Estimation of the visibility in Seoul, South Korea, based on particulate matter and weather data, using machine-learning algorithm

Bu-Yo Kim\textsuperscript{1*}, Joo Wan Cha\textsuperscript{1}, Ki-Ho Chang\textsuperscript{1}, Chulkyu Lee\textsuperscript{2}

\textsuperscript{1}Research Applications Department, National Institute of Meteorological Sciences, Seogwipo, Jeju 63568, South Korea
\textsuperscript{2}Observation Research Department, National Institute of Meteorological Sciences, Seogwipo, Jeju 63568, South Korea

* Corresponding author. Tel: +82-64-780-6675
E-mail address: kimbuyo@korea.kr

Abstract

Visibility is an important indicator of air quality and of any consequent meteorological and climate change. Therefore, visibility in Seoul, which is the most polluted city in South Korea, was estimated using machine learning (ML) algorithms based on meteorological (temperature, relative humidity, and precipitation) and particulate matter (PM\textsubscript{10} and PM\textsubscript{2.5}) data acquired from an automatic weather station, and the estimated visibility was compared with the observed visibility. Meteorological data, observed at 1-h intervals between 2018 and 2020, were used. Through learning and validation of each ML algorithm, the extreme gradient boosting (XGB) algorithm was found to be most suitable for visibility estimations (bias=0 km, root mean square error (RMSE)=0.08 km, and r=1 for training data set). Among the meteorological and particulate matter data used for learning the XGB algorithm, the relative importance of PM\textsubscript{2.5} and relative humidity variables were high (51\% and 19\%, respectively), whereas precipitation and wind speed had the low relative importance (approximately 1\%). The estimation accuracy for the test dataset was good (bias=–0.11 km, RMSE=2.08 km, and r=0.94); the estimation accuracy was higher in the dry season (bias=–0.06 km, RMSE=1.79 km, and r=0.96) than in the rainy season (bias=–0.17 km, RMSE=2.34 km, and r=0.91). The results of this study indicated a higher correlation than the results of previous visibility estimation studies. The proposed method promotes accurate estimation of visibility in areas with poor visibility, and thus, it can be used to assess public health in areas with poor air quality.
Keywords: Seoul, meteorological data, PM$_{10}$, PM$_{2.5}$, visibility estimation, machine learning, extreme gradient boosting algorithm.
1 INTRODUCTION

Visibility is a measure of the distance at which an object or light source can be identified, and is defined as the distance at which the light intensity is reduced to 5% of the original level (WMO 2014). Such visibility causes low visibility of several kilometers, depending on precipitation or atmospheric suspended matter. Low visibility can cause economic losses due to road, marine, and air traffic accidents and negatively affect public health and property (Huang and Zhang, 2017; Wu et al., 2019). In particular, particulate matter (PM) released from energy use activities in urban areas, industrial activities, and population growth deteriorate visibility; a decrease in visibility of 6 to 8 km increases the mortality rate associated with heart disease and bronchitis by 2–3% (Ozer et al., 2007; Lee et al., 2015; Jeong et al., 2017). In addition, as changes in visibility are associated with changes in the meteorological parameters and climate in general (Peterson et al., 2019; Li et al., 2020; Zong et al., 2020), visibility can serve as an indicator of past, present, and future air quality improvements (Lee et al., 2014).

Previously, visibility was observed manually by human-eyes, but present observations in most countries include the use of a visibility sensor. Visibility sensors have high precision and accuracy in measuring visibility, but constructing a dense visibility observation network is difficult owing to economic and geographic constraints (Kim et al., 2021b). Therefore, regional visibility is estimated or predicted using numerical prediction models, which overcome these limitations. Fita et al. (2019) predicted the visibility using the K94, RUC, and FRAM-L models,
and compared them with the observed visibility. Zong et al. (2020) analyzed the accuracy of visibility data predicted using WRF-Chem. However, although these numerical prediction models are suitable for calculating the spatiotemporal visibility, their accuracy is low (Singh et al., 2018).

Therefore, in addition to numerical prediction models, novel methods to estimate or predict visibility using correlation between observed visibility and meteorological variables are being presently used (Bari, 2018; Fita et al., 2019).

