Estimating Ground-Level PM$_{2.5}$ by Fusing Satellite and Station Observations: A Geo-Intelligent Deep Learning Approach

Tongwen Li$^1$, Huanfeng Shen$^{1,2,3}$, Qiangqiang Yuan$^{2,4}$, Xuechen Zhang$^1$, and Liangpei Zhang$^{2,5}$

$^1$School of Resource and Environmental Sciences, Wuhan University, Wuhan, China, $^2$The Collaborative Innovation Center for Geospatial Technology, Wuhan, China, $^3$The Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University, Wuhan, China, $^4$School of Geodesy and Geomatics, Wuhan University, Wuhan, China, $^5$The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China

Abstract Fusing satellite observations and station measurements to estimate ground-level PM$_{2.5}$ is promising for monitoring PM$_{2.5}$ pollution. A geo-intelligent approach, which incorporates geographical correlation into an intelligent deep learning architecture, is developed to estimate PM$_{2.5}$. Specifically, it considers geographical distance and spatiotemporally correlated PM$_{2.5}$ in a deep belief network (denoted as Geoi-DBN). Geoi-DBN can capture the essential features associated with PM$_{2.5}$ from latent factors. It was trained and tested with data from China in 2015. The results show that Geoi-DBN performs significantly better than the traditional neural network. The out-of-sample cross-validation $R^2$ increases from 0.42 to 0.88, and RMSE decreases from 29.96 to 13.03 $\mu$g/m$^3$. On the basis of the derived PM$_{2.5}$ distribution, it is predicted that over 80% of the Chinese population live in areas with an annual mean PM$_{2.5}$ of greater than 35 $\mu$g/m$^3$. This study provides a new perspective for air pollution monitoring in large geographic regions.

1. Introduction

PM$_{2.5}$ (particulate matter), with an aerodynamic diameter of less than 2.5 $\mu$m, is associated with many adverse health effects, such as respiratory problems and cardiovascular disease (Bartell et al., 2013; Brauer et al., 2012; Crouse et al., 2012). Previous studies (Chen et al., 2013) have shown that a 3 year reduction in average life expectancy and a 14% increase in overall mortality would result from a 100 $\mu$g/m$^3$ increase in the concentration of respirable particulate matter. As a result, PM$_{2.5}$ pollution has attracted widespread attention in recent years and has become the focal point of international air pollution research (Engel-Cox et al., 2013).

With the launch of satellites and the continuous improvements in data retrieval technology, estimating ground-level PM$_{2.5}$ using satellite remote sensing has become a promising approach for the monitoring of PM$_{2.5}$ pollution (Engel-Cox, Hoff, et al., 2004; Hoff & Christopher, 2009; Liu, Chen, et al., 2009; Martin, 2008). Three main kinds of methods have been applied to estimate PM$_{2.5}$ using satellite-derived aerosol optical depth (AOD); chemical simulation models (Liu et al., 2004; van Donkelaar et al., 2010), statistical models (Liu, Paciorek, & et al., 2009; Song et al., 2014), and semiempirical models (Lin et al., 2015). Among them, the statistical models are much easier to implement and can obtain a competitive accuracy in PM$_{2.5}$ estimation (Liu, 2014). As a result, many different statistical models have been developed to explore the quantitative relationship between satellite-derived AOD and ground-measured PM$_{2.5}$ (the AOD-PM$_{2.5}$ relationship) (Liu et al., 2007, 2008). For example, the linear regression model establishes a simple linear relationship between AOD and PM$_{2.5}$ (Zhang et al., 2009). Considering more meteorological parameters, the multiple linear regression model was developed by Gupta and Christopher (2009b). To account for the spatial heterogeneity of the AOD-PM$_{2.5}$ relationship, a geographically weighted regression model was introduced (Hu et al., 2013). Moreover, some more complex mixed-effect models (Lee et al., 2011) and generalized additive mixed models (Kloog et al., 2011; Liu et al., 2012) have also been developed to estimate ground-level PM$_{2.5}$. All these statistical models are used to represent the relationship between PM$_{2.5}$ and the latent factors.

