the top for classification or prediction purposes [18].

The inputs $PM_{2.5}$, $PM_{10}$, $T$, $P$, $H$, $WS$, $WD$, and $L$ refer to the concentrations of $PM_{2.5}$, $PM_{10}$, temperature, atmospheric pressure, humidity, wind speed, wind direction, and light, respectively. Time is represented by $t$, so the output, $PM_{2.5}(t+1)$, denotes the concentration of $PM_{2.5}$ at time $t+1$.

### 3.2.1 The Multilayer RBM Network in DBN-H

The nodes in each layer of the RBM network are independent and can only randomly adopt values of 0 or 1. The total probability distribution satisfies a Boltzmann distribution and the $h$ in the implicit layer can be acquired using the conditional probability $p(h\mid x)$ ; otherwise $x'$ in the visual layer is obtained using $p(x\mid h)$ . If $x'$ is equal to the previous $x$ by adjusting weights, the acquired implicit layer is another representation of the visual layer ($h$ is the states of the neurons in the implicit layer, $x$ is training sample, and $x'$ is the states of the neurons in the visual layer).

The joint distribution of the RBM, subject to a given model parameter ($\theta$), is:

$$p(v,h;\theta)=\exp\left[-E(v,h;\theta)\right]/Z\,, \tag{1}$$

where, $Z=\sum_{v}\sum_{h}\exp\left[-E(v,h;\theta)\right]$ is the normalizing factor and $\mathrm{E}$ the energy function defined as:

$$E(v,h;\theta)=-\sum_{ij}W_{ij}v_ih_j-\sum_{i}b_iv_i-\sum a_jh_j\ , \tag{2}$$

where, $i$ and $j$ denote nodes in the visual ($i$) and implicit ($j$) layers, and $W_{ij}$ is the connection weight between them. Moreover, $b_i$ and $a_j$ represent offsets.

Next, the RBM multilayer network has to be trained. To make this process efficient, Hinton, Osindero, & Teh obtained the necessary parameters using unsupervised learning and further determined the required weights by performing layer-by-layer training [18]. Here, an RBM implicit layer is taken to be the visual layer of the next layer. Using two items as an example, the training algorithm can therefore be outlined as follows:

Step 1: Initialization

1) The training sample $x_1$ is given;
2) The learning rate $\eta$ is set;
3) The offsets ($a$ and $b$) and weight matrix $W$ are initialized.

```
Step 2: Training
FOR $i = 1$ TO $n$ DO
    $P(h_{1i}=1\mid x_1)=\mathrm{sigmoid}\left(\sum_{j}^{m}W_{ij}x_{1j}+c_i\right)$
    sample $h_{1i}\in\{0,1\}$ from $P(h_{1i}=1\mid x_1)$
END FOR
FOR $j = 1$ TO $m$ DO
    $P(x_{2j}=1\mid h_1)=\mathrm{sigmoid}\left(\sum_{i}^{m}W_{ij}h_{1i}+b_j\right)$
    sample $x_{2j}\in\{0,1\}$ from $P(x_{2j}=1\mid h_1)$
END FOR
FOR $i = 1$ TO $n$ DO
    $P(h_{2i}=1\mid x_2)=\mathrm{sigmoid}\left(\sum_{j}^{m}W_{ij}x_{2j}+c_i\right)$
END FOR
$W=W+\eta\left[h_1x_1^{'}-P(h_{2}.=1\mid x_2)x_2^{'}\right]$
$b=b+\eta(x_1-x_2)$
$c=c+\eta\left[h_1-P(h_{2}.=1\mid x_2)\right]$
```

### 3.2.2 The BP Neural Network in DBN-H

In the DBN-H model, the BP network is a multilayer network divided into input, implicit, and output layers, and these layers are totally connected. The output of the anterior layer units cannot feed back to the fore-anterior layer and there are no connections between units in the same layer. The output of the nodes in the implicit layer in the BP network can be expressed as:

$$O_j=f\left(\sum W_{ij}x_i-a_i\right), \tag{3}$$

where, $a_i$ refers to the threshold of the neural cell and $f$ is the activation function (generally taken to be a sigmoid function). The output of the output node can be calculated using:

$$y_k=f\left(\sum T_{jk}o_j-b_k\right), \tag{4}$$

where, $b_k$ represents the threshold of the neural cell and $T_{jk}$ the strength of the connection between nodes in the implicit and output layers.