Previous studies have shown that among the weather variables, PM, relative humidity (RH), and wind speed (WS) significantly affect visibility changes (Lee et al., 2015; Qu et al., 2015; Kim, 2019). However, visibility is not linearly proportional to these meteorological parameters. The accuracy of visibility estimations based on correlation can significantly change based on meteorological conditions (radiation, turbulence, microphysics, chemistry, and surface conditions) (Won et al., 2020). Therefore, in addition to calculating visibility using linear (Du et al., 2013) or exponential (Ozer et al., 2007; Qu et al., 2015) relationships, visibility is being actively determined using non-linear machine learning (ML) methods (Cornejo-Bueno et al., 2017, 2020), which exhibit high computational speed and high computational accuracy (Kim et al., 2021b).

In this study, visibility in Seoul, which is the most polluted area in South Korea, was estimated using meteorological (temperature, RH, and precipitation) and particulate matter (PM$_{10}$ and PM$_{2.5}$)
data, and ML algorithms. Although Seoul comprises only 0.6% of the total area of South Korea, it is home to approximately 18% of the total national population; thus, the energy consumption of this metropolitan city for domestic and industrial applications is higher than that of other cities (Lee et al., 2014). In addition to local air pollution, Seoul has high concentrations of dust and air pollutants that are generated from deserts and other cities in China and Mongolia transported according to weather patterns (Peterson et al., 2019). This perspective is important for conducting research on air quality and data utilization (Yum and Cha, 2010; Lee et al., 2014; Kim and Lee, 2018). This study aimed to determine visibility in Seoul by adopting an ML algorithm optimized for visibility estimation using meteorological and particulate matter data. The proposed method allows accurate determination of visibility without installing a visibility sensor.

2 METHODS

2.1 Research data

To estimate the visibility, meteorological and particulate matter data were collected every 1-h from January 1, 2018, to December 31, 2020, from an automatic weather station (AWS; station No. 108, 37.57° N, 126.97° E) of Korea Meteorological Administration (KMA) and an air-quality measurement station (AMS; station No. 11121, 37.56° N, 126.96° E) of the Ministry of Environment (MOE), in Seoul, South Korea. The aerial distance between the two stations is
approximately 1.1 km. PM$_{10}$ and PM$_{2.5}$ data ($\mu$g m$^{-3}$) were measured using a continuous ambient particulate matter monitor (FH62C14 (Thermo Fisher Scientific Inc., USA) and BAM 1020 (Met One Instrument Inc., USA)) by AMS, and quality-controlled PM data were obtained from Air Korea (www.airkorea.or.kr) (MOE, 2021). Data for air temperature (Ta, °C), dew point temperature (Td, °C), atmospheric pressure (Pa, hPa), RH (%), wind direction (WD, °), WS (m s$^{-1}$), and precipitation (mm h$^{-1}$, accumulated over 1-h) were collected at the AWS. Visibility (km) data, measured using an automated synoptic observing system (ASOS), were used to analyze the accuracy of the estimated visibility. Manually observed visibility data acquired from the KMA were converted to sensor-based visibility for the second half of 2017 through an automatic observation pilot operation in 2017. Therefore, to use the same observation method and objective data, data from the last three years since 2018 were used. Further, PM$_{10}$, PM$_{2.5}$, Ta, Td, Pa, RH, WD, WS, and precipitation data were used as input data for the ML algorithm, and the existing visibility data were used to evaluate the accuracy of the estimated visibility. In previous studies (Jung et al., 2009; Thach et al., 2010; Wu et al., 2012; Guo et al., 2020), visibility was estimated in dry (less than 60–70% RH) weather conditions on days with no precipitation to exclude deliquescence or hygroscopic growth of PM that occurs under high RH (Guo et al., 2020). However, in this study, all data, excluding erroneous datasets were used to estimate visibility under all weather conditions.
The collected meteorological and particulate matter data were randomly sampled at a ratio of 5:3:2 for training, validation, and testing without replacing the entire data (100%), and each dataset was constructed (Xiong et al., 2020; Kim et al., 2021a). The optimal hyperparameters were set using the training and validation datasets, respectively. The accuracy of the visibility estimated by each ML algorithm was evaluated using the visibility data. Subsequently, the ML algorithm with the highest estimation accuracy was adopted, and the estimation results for the test dataset were compared with the observed visibility data and comprehensively analyzed. A comparison of the estimated visibility (VIS\text{est}) and the observed visibility by the visibility sensor (VIS\text{obs}) is shown in Eqs. (1)–(3). Accuracy was compared using bias, root mean square error (RMSE), and correlation coefficient (r).

