Communications for Statistical Applications and Methods
2023, Vol. 30, No. 5, 517–530

https://doi.org/10.29220/CSAM.2023.30.5.517
Print ISSN 2287-7843 / Online ISSN 2383-4757

Prediction of PM10 concentration in Seoul, Korea using Bayesian network

Minjoo Jo[a], Rosy Oh[b], Man-Suk Oh[1,a]

[a]Department of Statistics, Ewha Womans University, Korea;
[b]Department of Mathematics, Korea Military Academy, Korea

Abstract

Recent studies revealed that fine dust in ambient air may cause various health problems such as respiratory diseases and cancer. To prevent the toxic effects of fine dust, it is important to predict the concentration of fine dust in advance and to identify factors that are closely related to fine dust. In this study, we developed a Bayesian network model for predicting PM10 concentration in Seoul, Korea, and visualized the relationship between important factors. The network was trained by using air quality and meteorological data collected in Seoul between 2018 and 2021. The study results showed that current PM10 concentration, season, carbon monoxide (CO) were the top 3 effective factors in 24 hours ahead prediction of PM10 concentration in Seoul, and that there were interactive effects.

Keywords: air pollution, prediction, PM10, machine learning, Bayesian network

1. Introduction

Recently, air pollution has been a major environmental concern in Korea. Particularly, Seoul, the capital city of Korea with about 10 million residents, ranked 1217 out of more than 8,000 cities in the world in order of poor air quality, based on data collected during 2017–2022 by IQAir, the Swiss air quality technology company (www.iqair.com).

Fine dust is classified into primary fine dust and secondary fine dust according to the generation process. Primary fine dust is emitted through fuel combustion or vehicle exhaust, and secondary fine dust is generated through chemical reactions between already emitted substances (Jang, 2016). Studies have shown that fine dust causes cardiopulmonary diseases, lung cancer, respiratory diseases, and stroke (Shin, 2007; Han *et al*., 2019; Lee *et al*., 2020; Yang *et al*., 2020; Kim *et al*., 2022).

To prevent the adverse effects of fine dust on the human body, it is important to predict the concentration of fine dust in advance in order to minimize exposure to fine dust and to identify factors that are closely related to fine dust in order to reduce concentration of fine dust (Lim, 2019; Cha and Kim, 2018; Shin *et al*., 2007).

Various approaches to fine dust analysis have been proposed by many researchers. However, fine dust is affected by many factors such as weather conditions, population density, natural environment, industrial environment, and geographical conditions. Also, currently fine dust is measured every hour

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT, ME) (NRF-2022R1A2C1091256, 2020R1A2C1008699).

[1] Corresponding author: Department of Statistics, Ewha Womans University, 52 Ewha Womans University-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: msoh@ewha.ac.kr

Published 30 September 2023 / journal homepage: http://csam.or.kr

at observatories, producinpg a large amount of data. Therefore, there may be limitations in analyzing such a large amount of complex data using traditional statistical methods and fine dust analysis using machine learning techniques has recently attracted great attention of researchers (Cha and Kim, 2018; Cho *et al*., 2019; Lim, 2019).

Bayesian network is a type of machine learning technique, which is a probabilistic graphical model that uses conditional independence/dependence to create a directed acyclic graph (DAG) that expresses the relationship between variables. Bayesian network is a model often used for large amount of complex data in medical, environmental, and transportation area. Recent studies on predicting diseases using Bayesian network include Oh *et al.* (2022) and Park *et al.* (2018). Compared to other machine learning techniques, the key advantage of Bayesian networks is interpretability. Using the DAG created by Bayesian network, relationships between variables can be visualized and causative factors may be discovered. In addition, in contrast to other machine learning techniques that use only data when creating a DAG, Bayesian network has the flexibility to insert or delete relationships between specific variables by reflecting expert knowledge (Oh *et al*., 2022). For a detailed coverage of Bayesian networks, the reader is referred to Korb and Nicholson (2010), Daly *et al.* (2011), Sesen *et al.* (2013). In spite of these great advantages of Bayesian networks, there are few studies that have analyzed factors that affect fine dust concentration in Korea using Bayesian networks.

In this study, we constructed Bayesian network model for predicting concentration of fine dust in Seoul, Korea, identifying important factors that affect fine dust, and visualizing the relationship between variables. We used PM10 concentration as a measure of the concentration of fine dust. PM10 is dust with a diameter of 10 or less floating in the air, and it is a particle so thin and small that we cannot see (Korean Ministry of Environment, 2014). As predictors, we used variables related to air quality (Lim, 2019; Cho *et al*., 2019; Cha and Kim, 2018) such as ozone (O3), nitrogen dioxide (NO2), carbon monoxide (CO), sulfur dioxide (SO2), meteorological variables such as temperature, wind speed, wind direction, humidity, and temportal variables such as season, time, and PM10 concentration 24 hours before the prediction time. It is well known that air quality variables are related to the generation process of fine dust that determines atmospheric chemical components in the air (Kim *et al*., 2016), and meteorological factors affect the transportation and diffusion of pollutants in the air (Shin *et al*., 2007). It is also well known that PM10 is subject to seasonal and temporal influences (Oh and Park, 2022).