Input layer:

$$\mathrm{input}(i)\in S,\ S=\{I_1,I_2,\ldots,I_{c1}\}\,, \tag{5}$$

Implicit layer:

$$n=\sqrt{c_1+c_2}+\alpha\,, \tag{6}$$

where, $n$ is the number of neural cells and $\alpha$ is a constant, and Output layer:

$$\sum_{j=1}^{c2}\mathrm{output}(j)\ \ ,\mathrm{output}(j)\in L,L=\{O_1,O_2,\ldots O_{c2}\}\,, \tag{7}$$

The model used to predict air pollution was established using a three-layer (input, implicit, and output) BP network. Various pollutants appear in the input layer. The number of neural cells in the implicit layer can be determined utilizing the appropriate formula while the optimal value was generally calculated according to trial and error. In other words, the number of cells in the implicit layer that yield a model with optimal prediction effect was selected via tests. The output layer displays the corresponding pollution value as determined by the various factors in the input layer. The model is manifested schematically in Fig. 3, in which $I_1$, $I_2$, $I_3$, $Ic$, $Ic_1$, $Ic_2$, and $Ic_3$ refer to the inputs to the neural network and $O$ denotes the output.

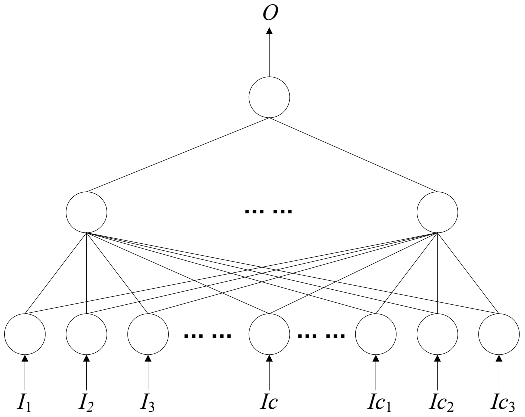


**Figure 3** The structure of the BP neural network in the DBN-H model

The BP algorithm in the DBN-H model can be outlined as follows:

Step 1: The outputs of each layer are acquired according to practical inputs.

Step 2: According to the outputs of each layer, the result $O_k$ of the output layer is calculated, as well as the loss value. The loss function ($LF$) is expressed as:

$$LF(\omega)=\frac{1}{2}\sum_{k=outputs}(t_k-O_k)^2\ , \tag{8}$$

Step 3: The loss value ($LV$) of neural cell $j$ in the implicit layer can be calculated:

$$LV(h_j)=O_j\cdot(1-O_j)\sum W_{ji}O_j\ , \tag{9}$$

Step 4: The weight is then updated:

$$W_{ji}=W_{ji}+\eta O_j X_{ji}\ , \tag{10}$$

where $\eta$ is the learning rate.

### 3.3 The Competitive Adaptive-Reweighed Prediction Method

The data input to the DBN-H model, especially those relating to meteorological elements are closely correlated with the diffusion of the haze pollutants. In this study, the urban haze level was calculated by predicting the future meteorological factors and evaluating the correlation between these factors and haze. Therefore, the accuracy of the input data has an important effect on the final forecasts. By using various algorithms (including multiple regression, ARMA, CART, and NN), a competitive adaptive-reweighed algorithm is presented to predict the input meteorological data (Fig. 4). Tests certified that the prediction accuracy of this procedure is better than that obtained adopting a single algorithm.

The output of the prediction process can be expressed as follows:

$$AQI=\frac{1}{4}\sum_{i=1}^{4}\alpha_i x_i\ , \tag{11}$$

where $\alpha_i$ is the weight assigned to prediction algorithm $i$, so that $\sum_{i=1}^{4}\alpha_i=1$, and $x_i$ is the output from that method.