\[
\text{bias} = \frac{\sum_{i=1}^{N} (\text{VIS}_{\text{est}i} - \text{VIS}_{\text{obs}i})}{N} \tag{1}
\]

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\text{VIS}_{\text{est}i} - \text{VIS}_{\text{obs}i})^2}{N}} \tag{2}
\]

\[
\text{r} = \frac{\sum_{i=1}^{N} (\text{VIS}_{\text{est}i} - \overline{\text{VIS}_{\text{est}}})(\text{VIS}_{\text{obs}i} - \overline{\text{VIS}_{\text{obs}}})}{\sqrt{\sum_{i=1}^{N} (\text{VIS}_{\text{est}i} - \overline{\text{VIS}_{\text{est}}})^2 \sum_{i=1}^{N} (\text{VIS}_{\text{obs}i} - \overline{\text{VIS}_{\text{obs}}})^2}} \tag{3}
\]

Here, N is the sample number.
2.2 Machine learning algorithms

The ML algorithms used in this study were artificial neural network (ANN), extreme learning machine (ELM), k-nearest neighbor (kNN), random forest (RF), support vector regression (SVR), and extreme gradient boosting (XGB) among supervised learning regression methods. Each hyperparameter of these algorithms was repeatedly grid-searched with fine resolution (Bergstra and Bengio, 2012; Kim et al., 2021a). A brief description of the hyperparameter options for each ML algorithm used in this study are as follows:

2.2.1 Artificial neural network

ANN is a single perceptron composed of an input layer, an output layer, and a hidden layer between the two layers (Fig. 1a) (Rosa et al., 2020). In this study, the R “nnet” package (Ripley and Venables 2021a) was used, and the hyperparameters were set as follows: size (number of hidden nodes)=10, maxit (maximum number of iterations)=900, and decay (parameter for weight decay)=0.5.

2.2.2 Extreme learning machine

ELM is a multi-perceptron composed of an input node, a hidden node, and an output node (Fig. 1b). It predicts by applying a weight \(w\) and bias \(b\) between the input node and the hidden node
and a weight ($\beta$) between the hidden node and the output node (Huang et al., 2006). Additionally, the ELM can be predicted by applying weights and biases to input and output vectors using a single hidden layer feed-forward neural network training method, as shown in Eq. (4) (Wang et al., 2021). The number of hidden nodes, which is a hyperparameter of the ELM algorithm, was set to 1000.

$$y_i = \sum_{j=1}^{L} \beta_j g(w_j x_i + b_j), \quad i = 1, \ldots, N$$ (4)

2.2.3 k-nearest neighbor

kNN determines the $k$ neighbors closest to the query in the data feature space (Fig. 1c) and predicts the query using distance-based weights (Zhang et al., 2018). In this study, the R “class” package (Ripley and Venables, 2021b) was used, and the hyperparameter $k$ was set to 12.

2.2.4 Support vector regression

SVR determines a hyperplane composed of support vectors that can classify the maximum margin for the distance between vectors (Fig. 1d) and returns the data based on the $\varepsilon$-insensitive loss function (Taghizadeh-Mehrjardi et al., 2017). In this study, the R “e1071” package (Meyer et al., 2021) and the radial basis function (RBF) kernel of SVR were used, and the hyperparameters were set as follows: epsilon ($\varepsilon$)=0.1, gamma ($\gamma$)=0.2, and cost (C)=3.
2.2.5 Random forest

RF constructs N decision trees by combining randomly selected variables for each node (Fig. 1e), and predicts the results by ensembles the results of each decision tree (Wright and Ziegler, 2017). In this study, the R “ranger” package (Wright et al., 2020) was used and the hyperparameters were set as follows: num.trees (number of trees)=790, mtry (number of variables randomly sampled at each node)=8, and min.node.size (minimal node size)=4.

2.2.6 Extreme gradient boosting

XGB improves the predictive power by sequential reinforcement learning of a decision tree through boosting (Fig. 1f). In this study, the R “xgboost” package (Chen et al., 2019) and the Gaussian distribution function kernel were used; additionally, the hyperparameters were set as follows: n.rounds (maximum number of iterations)=1080, max_depth (maximum depth of binary tree)=8, and eta (learning rate)=0.1.