This article consists of of 5 sections. Section 2 briefly describes the Bayesian network models used in this study, and Section 3 presents a description and preparation of the data. Section 4 describes construction of the Bayesian network prediction model and interpretation of the network, and Section 5 presents conclusions and comments.

2. Bayesian network

A Bayesian network is a probabilistic graphical model based on a DAG. The nodes of DAG represent variables, and the connected arcs represent conditional dependencies between variables (Pearl, 2011). The joint probability distribution for the nodes of the Bayesian network can be expressed as equation (2.1), where $P_B(X_1, \ldots, X_n)$ is the joint probability distribution of n random variables $X = \{X_1, \ldots, X_n\}$, Pa_i is the parent node of X_i, and $P(X_i|Pa_i)$ is the conditional distribution of X_i given Pa_i.

$$P_B\left(X_1, \ldots, X_n\right) = \prod_{i=1}^{n} P\left(X_i \mid Pa_i\right). \tag{2.1}$$

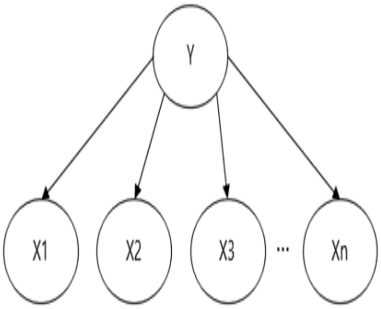

Figure 1: *The structure of Naive Bayes.*

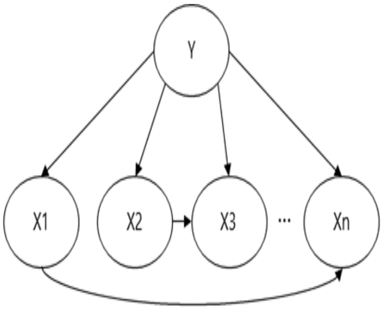

Figure 2: *The structure of Tree Augmented Naive Bayes.*

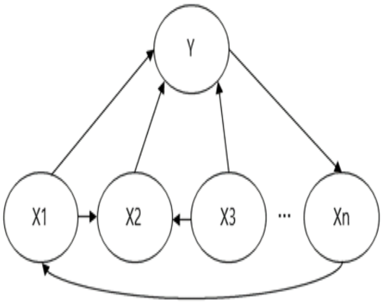

Figure 3: *The structure of General Bayesian Network.*

Bayesian networks make it easy to identify the conditional independence/dependence between variables by visualizing the relationship between variables, and have the advantage of being able to utilize expert knowledge when creating DAGs (Yang and Zhang, 2000). Especially, Bayesian networks can be used as a classifier to determine the probability distribution of a class node based on the conditional probability of other attributes. The classification is based on the highest posterior probability distribution. Bayesian networks are useful for classification because they allow for a fast and intuitive understanding of the relationships between nodes (Ang *et al.*, 2016).

Commonly used Bayesian network structures in classification are Naive Bayes (NB), Tree Augmented Naive Bayes (TAN), and General Bayesian Network (GBN) (Oh *et al.*, 2022). In this study, the concentration of PM10 prediction model was developed by applying these three structures. R package bnlearn was used to fit Bayesian network models.

Table 1: Description and discretization of variables

Variable	Description	Category
fPM10	PM10 concentration after 24 hours ($\mu g/m^3$)	good(0-30); normal(31-80); bad(81-)
cPM10	PM10 concentration ($\mu g/m^3$)	good(0-30); normal(31-80); bad(81-)
O3	O3 concentration (ppm)	good(0-0.03); normal(0.031-0.09); bad(0.091-)
NO2	NO2 concentration (ppm)	good(0-0.03); normal(0.031-0.06); bad(0.061-)
CO	CO concentration (ppm)	<Q1, Q1-Q2, Q2-Q3, >Q3
SO2	SO2 concentration (ppm)	<Q1, Q1-Q3, >Q3
Temperature	temperature (°C)	<Q1, Q1-Q2, Q2-Q3, >Q3
Wind.speed	wind speed (m/s)	<Q1, Q1-Q2, Q2-Q3, >Q3
Humidity	humidity (%)	<Q1, Q1-Q2, Q2-Q3, >Q3
Season	season	spring(Apr-May); summer-fall(Jun-Oct); winter(Nov-Mar)
Time	time zone	dawn; morning; afternoon; night
Wind.direction	wind direction	E; W; S; N; NE; NW; SE; SW

- NB is the simplest Bayesian network structure and is widely used in various fields such as medical and computer networks. NB assumes that all predictor variables affect the response variable independently, and treats only the response variable as the only parent node of each predictor variable. It is easy to interpret due to the simple structure. However, it rarely reflects the true relationship between variables in real data (Ang *et al*., 2016). The structure of NB is illustrated in Figure 1.