The outputs from the four prediction algorithms were separately compared with the system output. Weights (i.e. $\alpha_i$) that have a low variance were constantly strengthened as a reward (by adjusting the weights) so that the optimum prediction accuracy can be realized. Learning is continued until the total error ($TE$) of the sample set

$$TE=\frac{1}{2}\sum_{i=1}^{m}(y^i-c^i)^2\ , \tag{12}$$

where $y^i - c^i$ is the absolute error of the system (i.e. difference between expected and actual outputs) and $m$ the number of learning specimens, meets the given precision requirement. That is, until $TE<\varepsilon$, where $\varepsilon$ is the preset precision level. Then, the final adjusted weights need to be recorded.

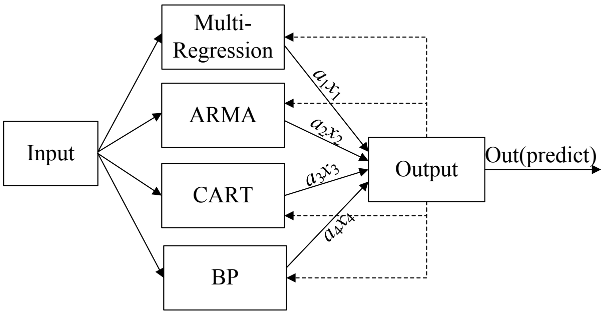


**Figure 4** The model used in the competitive adaptive-reweighed prediction method

The depth of the DBN network has a significant effect on the performance of the model. As the network depth increases, the data-mining ability of the model becomes stronger and the more abstract the feature extraction process. Thus, the better prediction performance of the network can be achieved. However, an excessive number of layers may result in problems due to over-fitting. In contrast, as there is a sufficient amount of information about haze ($PM_{2.5}$) and the meteorological factors that affect haze diffusion, a few layers can provide a good feature mining effect. In the current context, the tests in this study indicate that a three-layer network structure can exhibit a favourable prediction result.

The number of nodes in the implicit layers of the DBN can also have a certain effect on the performance of the model. Using an insufficient number of nodes can result in a weak data mining performance while too many lead to over-fitting problems. In the present case (aimed at providing a model for predicting haze), a contrastive analysis illustrates that a good prediction performance is achieved when the number of nodes in the implicit layer is equal to 100.

### 3.4 Training the DBN-H Model

The weights required in the generative model were obtained under independent and unsupervised conditions by the unsupervised greedy method. The RBM network in each layer adjusted the weights in its own layer to acquire the probability distribution of the training samples further by determining the weights. In this way, the probability of the training samples shows the highest level. The mapping of feature vectors reaches the optimal level when the feature vectors are mapped to different feature spaces. The BP network receives the output feature vectors from the RBMs as the input feature vectors to train the classifier under supervision. The BP network transmits wrong information in a top-down fashion to the RBM in each layer to adjust finely the whole DBN to achieve optimal global mapping.

The training algorithm RBM update($x_1$, $\eta$, $W$, $b$, $c$) of the model can be outlined as follows:

Step 1: Initialization

1) $k = l$ , where $k$ refers to the layer number being trained;

2) $W^k$ =0, the weight matrix in layer $k$;

3) $b^k$ =0, the offset in the visual layer of layer $k$;

4) $c^k$ =0, the offset in the implicit layer of layer $k$.

Step 2: Training

```
FOR $k=1$ TO $l$
    WHILE not stopping criterion DO
        sample $h^0 = x$ from $\hat{S}$ // $\hat{S}$ denotes the training distribution sequence
        FOR $i=1$ TO $k-1$
            IF mean_field_computation THEN
                assign $h_j^i$ to $P(h_j^i = 1 | h^{i-1})$ for all elements $j$ of $h^i$
            ELSE
                sample $h_j^i$ to $P(h_j^i = 1 | h^{i-1})$ for all elements $j$ of $h^i$
            END IF
        END FOR
        RBM update($h^{k-1}, \eta, W^k, b^k, c^k$)
    END WHILE
END FOR
```

## 4 ANALYSING RESULTS AND DISCUSSION

### 4.1 Data Acquisition

The data was obtained in two main ways: (1) by downloading data released by various official websites. These include the daily national urban air quality reports released by the Ministry of Environmental Protection of the People's Republic of China (http://datacenter.mep.gov.cn/index), the National Meteorological Datacentre (http://data.cma.cn/), the China Meteorological Administration (http://www.cma.gov.cn/), and the weather forecasts for Weihai in Shandong province, China. (https://www.tianqi.com/weihai/); (2) by continuously acquiring data during 2016-2017 using a home-made, integrated monitoring station (Fig. 5) to monitor the air quality and meteorological factors at eight monitoring points in Weihai.