2.3 Time series of meteorological and particulate matter variables in Seoul

The daily mean time series showed varying trends in the collected data (2018–2020) (Fig. 2). PM$_{10}$ and PM$_{2.5}$ concentrations in Seoul increase in the dry season (December–May) (Hur et al., 2016),
during which yellow sand and air pollutants arising from fossil fuel use and industrial activities in China, Mongolia, and surrounding urban areas are carried into Seoul by the westerlies (Ghim et al., 2015; Oh et al., 2015; Jeong et al., 2017; Peterson et al., 2019; Oh et al., 2020; Hur et al., 2021). According to the air pressure pattern, high PM concentration is maintained for long periods, which deteriorates visibility (Lee et al., 2013; Kim and Chun, 2013; Park et al., 2019). In particular, in the case of PM$_{2.5}$, the higher the concentration, the greater is the decrease in visibility owing to strong scattering in the atmosphere (Ma et al., 2020). PM$_{10}$ and PM$_{2.5}$ increases with an increase in surface temperature and WS (Kim et al., 2017; Kim, 2019; Plocoste and Galif, 2021), but decreases due to precipitation in the rainy season (June–November) (Lee et al., 2013; Kim and Kim, 2020). In addition, high RH scatters light from hygroscopically grown PM, which further reduces the visibility (Jung et al., 2009; Lee et al., 2014; Qu et al., 2015; Ma et al., 2020). Ta–Td and RH shows a negative correlation ($r=–0.97$), and as Ta–Td approaches 0 K, the atmosphere becomes wetter and condensation occurs, which acts as an important factor in the deterioration of visibility (Yu et al., 2019). Therefore, visibility is impaired during periods of high PM (especially PM$_{2.5}$) concentrations and high RH (low Ta–Td) (Zhang et al., 2010). During precipitation, visibility deteriorates, but visibility is also improved by the cleaning effect of suspended matter in the atmosphere (Founda et al., 2016; Kim et al., 2021b). The occurrence frequency of the monthly mean low visibility in Seoul is shown in Fig. 3. The occurrence frequency of low visibility was
high in the dry season when PM concentration was high (<10 km: 19–37%, <5 km: 6–15%), and showed a low distribution in the rainy season when the PM$_{10}$ concentration was low (<10 km: 7–24%, <5 km: 3–9%). This pattern was similar to the low visibility frequency distribution observed in Hebei, China, wherein SO$_2$ emissions and the occurrence frequency of low visibility were proportionally related (Fu et al., 2014). In addition, the occurrence frequency of low visibility was the lowest in August, due to multiple precipitation events (approximately 38 cases) with high rainfall intensity (approximately 2.28 mm h$^{-1}$).

3 RESULTS AND DISCUSSION

3.1 Training and validation results of machine learning algorithms

Table 1 shows the visibility estimation results for each ML algorithm with hyperparameters optimized using the training and validation datasets. The decision tree-based XGB and RF algorithms performed better than the neural network algorithm-based ANN and ELM, while the kNN algorithm based on the vector distance between data showed the poorest performance. The XGB algorithm showed the best output performance in the training and validation datasets and was the most suitable for estimating the visibility using the present data. In addition, compared to other algorithms, XGB showed very fast learning and predictions (requiring only a few seconds). The results of the XGB algorithm were in good agreement with the visibility observations and 1:1 line (Fig. 4), showing a small difference (bias=0 km, RMSE=0.08 km, and $r$ =1). Therefore,
XGB, which showed low computational cost and high accuracy, was selected as the visibility estimation algorithm for this study. The relative importance of the input variables for learning the XGB algorithm is shown in Fig. 5. Relative importance indicates the relative contribution of a feature to the predicted result based on impurity variance to each split leaf (data feature) in the process of growing each node of the tree. The relative importance of PM$_{2.5}$ (51.05%) and RH (18.82%), and Ta–Td (12.18%) variables were the high, while that of precipitation (1.17%) and WS (0.96%) was the low. PM concentration and RH have a significant influence on changes in visibility (Maurer et al., 2019; Kim et al., 2021b). PM$_{2.5}$ was shown to be the most important variable in estimating visibility because suspended matter of size smaller than PM$_{10}$ caused larger scattering in the atmosphere and more frequent disturbances in visibility such as haze (Cheng et al., 2017, Ma et al., 2020). RH and Ta–Td act as important factors in reducing visibility by changing the characteristics of atmospheric aerosols and causing condensation in the wet atmosphere (Yu et al., 2019). Julian day reflects the monthly and seasonal periodicity of variations in visibility, and temperature and pressure are related to overall weather (Kim et al., 2021b). The hour variable can also reflect variations in visibility or the daily periodicity of weather variables that affect visibility variations. However, it showed a relatively low importance because visibility varies directly according to variables such as PM$_{2.5}$ and RH. In the case of wind direction, visibility varies in relation to the
inflow of air pollutants and dry or wet air; visibility is improved according to variations in wind speed (Ma et al., 2020). However, in this study, the variation in wind speed was not large and the contribution to the visibility estimation by other variables (such as PM$_{2.5}$ and RH) were high, indicating its relatively low importance. When precipitation occurs, changes in weather conditions occurred, such as high RH (low Ta–Td) and a decrease in PM concentration, and the visibility deteriorated or improved, thereby reducing the corresponding relative importance. When precipitation occurs, changes in weather conditions occurred, such as high RH (low Ta–Td) and a decrease in PM concentration, and the visibility deteriorated or improved, thereby reducing the corresponding relative importance.