However, the levels of PM$_{2.5}$ concentration are related to many factors (Hystad et al., 2011; Paciorek et al., 2008; van Donkelaar et al., 2006; Weber et al., 2010), such as meteorological conditions (e.g., temperature, wind speed, and relative humidity), land use type, population, road networks, and so on. This situation has increased the difficulty of using the traditional statistical models to estimate PM$_{2.5}$. Unlike the traditional methods, nonlinear and nonparametric machine learning algorithms may have the potential to address...
this problem (Lary et al., 2014). For example, Gupta and Christopher (2009a) used a back-propagation neural network (BPNN) to estimate surface-level PM$_{2.5}$ in the southeastern United States; an artificial neural network algorithm was trained with Bayesian regularization to estimate PM$_{2.5}$ in eastern China (Wu et al., 2012), and a generalized regression neural network (GRNN) was reported to outperform the traditional models at national scale in China (Li et al., 2017). These neural network models show great advantages in estimating ground-level PM$_{2.5}$.

Deep learning, which is considered to be the second generation of neural network, may be a potential way to address this situation (Hinton & Salakhutdinov, 2006). However, to date, deep learning has seldom been applied in the estimation of ground-level PM$_{2.5}$; only a few attempts (Ong et al., 2015) have been made to predict time series PM$_{2.5}$ over monitoring stations. On the other hand, intelligent algorithms are commonly used to describe numerical relationships, but they neglect the geographical correlation of environmental variables. Meanwhile, it has been reported that PM$_{2.5}$ concentrations show significant autocorrelation in time and space (Wu et al., 2015). The nearby PM$_{2.5}$ from neighboring stations and the PM$_{2.5}$ observations from nearby days for the same station are informative for estimating PM$_{2.5}$. It is therefore important to incorporate this geographical correlation relationship into the intelligent algorithms.

Consequently, the objective of this study is to develop a geo-intelligent deep learning model to estimate ground-level PM$_{2.5}$. This model is established and evaluated based on satellite observations, meteorological parameters, and ground-level PM$_{2.5}$ measurements from China, which is suffering from serious PM$_{2.5}$ pollution (Peng et al., 2016). This study will provide a new perspective to investigate the spatiotemporal characteristics of air pollution in a large geographic region.

2. Materials and Methods

2.1. Study Region and Data

The study region is China (see Figure S1 in the supporting information). The study period is from 1 January 2015 to 31 December 2015.

The data include five main parts. (1) Ground-level PM$_{2.5}$: hourly PM$_{2.5}$ for 2015 was obtained from the China National Environmental Monitoring Center (CNEMC) Web site (http://www.cnemc.cn). The number of stations was ~1,500 by the end of 2015. We averaged hourly PM$_{2.5}$ to daily mean PM$_{2.5}$ for the estimation of PM$_{2.5}$. (2) Moderate Resolution Imaging Spectroradiometer (MODIS) AOD: both Terra and Aqua MODIS AOD products were downloaded from the Level 1 and Atmosphere Archive and Distribution System (LAADS, https://lads-web.modaps.eosdis.nasa.gov/). We adopted Collection 6 10 km AOD products, which are retrieved by combining dark target and deep blue algorithms (Levy et al., 2013). The average of the Terra and Aqua AOD products was employed to estimate daily average PM$_{2.5}$. (3) Meteorological parameters: we extracted relative humidity (RH, %), air temperature at a 2 m height (TMP, K), wind speed at 10 m above ground (WS, m/s), surface pressure (PS, kPa), and planetary boundary layer height (PBL, m) from MERRA-2 meteorological reanalysis data, which were downloaded from the NASA Web site (http://gmao.gsfc.nasa.gov/GMAO_products/). (4) MODIS normalized difference vegetation index (NDVI): MODIS NDVI products (MOD13) were downloaded from the LAADS Web site. (5) Population and road data: these data were obtained from the Socioeconomic Data and Applications Center (SEDAC, http://sedac.ciesin.columbia.edu/data/collection/gpw-v4) and the National Geomatics Center of China (NGCC, http://ngcc.sbsm.gov.cn/), respectively.