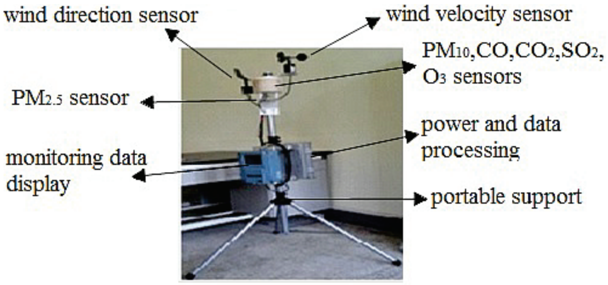
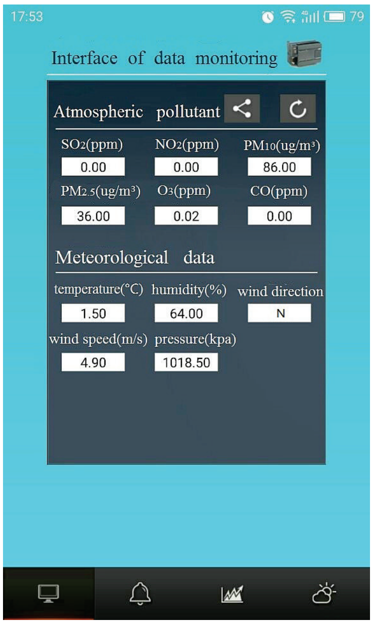


**Figure 5** The integrated monitoring equipment used to determine air quality and meteorological factors



Interface of data monitoring

Atmospheric pollutant

| $SO_2$(ppm) | $NO_2$(ppm) | $PM_{10}$(ug/m³) |
|---|---|---|
| 0.00 | 0.00 | 86.00 |
| $PM_{2.5}$(ug/m³) | $O_3$(ppm) | CO(ppm) |
| 36.00 | 0.02 | 0.00 |

Meteorological data

| temperature(°C) | humidity(%) | wind direction |
|---|---|---|
| 1.50 | 64.00 | N |
| wind speed(m/s) | pressure(kpa) | |
| 4.90 | 1018.50 | |

**Figure 6** Collection and monitoring of data using mobile phones

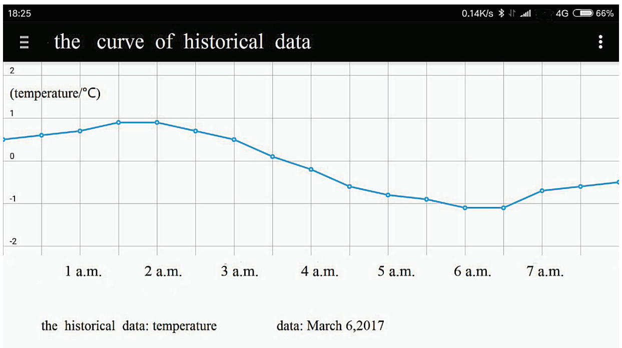


**Figure 7** Historical trends in temperature data recorded using mobile phones

The equipment employed is solar powered and can simultaneously collect data relating to meteorological factors and air quality. It can also obtain data continuously after setting a suitable time interval. By doing so, the consistency of the spatiotemporal relationship between the meteorological and air quality data can be guaranteed. The integrated monitoring equipment can realize continuous and real-time monitoring and effectively store the data on air pollution (e.g. $PM_{2.5}$) and meteorological factors (temperature, humidity, etc.) using ultra-low-power consumption technology. Furthermore, the acquired data can be transmitted to a cloud computing center in real time through the sensor network, Internet, or 3G/4G networks for big-data analysis and air-quality prediction. Additionally, real-time collection and monitoring of data and analysis of trends in the historical data can be carried out on mobile phones, as manifested in Fig. 6 and Fig. 7.