3.2 Analysis of the visibility estimation results

The scatter plot of visibility (VIS$_{XGB}$) estimated by the XGB algorithm using the test dataset, and the observed visibility (VIS$_{obs}$) (bias=$-0.11$ km, RMSE=2.08 km, and r=0.94) is shown in Fig. 6. The correlation was stronger than that determined by Du et al. (2013), which assessed the linear relationship between visibility and the meteorological variables in metropolitan areas in China ($r=0.62$), Qu et al. (2015) and Won et al. (2020), which assessed the exponential relationship between visibility and PM$_{10}$ ($r=0.79$) and PM$_{2.5}$ ($r=0.87$), and Zong et al. (2020), which used WRF-Chem and ML algorithms ($r=0.42$). In addition, the correlation was higher than
that determined by Sohn and Kim (2015) \( (r=0.71) \), which estimated the visibility in Seoul. These previous studies estimated visibility during sunny days or during periods of low RH \(<60\%\)\. Fig. 7 shows Fig. 6 as a daily mean time series, wherein \( \text{VIS}_{XGB} \) shows a trend similar to \( \text{VIS}_{\text{obs}} \), with a small difference and a high correlation coefficient \( (\text{bias}=–0.11 \text{~km}, \text{RMSE}=0.68 \text{~km}, \text{and} \ r=0.98) \). That is, the visibility estimated using the ML algorithm and meteorological and particulate matter data in this study could explain approximately 96\% of the observed visibility \( (r^2=0.96) \).

Table 2 shows the monthly mean meteorological and particulate matter variables and the visibility estimation accuracy for the test dataset. The estimation accuracy was relatively higher in the dry season \( (\text{bias}=–0.06 \text{~km}, \text{RMSE}=1.79 \text{~km}, \text{and} \ r=0.96) \) than in the rainy season \( (\text{bias}=–0.17 \text{~km}, \text{RMSE}=2.34, \text{and} \ r=0.91) \). During the dry season, visibility was low in Seoul. The dry season was characterized by higher \( \text{PM}_{10} \) and \( \text{PM}_{2.5} \) concentrations, higher \( \text{Ta}–\text{Td} \), lower RH, and less precipitation than the rainy season \( (\text{Kim et al., 2021b}) \). The visibility estimation accuracy was lower in summer (June–August) than in other seasons \( (\text{bias}=–0.17 \text{~km}, \text{RMSE}=2.43 \text{~km}, \text{and} \ r=0.91) \) because of multiple precipitation days and strong precipitation intensity, such as in August. Fig. 8 compares the \( \text{VIS}_{\text{obs}} \) and \( \text{VIS}_{XGB} \) accuracies for each variable interval for each meteorological and particulate matter variable. As for the accuracy of estimating the visibility for each section of \( \text{PM}_{10} \) and \( \text{PM}_{2.5} \), the RMSE decreased and \( r \) increased as the PM concentration...
increased. That is, a high PM concentration (especially PM$_{2.5}$) has a great effect on the decrease in visibility; therefore, the estimated results were highly accurate (Won et al., 2020). RH and Ta–Td variables showed relatively high visibility estimation accuracy, except during very dry or wet weather conditions. In particular, since the scattering characteristics of aerosols in the atmosphere change as the atmosphere becomes wetter, a large error may occur in the visibility estimation (Jung et al., 2009). In the case of wind speed, the visibility estimation error was relatively large at 6 m s$^{-1}$ or more. The data characteristics of this section had the highest precipitation (0.45 mm h$^{-1}$) and the lowest PM concentration (PM$_{10}$: 26.92 μg m$^{-3}$, PM$_{2.5}$: 15.23 μg m$^{-3}$) compared to other sections. Conversely, in the case of precipitation, the difference increased according to the presence and absence of precipitation and the intensity of precipitation, while correspondingly, the correlation coefficient decreased significantly. That is, visibility can improve or deteriorate depending on the precipitation characteristics (type and intensity), which causes difficulties in estimating the visibility (Gultepe and Milbrandt, 2010). Nevertheless, the monthly mean visibility estimation accuracy was as follows: bias=–0.11 km, RMSE=2.05 km, and $r$=0.93. These results showed lower variability and higher accuracy than previous studies that quantitatively estimated precipitation using satellite-based (Nguyen et al., 2021), radar-based (Shin et al., 2019), and numerical model-based (Ko et al., 2020) data using ML algorithms. Therefore, the application of
visibility estimations using ML and meteorological and particulate matter data is expected to be high.