Additional information and integration of the input data are provided in Text S1 in the supporting information.

2.2. The Deep Belief Network (DBN) Model for the Estimation of PM$_{2.5}$

The deep belief network (DBN) model, which is one of the most typical deep learning models, was introduced in 2006 (Hinton et al., 2006). A DBN consists of multiple restricted Boltzmann machine (RBM) layers and a back-propagation (BP) layer and can be used for classification or prediction problems. For example, the structure of a DBN with two RBM layers is shown in Figure 1a.

An RBM consists of a visible layer and a hidden layer, where the hidden layer of the prior RBM is the visible layer of the next RBM. Taking the first RBM as an example, from the visible layer ($v$) to the hidden layer ($h^1$),
$h_i^1 = \begin{cases} 
1, f(W_0i x + b_i) \geq \mu \\
0, f(W_0i x + b_i) < \mu 
\end{cases}$

(1)

where $i$ refers to the number of the $i$th neuron and $b_i$ denotes the bias for neuron $i$. $f(\cdot)$ indicates the transfer function, $f(x) = \frac{1}{1 + e^{-x}}$. It is the same for the calculation of the visible layer from the hidden layer. The contrastive divergence algorithm is usually used for training an RBM (Hinton & Salakhutdinov, 2006). The weights are updated in the $n$th iteration as

$W_0^{n+1} = W_0^n + \varepsilon \left( (h_1^n)^T x - (h_2^n)^T v_1 \right)$

(2)

where $\varepsilon$ is the learning rate, $v_1$ denotes the reconstruction from hidden layer ($h_1^n$), and $h_1^n, h_2^n$ are generated from $x, v_1$ using equation (1), respectively. The RBMs are pretrained one by one, without supervision, and the trained weights are used to initialize the multilayer neural network. The DBN model then works as a feedforward neural network (Yue et al., 2017), whereas the error reduction using the BP algorithm is referred to here as “fine-tuning.”

Specifically, in our case, the schematics of the geo-intelligent DBN (Geoi-DBN) used to estimate ground-level PM$_{2.5}$ are presented in Figure 1b. The input variables are the satellite-derived AOD, meteorological parameters, NDVI, and spatiotemporally informative terms. Because of the autocorrelation, the nearby $n$ grids of PM$_{2.5}$ measurements and the PM$_{2.5}$ observations from the $m$ prior days for the same grid are informative for estimating PM$_{2.5}$. The nearer observations are more informative than further ones (Tobler, 1970; Yuan et al., 2012). For a specific grid, the spatiotemporally informative terms are represented as...
where $d_s, d_t$ refer to the spatial and temporal distances, respectively. $m, n$ are 3 and 10, respectively (Table S1 in the supporting information). The geographical distance ($DIS$) is used to reflect the heterogeneity of uneven station distribution. Two hidden layers (two RBMs) are then used, and the number of neurons in each hidden layer is 15 (Text S2). The RBM layers are stacked one by one to transfer the input signals to the higher layer. The output layer is a BP layer, which has only one node (PM$_{2.5}$ measurements).

The relationship PM$_{2.5} = f$(AOD, RH, WS, TMP, PBL, PS, NDVI, S-PM$_{2.5}$, T-PM$_{2.5}$, DIS) is wished to learn from the data records. The process can be divided into three steps:

1. Pretraining. Using the collected data records, the RBMs are trained layer by layer, without supervision. This unsupervised training can extract the essential features associated with PM$_{2.5}$, and they are transferred from the prior RBM to the next RBM layer. Therefore, the higher layer can extract the deeper features related to PM$_{2.5}$.

2. Fine-tuning. Through the prior pretraining step, the initial weights of Geo-i-DBN are generated and we can obtain the calculated PM$_{2.5}$. Compared with in situ PM$_{2.5}$ measurements, an estimation error can be obtained, and it is sent back to the Geo-i-DBN model to fine-tune the weight coefficients using the BP algorithm.