## 4.2 Data Preprocessing

### 4.2.1 Time Registration

The meteorological factors and air quality are all recorded in the form of time series. The complete set of time series data for the same variable may be derived from different heterogeneous data sources and the sampling frequencies used therein may not be consistent with the time base. Moreover, latency problems may appear during the transmission of data. These problems mean that multi-source time series will not be matched in time. Such series can be divided into sub-sequences using translation registration of a sliding window. By comparing the related sub-sequences with the related collected data, the time differences between two time series can be calculated. Then, according to the time difference found, the collected data can be translated (based on the time series of highest frequency) to further realize time registration between the series [19].

### 4.2.2 Data Cleaning

The processed data is mainly composed of meteorological and environmental data. Monitoring data obtained using sensors is subject to various phenomena resulting in some parts of the data missing and some data being redundant. Data cleaning is therefore conducted on the data using a multiple interpolation approach. Additionaly, *N* estimates of each missing value are constructed producing *N* complete data sets. Afterwards, all the complete data sets are processed by adopting the same method to obtain *N* processing results. Finally, by analyzing the *N* processed results, the final estimates of the target variables can be acquired thus accomplishing the cleaning of the haze data.

## 4.3 Evaluating the Predictive Ability of the DBN-H Model

In order to evaluate the predictive ability of the DBN-H model, correlation coefficients (*Corr*) and mean absolute error (*MAE*) are regarded as evaluation indices. The correlation coefficient reflects the correlation between the practical and predicted values of the haze concentration, and is calculated according to:

$$Corr = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{R_i - \overline{R}}{\sigma_R} \times \frac{P_i - \overline{P}}{\sigma_P}\right), \tag{13}$$

where $n$ is the number of prediction samples, $R_i$ the practical haze concentration, and $P_i$ the predicted haze concentration. Moreover, $\overline{R}$, $\sigma_R$, $\overline{P}$, and $\sigma_P$ refer to the mean value and standard deviation of the practical haze concentration $R_i$ and those of the predicted haze concentration $P_i$, respectively.

*MAE* is considered to reflect the average error between the actual and predicted values of the haze, that is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| R_i - P_i |, \tag{14}$$

In order to verify the accuracy of the proposed model, a comparison is made between DBN-H and the traditional models, including MLRM, ARMA, CART and NN, in terms of correlation coefficients and MAEs. The results are demonstated in Tab. 1 and Fig. 8, where MLRM stands for the multivariable linear regression model.

**Table 1** A comparison of the accuracy of various prediction models

| Term | DBN-H | MLRM | ARMA | CART | NN |
|---|---|---|---|---|---|
| *Corr* | 0,767 | 0,701 | 0,213 | 0,687 | 0,745 |
| *MAE* $\mu g/m^3$ | 26,521 | 30,471 | 75,618 | 33,176 | 29,547 |