4 CONCLUSIONS

In this study, visibility in Seoul, South Korea, was estimated using meteorological and particulate matter data acquired from the AWS of KMA and AMS of MOE observatory, and using an ML algorithm; moreover, the estimated visibility and visibility observed by the visibility sensor were compared and analyzed. Weather information (temperature, RH, and precipitation) observed by the AWS, and PM$_{10}$ and PM$_{2.5}$ observed by AMS data were used. The visibility estimation performance of the ML and XGB algorithms was superior to that of the RF, neural networks (ANN and ELM), and vector distance-based algorithms (kNN and SVR). The relative importance of the variables input in this process was approximately 51% and 19% for PM$_{2.5}$ and RH, respectively. Conversely, the relative importance of precipitation and WS were the low (approximately 1%). Visibility estimated using the test dataset (bias=–0.11 km, RMSE=2.08 km, and $r=0.94$) showed higher accuracy than the results of previous studies. Moreover, in this study, the meteorological conditions (low RH, no precipitation, and sunny days) were not restricted for visibility estimations, and thus, visibility was estimated for all weather conditions. Although the estimated accuracy of visibility differed based on the month and season, the accuracy of the estimated visibility was high during the dry season (December–May), when the frequency of low visibility was high. The
accuracy of estimating the monthly mean visibility was high (bias=–0.12 km, RMSE=2.05 km, and 
r=0.93); thus, visibility could be estimated with high accuracy using meteorological and particulate 
matter variables and ML algorithms. Large metropolitan cities with high floating populations, such 
as Seoul, are extremely sensitive to public health issues related to air quality (Kim and Lee 2018). 
However, densely populated high-rise buildings and high real estate prices in these areas increase 
the difficulty in establishing observation stations to measure visibility. Therefore, the method 
proposed in this study can assist in visibility estimations in areas where visibility cannot be 
observed using only the available meteorological and particulate matter data.

ACKNOWLEDGMENTS

This work was funded by the Korea Meteorological Administration Research and Development 
Program “Research on Weather Modification and Cloud Physics” under Grant (KMA2018-00224).

DISCLAIMER

The authors declare that they have no known competing financial interests or personal 
relationships that could have appeared to influence the work reported in this paper.

REFERENCES


Low-Visibility Events at Valladolid Airport, Spain. Symmetry 12(6), 1045. https://doi.org/10.3390/sym12061045


Table 1. Visibility estimation results of the training and validation datasets for each machine learning (ML) algorithm.