3. Prediction. This step evaluates the performance of the Geo-i-DBN model established on the input data records and predicts the PM$_{2.5}$ values for those locations with no ground stations. Thus, spatially continuous PM$_{2.5}$ data can be reconstructed.

To evaluate the model performance, a sample-based 10-fold cross-validation technique (Rodriguez et al., 2010) was applied to test the model overfitting and predictive power. Furthermore, spatial hold-out cross-validation strategies, that is, site-based cross-validation and leave one province out cross-validation, were used to assess the spatial performance. Details of the model validation are provided in Text S3. We adopted the statistical indicators (Text S4) of the coefficient of determination ($R^2$), the root-mean-square error (RMSE, $\mu$g/m$^3$), the mean prediction error (MPE, $\mu$g/m$^3$), and the relative prediction error (RPE, %) to evaluate the model performance.

### 3. Results and Discussion

#### 3.1. Evaluation of the Model Performance

To evaluate the deep learning model, we compared the DBN model with BPNN (Gupta & Christopher, 2009a) and GRNN (Li et al., 2017). BPNN, GRNN, and DBN are all data-driven learning models. However, BPNN has a relatively simple structure with three layers; GRNN uses plenty of neurons in the hidden layer to fully approximate functions, and DBN has a more complicated structure and initializes itself by the unsupervised pretraining. As presented in Table 1, when the geographical correlation is not incorporated into the models (original models), Ori-GRNN obtains the best performance (sample-based and site-based cross-validation $R^2$ of 0.60 and 0.58, respectively), while Ori-BPNN performs the worst. The Ori-DBN model, which has a more complicated structure, is expected to achieve a better performance; however, it obtains a relatively poor result. A possible reason could be that the AOD-PM$_{2.5}$ relationship is not complicated enough for deep learning. From the original models to the geo-intelligent models, the performance is greatly improved. Among these models, the Geo-i-DBN model performs the best, whereas the Geo-i-GRNN model performs the worst. This differs from the results of the original models. A possible reason is that the spatiotemporally informative
terms greatly increase the complexity of the AOD-PM$_{2.5}$ relationship. Benefiting from layer-by-layer pretraining, the more complicated relationship between PM$_{2.5}$ and the predictors is better learned in Geoi-DBN than in Geoi-GRNN. Therefore, the Geoi-DBN model achieves the best performance, with out-of-sample cross-validation $R^2$ and RMSE values of 0.88 and 13.03 $\mu$g/m$^3$, respectively. Furthermore, the spatial hold-out cross-validation results also show that Geoi-DBN has a satisfactory spatial predictive power (Tables 1 and S3). These results demonstrate that the Geoi-DBN model is a promising approach for the estimation of PM$_{2.5}$.

Figure 2 shows the scatterplots for the cross-validation of the Ori-DBN and Geoi-DBN models. For the Ori-DBN model, the sample-based cross-validation $R^2$ and RMSE values are 0.54 and 25.86 $\mu$g/m$^3$, respectively. The $R^2$ and RMSE values of the model fitting are 0.59 and 24.53 $\mu$g/m$^3$ (Table S4), respectively. When considering the geographical correlation, the spatiotemporal characteristics of atmospheric PM$_{2.5}$ are better described in the AOD-PM$_{2.5}$ modeling. The model performance is therefore significantly improved. On the other hand, the out-of-sample cross-validation slope for the Geoi-DBN model is 0.88, with an intercept of 6.39 $\mu$g/m$^3$. These findings indicate that the Geoi-DBN model tends to underestimate when the ground-level PM$_{2.5}$ is greater than ~60 $\mu$g/m$^3$. Therefore, the higher PM$_{2.5}$ concentrations may not be sufficiently explained. This issue is further discussed in section 3.4. However, it should be noted that the cross-validation slope of the Geoi-DBN model (0.88) is much greater than that of the Ori-DBN model (0.55). This means that the Geoi-DBN model shows a much lower extent of underestimation than the Ori-DBN model.