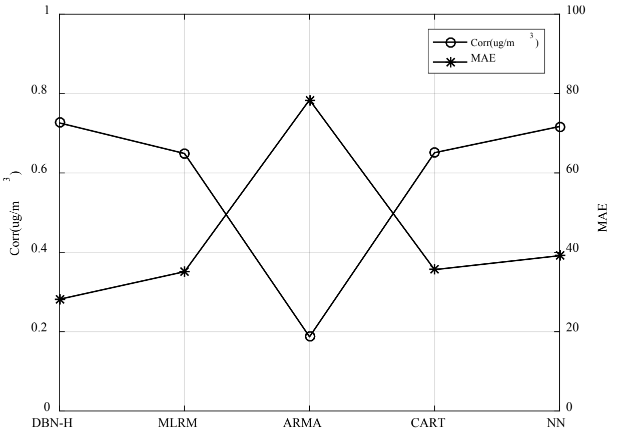


**Figure 8** A comparison of the accuracy of various prediction models in terms of correlation coefficient and *MAE*

It can be seen from Tab. 1 and Fig. 8 that the DBN-H is superior to the other prediction algorithms in respect of correlation coefficient and MAE. The results indicate that the correlation coefficient and MAE between the results predicted using the DBN-H and the actual results are 0.767 and 26.5 $\mu g/m^3$, respectively. The DBN-H exhibits relatively favorable prediction results in terms of correlation coefficient and *MAE* and can deeply mine the correlations between haze and meteorological factors. Besides, the model can determine the primary meteorological factors affecting haze, and effectively decrease the noise in the numerical values after a stable prediction model has been formed via training, to further increase the prediction accuracy.

Haze pollution is closely related to season. In winter, people need to heat their home more and so, compared with other times of the year, haze pollution is more serious. In order to further corroborate the prediction performance of the DBN-H under different environmental conditions, predictions were carried out separately utilizing haze data taken from the winter heating period (December - February) and the summer (June - August). The results are shown in Tab. 2 and Tab. 3, and Fig. 9 and Fig. 10.

**Table 2** A comparison of the performance of the various prediction models using data recorded in the winter heating period (December-February)

| Term | DBN-H | MLRM | ARMA | CART | NN |
|---|---|---|---|---|---|
| *Corr* | 0,726 | 0,649 | 0,187 | 0,651 | 0,717 |
| *MAE* μg/m$^3$ | 28,154 | 35,107 | 78,416 | 35,619 | 39,141 |

**Table 3** A comparison of the performance of the various prediction models using data recorded in the summer (June-August)

| Term | DBN-H | MLRM | ARMA | CART | NN |
|---|---|---|---|---|---|
| *Corr* | 0,816 | 0,712 | 0,207 | 0,705 | 0,750 |
| *MAE* μg/m$^3$ | 24,315 | 40,121 | 70,155 | 30,006 | 27,153 |

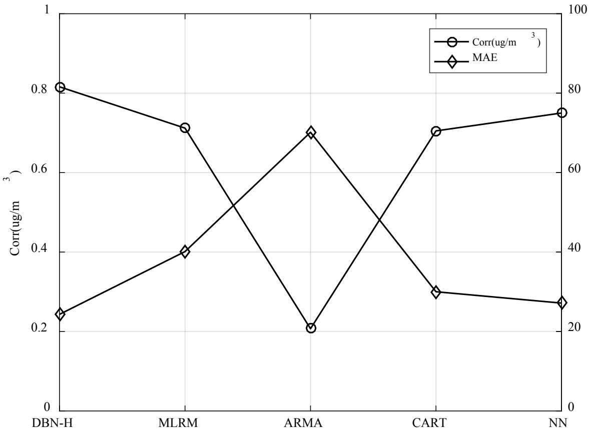


**Figure 9** A comparison of the various prediction models in terms of correlation coefficient and MAE during the winter heating period (December-February)

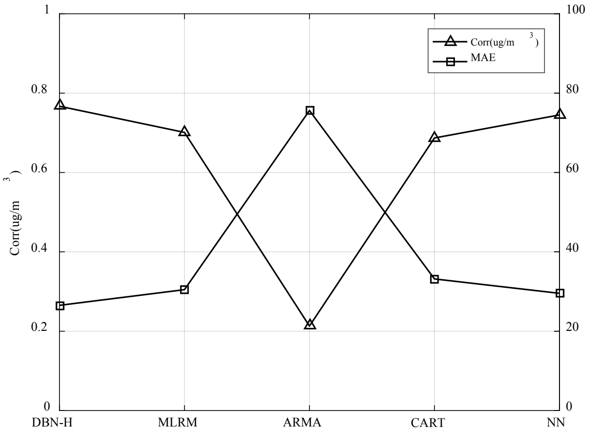


**Figure 10** A comparison of the various prediction models in terms of correlation coefficient and MAE in the summer (June-August)

The results indicate that the haze prediction in summer is better than the prediction in winter. The correlation coefficient of the haze forecast made by the DBN-H falls from 0.767 to 0.726 in the winter heating period (decreasing by 4.1 %). In contrast, the correlation coefficient in summer was 0.816 (increasing by 8.9 %). This is because the factors causing the haze pollution in summer were reduced, but the pollution was more serious during the heating period in winter and that the factors influencing the concentration of haze pollution are slightly increased. These resulted in a slight influence in accuracy of the haze pollution predictions. Overall, the proposed DBN-H model is better than the traditional prediction models in terms of prediction accuracy.