<table>
<thead>
<tr>
<th>Data set</th>
<th>VIS\textsubscript{obs} mean (km)</th>
<th>Bias (km)</th>
<th>RMSE (km)</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGB</td>
<td>Training</td>
<td>14.42</td>
<td>-0.00</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>14.41</td>
<td>-0.07</td>
<td>2.12</td>
</tr>
<tr>
<td>RF</td>
<td>Training</td>
<td>14.42</td>
<td>-0.01</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>14.41</td>
<td>-0.02</td>
<td>2.23</td>
</tr>
<tr>
<td>SVR</td>
<td>Training</td>
<td>14.42</td>
<td>0.03</td>
<td>1.94</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>14.41</td>
<td>0.02</td>
<td>2.32</td>
</tr>
<tr>
<td>ANN</td>
<td>Training</td>
<td>14.42</td>
<td>-0.06</td>
<td>2.39</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>14.41</td>
<td>-0.06</td>
<td>2.42</td>
</tr>
<tr>
<td>ELM</td>
<td>Training</td>
<td>14.42</td>
<td>-0.15</td>
<td>2.38</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>14.41</td>
<td>-0.20</td>
<td>3.64</td>
</tr>
<tr>
<td>kNN</td>
<td>Training</td>
<td>14.42</td>
<td>0.13</td>
<td>2.10</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>14.41</td>
<td>0.20</td>
<td>2.83</td>
</tr>
</tbody>
</table>
Table 2. Monthly mean VIS$_{\text{obs}}$, PM$_{10}$, PM$_{2.5}$, Ta–Td, RH, WS, and precipitation data and monthly estimation accuracy of VIS$_{\text{XGB}}$. Figures in parentheses indicate standard deviations.

<table>
<thead>
<tr>
<th>Month (N)</th>
<th>VIS$_{\text{obs}}$ mean (km)</th>
<th>PM$_{10}$ (μg m$^{-3}$)</th>
<th>PM$_{2.5}$ (μg m$^{-3}$)</th>
<th>Ta–Td (K)</th>
<th>RH (%)</th>
<th>WS (m s$^{-1}$)</th>
<th>Precip. (mm h$^{-1}$)</th>
<th>Bias (km)</th>
<th>RMSE (km)</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (430)</td>
<td>13.47 (6.10)</td>
<td>49.23 (31.15)</td>
<td>33.37 (24.00)</td>
<td>9.74 (4.67)</td>
<td>50.38 (17.55)</td>
<td>1.91 (1.30)</td>
<td>0.04 (0.36)</td>
<td>−0.22</td>
<td>1.55</td>
<td>0.97</td>
</tr>
<tr>
<td>2 (400)</td>
<td>13.81 (5.96)</td>
<td>46.17 (22.06)</td>
<td>30.20 (17.44)</td>
<td>10.49 (5.31)</td>
<td>48.59 (19.15)</td>
<td>1.93 (1.18)</td>
<td>0.03 (0.22)</td>
<td>−0.04</td>
<td>1.63</td>
<td>0.96</td>
</tr>
<tr>
<td>3 (416)</td>
<td>12.69 (6.15)</td>
<td>46.91 (29.53)</td>
<td>31.00 (21.84)</td>
<td>10.32 (6.23)</td>
<td>52.11 (20.95)</td>
<td>2.14 (1.45)</td>
<td>0.02 (0.13)</td>
<td>−0.03</td>
<td>1.54</td>
<td>0.97</td>
</tr>
<tr>
<td>4 (404)</td>
<td>14.78 (5.44)</td>
<td>38.71 (20.94)</td>
<td>21.16 (11.91)</td>
<td>10.95 (10.35)</td>
<td>51.64 (21.43)</td>
<td>2.22 (1.45)</td>
<td>0.12 (0.70)</td>
<td>−0.08</td>
<td>2.02</td>
<td>0.93</td>
</tr>
<tr>
<td>5 (389)</td>
<td>13.01 (6.20)</td>
<td>41.09 (24.07)</td>
<td>22.34 (13.81)</td>
<td>9.86 (6.44)</td>
<td>57.11 (22.01)</td>
<td>2.03 (1.35)</td>
<td>0.11 (0.50)</td>
<td>−0.10</td>
<td>1.83</td>
<td>0.96</td>
</tr>
<tr>
<td>6 (412)</td>
<td>13.67 (5.86)</td>
<td>30.67 (15.14)</td>
<td>20.25 (10.18)</td>
<td>8.43 (5.20)</td>
<td>62.23 (18.30)</td>
<td>1.95 (1.19)</td>
<td>0.12 (0.87)</td>
<td>−0.12</td>
<td>2.41</td>
<td>0.91</td>
</tr>
<tr>
<td>7 (429)</td>
<td>14.57 (5.99)</td>
<td>24.45 (15.61)</td>
<td>16.77 (11.92)</td>
<td>6.54 (4.29)</td>
<td>69.57 (16.19)</td>
<td>1.91 (1.13)</td>
<td>0.11 (0.92)</td>
<td>−0.05</td>
<td>2.27</td>
<td>0.93</td>
</tr>
<tr>
<td>8 (357)</td>
<td>16.10 (4.90)</td>
<td>22.91 (11.50)</td>
<td>15.67 (9.20)</td>
<td>6.52 (4.42)</td>
<td>70.14 (16.81)</td>
<td>1.64 (1.02)</td>
<td>0.19 (1.00)</td>
<td>−0.36</td>
<td>2.65</td>
<td>0.85</td>
</tr>
<tr>
<td>9 (389)</td>
<td>17.38 (4.58)</td>
<td>20.00 (10.55)</td>
<td>11.46 (7.19)</td>
<td>7.23 (4.38)</td>
<td>65.89 (16.67)</td>
<td>2.01 (1.30)</td>
<td>0.16 (0.83)</td>
<td>−0.22</td>
<td>2.51</td>
<td>0.84</td>
</tr>
<tr>
<td>10 (347)</td>
<td>15.95 (5.07)</td>
<td>28.12 (17.56)</td>
<td>14.71 (8.38)</td>
<td>8.23 (4.74)</td>
<td>60.64 (17.95)</td>
<td>1.83 (1.21)</td>
<td>0.10 (0.73)</td>
<td>0.04</td>
<td>1.89</td>
<td>0.93</td>
</tr>
<tr>
<td>11 (406)</td>
<td>14.61 (5.96)</td>
<td>43.61 (27.04)</td>
<td>23.67 (15.76)</td>
<td>8.11 (4.96)</td>
<td>59.54 (19.31)</td>
<td>1.97 (1.21)</td>
<td>0.12 (0.84)</td>
<td>−0.32</td>
<td>2.23</td>
<td>0.93</td>
</tr>
<tr>
<td>12 (416)</td>
<td>14.71 (6.03)</td>
<td>43.56 (21.11)</td>
<td>28.11 (16.78)</td>
<td>9.04 (4.54)</td>
<td>53.15 (17.95)</td>
<td>2.12 (1.19)</td>
<td>0.03 (0.25)</td>
<td>0.11</td>
<td>2.08</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Fig. 1. Schematic diagram of each machine learning algorithm: a) ANN, b) ELM, c) kNN, d) SVR, e) RF, and f) XGB (Kim et al., 2021a).
a) Visibility (km), PM$_{10}$ ($\mu$g m$^{-3}$), and PM$_{2.5}$ ($\mu$g m$^{-3}$)