The seasonal and spatial performances of the Geoi-DBN model were also evaluated. The out-of-sample $R^2$ values are 0.84, 0.82, 0.87, and 0.90 for spring, summer, autumn, and winter, respectively. The spatial

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample-based cross-validation</th>
<th>Site-based cross-validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>RMSE</td>
</tr>
<tr>
<td>Ori-BPNN</td>
<td>0.42</td>
<td>29.96</td>
</tr>
<tr>
<td>Ori-GRNN</td>
<td>0.60</td>
<td>24.22</td>
</tr>
<tr>
<td>Ori-DBN</td>
<td>0.54</td>
<td>25.86</td>
</tr>
<tr>
<td>Geoi-BPNN</td>
<td>0.84</td>
<td>15.23</td>
</tr>
<tr>
<td>Geoi-GRNN</td>
<td>0.82</td>
<td>16.93</td>
</tr>
<tr>
<td>Geoi-DBN</td>
<td>0.88</td>
<td>13.03</td>
</tr>
</tbody>
</table>

Figure 2. Scatter plots of the cross-validation results: (a and c) sample-based cross-validation results of the Ori-DBN and Geoi-DBN models; (b and d) site-based cross-validation results of the Ori-DBN and Geoi-DBN models.
distribution of R² (Figure S3) shows that higher values of R² are found over the areas (e.g., East China) with denser monitoring stations.

3.2. Mapping of PM$_{2.5}$ Concentrations

The annual and seasonal mean distributions of PM$_{2.5}$ in China are mapped in Figures 3a and S4, using our previous mapping strategy (Li et al., 2017). The annual distribution has a similar spatial distribution to the ground-measured PM$_{2.5}$, which is presented in Figure 3b. We also compared our results with previous satellite-derived PM$_{2.5}$ data (van Donkelaar et al., 2016). The results are shown in Figure S5. In our results, the levels of PM$_{2.5}$ are higher in the northern regions than in the southern regions. Meanwhile, a heavily polluted region is identified in the North China Plain. As reported in a previous study (Chen et al., 2008), the climate of this region is characterized by stagnant weather, with weak wind and a relatively low boundary layer height, which results in the atmospheric conditions for the accumulation, formation, and processing of aerosols. This is one of the main reasons for the serious PM$_{2.5}$ pollution in this area. Additionally, the PM$_{2.5}$ concentrations are generally higher in the inland regions (e.g., Hunan, Hubei, and Hunan) and lower in the coastal regions (e.g., Guangdong and Fujian). The regions with the least PM$_{2.5}$ pollution are in Hainan and Yunnan provinces, which benefit from the low levels of anthropogenic emissions and favorable meteorological conditions for atmospheric dispersion. Last but not least, a very high level of PM$_{2.5}$ pollution is found in the northwest, especially the Xinjiang Autonomous Region. A possible reason is that the dust particles in this desert region make a significant contribution to the accumulation of PM$_{2.5}$ (Fang et al., 2016).

3.3. Exposure Analysis Over China

The distribution of population in China in 2015 is presented in Figure S6. The World Health Organization (WHO) interim target (IT)-1 and IT-3 for annual mean PM$_{2.5}$ are 35 and 15 μg/m$^3$, respectively (WHO, 2006). As shown in Figure 4, the population-weighted estimated annual mean PM$_{2.5}$ is 53.34 μg/m$^3$, which greatly exceeds the WHO IT-1 standard. Almost all regions of China (except for the northwest) show population-weighted averages that are greater than the spatial averages. These findings indicate that more people are living in more polluted regions. Figure 4 also shows that over 80% of the Chinese population live in areas that exceed the WHO IT-1 standard. Spatially, South China has the highest percentage of population living in areas meeting the WHO IT-1 standard, whereas Central and North China have the lowest.