## 5 CONCLUSION

To enhance the accuracy of haze pollution predictions, a novel DBN based model was proposed to better analyze the correlations between the relevant meteorological factors and haze pollution and the diffusion laws of the haze. The following conclusions could be drawn:

(1) A DL method used for predicting and evaluating the haze was applied to investigate the correlation and internal relationships between the meteorological factors and haze. The results indicate that wind direction, wind speed, and season have the most significant influence on the changes occurring to the haze. As the factors affecting haze increase in the winter heating period, the prediction accuracy in winter is inferior to that in summer.

(2) Compared to other prediction models, the DBN-H model gives a correlation result that is 18 % better, while the MAE declines by 15.7 μg/m$^3$.

(3) The accuracy of the input data significantly affects the prediction results. Missing values can be dealt with at the preprocessing stage using time registration. The input data on the meteorological factors can be predicted utilizing a competitive adaptive-reweighed prediction method, which can further increase the accuracy of the haze predictions.

(4) In the case of sufficient information available on the haze ($PM_{2.5}$) and meteorological elements, favorable prediction performance can be guaranteed using a three-layer network structure and 100 nodes in the implicit layer of the prediction model based on the DL technology.

Thus, a comprehensive analysis of the internal correlation between meteorological factors and haze based on big data, and a competitive adaptive-reweighed prediction method can reflect the change law of haze more authentically, and then can provide scientific support for urban haze prevention and governance. However, the amount of data on air quality and meteorological factors collected at the same time and in the same place is still not perfectly sufficient. This problem makes a limitation to the prediction effect of the proposed method. Therefore, how to take more advantage of the various sets of environmental data and establish a perfect integrated acquisition network for collecting haze and meteorological data should be considered in future studies.

### Acknowledgements

This work was supported by the China Scholarship Council, the Foundation for Key Sci & Tech Commission Research Project of Jilin Province, China (20150204005GX), Key Sci & Tech Commission Research Project of Changchun City, China (14KG082), and the National Natural Science Foundation of China (61472049).

Special thanks are given to Prof. Jianlin Cheng, who invited Dr. Huimin Lu to do research as a visiting scholar in his lab at the Department of Electrical Engineering and Computer Science, University of Missouri-Columbia, USA.

## 5 REFERENCES

[1] Rohde, R. A. & Muller, R. A. (2015). Air pollution in China: mapping of concentrations and sources. *Plos One*, *10*(8), 1-15. https://doi.org/10.1371/journal.pone.0135749

[2] Saeed, S., Hussain, L., Awan, I. A., & Idris, A. (2017). Comparative analysis of different statistical methods for prediction of $PM_{2.5}$ and $PM_{10}$ concentrations in advance for several hours. *IJCSNS International Journal of Computer Science and Network Security*, *17*(11), 45-52.

[3] Singh, K. P., Gupta, S., Kumar, A., & Shukla, S. P. (2012). Linear and nonlinear modeling approaches for urban air quality prediction. *Science of the Total Environment*, *426*(1), 244-255. https://doi.org/10.1016/j.scitotenv.2012.03.076

[4] Li, X., Peng, L., Shao, J., Cui, S. L., & Tian, H. F. (2016). Air pollution forecast based on wavelet decomposition and ARMA model. *Environmental Engineering*, *34*(8), 110-113.

[5] Chen, F. (2016). Learning of index of air quality prediction model based on CART. *Chinese Journal of Shangrao Normal University*, *36*(6), 16-21.

[6] Singh, K. P., Gupta, S., & Rai, P. (2013). Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment*, *80*(1), 426-437. https://doi.org/10.1016/j.atmosenv.2013.08.023

[7] Russo, A., Raischel, F., & Lind, P. G. (2013). Air quality prediction using optimal neural networks with stochastic variables. *Atmospheric Environment*, *79*(7), 822-830.