b) Ta ($^\circ$C), Ta−Td (K), and RH (%)

c) Precipitation (mm h$^{-1}$), WS (m s$^{-1}$), and WD ($^\circ$)

**Fig. 2.** Time series of daily mean meteorological and particulate matter data in Seoul: a) Visibility (black), PM$_{10}$ (red), and PM$_{2.5}$ (blue); b) Ta (red), Ta−Td (blue), and RH (black); and c) precipitation (red), wind speed (blue), and wind direction (black arrow).
Fig. 3. Occurrence frequency of monthly mean low visibility in Seoul (<10 km: black, <5 km: gray).
Fig. 4. Scatter plots of the observed visibility (VIS_{obs}) and the estimated visibility (VIS_{XGB}) by the XGB algorithm for the training dataset. The red line is the 1:1 line.
Fig. 5. Variable relative importance of the XGB algorithm on training.
Fig. 6. Scatter plots of observed visibility (VIS$_{obs}$) and estimated visibility (VIS$_{XGB}$) for the test dataset. The red line is the 1:1 line.
Fig. 7. Daily mean time series of observed visibility (VIS$_{obs}$) (black) and estimated visibility (VIS$_{XGB}$) (red).
Fig. 8. Visibility estimation accuracy for each interval for each meteorological and particulate matter variable (a) PM$_{10}$, b) PM$_{2.5}$, c) RH, d) Ta–Td, e) WS, and f) precipitation). The number in parentheses below the interval represents the data ratio (%) to the total data.