3.4. Discussion

PM$_{2.5}$ assessment by fusing satellite and station observations involves lots of different factors, which inherently results in big data. In this situation, a deep learning model may better estimate PM$_{2.5}$. Therefore, we made efforts to incorporate urban big data (i.e., population data and road network) into the deep learning model. The results show that these predictors have almost no positive (or even passive) effect on model performance (Table S5). The possible reason could be that these predictors are not real time and cannot reflect the temporal variation of the AOD-PM$_{2.5}$ relationship. Moreover, they have too coarse spatial resolutions to provide additional information. It is possible that the model performance would be greatly improved if fine-scale and real-time data (e.g., daily street-level traffic flow) could be obtained.
The Geo-iDBN cross-validation slope of observed PM$_{2.5}$ versus estimation is 0.88, indicating some evidence of bias. However, it should be noted that the national-scale estimates of PM$_{2.5}$ in China (Fang et al., 2016; Lin et al., 2015; Ma et al., 2016; You et al., 2016) are mostly underestimated (slopes of 0.79–0.83) when ground-level PM$_{2.5}$ is greater than 60 μg/m$^3$. This underestimation may be down to several reasons, including the possibility of mixed types and layers of aerosols in the atmosphere (Guo et al., 2016) and the hygroscopicity of urban aerosols (Gupta & Christopher, 2009a). Another possible reason is that we used point-based monitoring data and a spatially averaged modeling framework. The sampling distribution of monitors may not give a great estimation of the spatially averaged concentration for that grid. Therefore, this underestimation is probably a systematic error related to the complicated aerosols in China and the modeling framework.

We summarize the previous studies using machine learning for PM$_{2.5}$ in China in Table S6. Compared with these studies, this study has three main advantages. First, we estimate the spatial PM$_{2.5}$ concentrations using machine learning by incorporating satellite data, while previous study predicted PM$_{2.5}$ over stations (Li et al., 2016). Differing from the traditional machine learning methods (Yao & Lu, 2014), a more advanced deep learning architecture is then utilized for the satellite-based PM$_{2.5}$ estimation. Lastly, while previous studies have used machine learning to simulate the AOD-PM$_{2.5}$ relationship (Wu et al., 2012), this study further considers the geographical correlation to greatly improve model performance.

In this study, a deep learning architecture has been established to estimate ground-level PM$_{2.5}$, achieving a satisfactory performance. However, it should be noted that we applied only one type of deep learning model (i.e., DBN) to model the AOD-PM$_{2.5}$ relationship. Would any other deep learning model work better with this problem? Deep learning has more hidden layers to better represent complex nonlinear relationships, so whether or not we can estimate PM$_{2.5}$ using original satellite reflectance rather than satellite-derived AOD to avoid intermediate error deserves further study. Deep learning is a promising approach for AOD-PM$_{2.5}$ modeling, but there is still room for improvement.

4. Conclusions

Despite the potential application of satellite-based AOD for air quality studies (Engel-Cox, Holloman, et al., 2004; Wang & Christopher, 2003), the estimation of PM$_{2.5}$ involves a large number of factors (Zheng et al,
2017). We therefore developed a geo-intelligent deep learning model to better represent the AOD-PM$_{2.5}$ relationship. This study introduced the layer-by-layer pretraining technique to the satellite remote sensing assessment of PM$_{2.5}$. In addition, the geographical correlation was adopted to significantly improve the estimation accuracy. The deep learning-based AOD-PM$_{2.5}$ modeling of China accurately estimated PM$_{2.5}$, with out-of-sample cross-validation $R^2$ and RMSE values of 0.88 and 13.03 µg/m$^3$, respectively. It is predicted that over 80% of the Chinese population live in areas with an annual mean PM$_{2.5}$ greater than the WHO IT-1 standard in 2015. Overall, we can say that the proposed approach is promising for air pollution monitoring in large geographical regions.

Acknowledgments

This work was funded by the National Key R&D Program of China (2016YFC020900) and the National Natural Science Foundation of China (41422018). We are grateful to the CNEMC, the Goddard Space Flight Center Distributed Active Archive Center (GSFC DAAC), the NASA Data Center, the NGCC, and the SEDAC for providing the foundational data. The data used are listed in section 2.1 and the supporting information.

References


---

LI ET AL. DEEP LEARNING FOR PM2.5 ESTIMATION 9