[8] Nejadkoorki, F. & Baroutian, S. (2012). Forecasting extreme $PM_{10}$ concentrations using artificial neural networks. *International Journal of Environmental Research*, *6*(1), 277-284. https://doi.org/10.1016/j.atmosenv.2013.07.072

[9] Gennaro, G. D., Trizio, L., Gilio, A. D., Pey, J., Pérez, N., Cusack, M., Alastuey, A., & Querol, X. (2013). Neural network model for the prediction of $PM_{10}$ daily concentrations in two sites in the western Mediterranean. *Science of the Total Environment*, 463-464, 875-883. https://doi.org/10.1016/j.scitotenv.2013.06.093

[10] Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., & Kolehmainen, M. (2011). Intercomparison of air quality data using principal component analysis, and forecasting of $PM_{10}$ and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki. *Science of the Total Environment*, *409*(7), 1266-1276. https://doi.org/10.1016/j.scitotenv.2010.12.039

[11] Corani, G. (2005). Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. *Ecological Modelling*, *185*(2-4), 513-529. https://doi.org/10.1016/j.ecolmodel.2005.01.008

[12] Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks*, 61, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

[13] Peng, H. P. (2015). Air quality prediction by machine learning methods. Ph.D. Thesis of University of British Columbia, Canada.

[14] Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research International*, *23*(22), 22408-22417. https://doi.org/10.1007/s11356-016-7812-9

[15] Ahn, J., Shin, D., Kim, K., & Yang, J. (2017). Indoor air quality analysis using deep learning with sensor data. *Sensors*, *17*(2476), 1-13. https://doi.org/10.3390/s17112476

[16] Li, Y. T., Yan, J. H., Sun, F., Zhang, D. W., Xia, X., Rui, X. G., Bai, X. X., & Yin, W. J. (2017). Air quality forecasting platform based on big data analytics & cognitive technology. *Chinese Journal of Environmental Management*, *9*(2), 31-36.

[17] Yin, W. J., Zhang, D. W., Yan, J. H., Zhang, C., Li, Y. T., & Rui, X. G. (2015). Deep learning based air pollutant forecasting with big data. *Chinese Journal of Environmental Management*, *7*(6), 46-52.

[18] Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527-1554. https://doi.org/10.1162/neco.2006.18.7.1527

[19] Lu, H. M., Di, T. Y., Song, J. J, Sun, H. Y., Han, X. M., & Liu, G. (2016). Translation registration algorithm for multi-source time series data based on the sliding window. *Journal of Engineering Science and Technology*, *9*(5), 44-50.

**Contact information:**

**Huimin LU,** Post Doctor, Full professor
(Corresponding author)
School of Computer Science and Engineering,
Changchun University of Technology, Old Library,
No. 2055 Yan'an Road, Changchun City, Jilin Province, 130012 China
E-mail: luhm.cc@gmail.com

**Jingjing SONG,** Postgraduate Student
School of Computer Science and Engineering,
Changchun University of Technology, Old Library,
No. 2055 Yan'an Road, Changchun City, Jilin Province, 130012 China
E-mail: 201602032@stu.ccut.edu.cn

**Tianyi DI,** Postgraduate Student
School of Computer Science and Engineering,
Changchun University of Technology, Old Library,
No. 2055 Yan'an Road, Changchun City, Jilin Province, 130012 China
E-mail: 201602042@stu.ccut.edu.cn

**Jamshid MORADI KURDESTANY,** Post Doctor
Department of Radiation Physics,
The University of Texas MD-Anderson Cancer Center,
1400 Pressler Street, Unit 1420, FCT8.6103, Houston, Texas, 77030 USA
E-mail: JMoradi@mdanderson.org

**Hongzhi WANG, PhD, Full professor**
School of Computer Science and Engineering,
Changchun University of Technology, Old Library,
No. 2055 Yan'an Road, Changchun City, Jilin Province, 130012 China
E-mail: wanghongzhi@ccut.edu.cn