- TAN relaxes the structural restriction of NB by allowing the dependencies between predictor variables, hence the reliability of the structure and classification accuracy are usually higher than that of NB (Ang *et al*., 2016). Since TAN allows adding a second parent node to each predictor variable, every predictor variable has a maximum of two parent nodes (Witten *et al*., 2011). It is a model that compromises simplicity and practicality by allowing up to one additional node compared to NB. The structure of TAN is illustrated in Figure 2.

- GBN has no limit on the number of parent nodes in the DAG structures. In addition, it does not distinguish between the response and the predictor variable, and the response variable is considered as another predictor variable (Madden, 2009). While GBN has the advantage of allowing more flexibility, it frequently fails to provide an interpretive model in real data. The structure of GBN is illustrated in Figure 3. There are several ways to find a GBN structure suitable for a given data. In this article, the *hc* function of the *bnlearn* package was used to find the structure of GBN. The function *hc* uses a hill climbing algorithm, which is a greedy search method that finds the optimal DAG through the process of adding, removing, or changing arcs (Scutari *et al*., 2022).

3. Data

3.1. Data collection

Data on variables related to air quality - the concentrations of PM10 ($\mu g/m^3$), O3 (ppm), NO2 (ppm), CO (ppm), and SO2 (ppm) - were obtained from Air Korea (https://www.airkorea.or.kr/). We used hourly data collected from April 1, 2018 to March 31, 2021 at a station located at 169 Jong-ro, Jongno-gu, Seoul (in front of Jongmyo parking lot).

Data on meteorological variables - temperature (°C), wind speed (m/s), wind direction (16 directions), and humidity (%) - were obtained from the Automated Synoptic Observing System (ASOS) data of the Korea Meteorological Administration Meteorological Data Open Portal (https://data.kma.go.kr/). Specifically, we used hourly data collected from April 1, 2018 to March 31, 2021 at Seoul Ob-

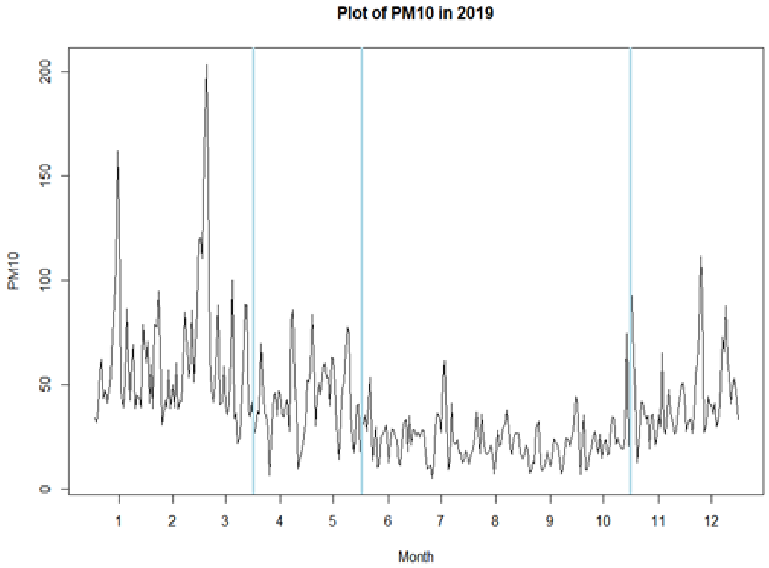

Figure 4: *Average daily PM10 concentration in 2019.*

servatory of Korea Meteorological Administration, 52 Songwol-gil, Jongno-gu, Seoul, which is very close to the station where the observations of the air quality variables were collected.

3.2. Data pre-processing

Considering that PM10, O3, NO2, CO, SO2, temperature, wind.speed, and humidity data were collected over time, missing values of these variables were imputed using linear interpolation (Roh, 2017), using na.interp function in R package forecast. Missing values of a categorical variable wind.direction were replaced with the most recent values.

Since weather forecasts usually predict the PM10 concentration of the next day, at each time point the concentration of PM10 after 24 hours was used as the target variable and it was named fPM10. As predictor variables, air quality variables (PM10, O3, NO2, CO, SO2) and meteorological variables (temperature, wind speed, wind direction, humidity) at each time point were used as predictors. The concentration of PM10 at each time point was named cPM10 to distinguish it from fPM10. The letters “c” and “f” in cPM10 and fPM10 stand for “current” and “future”, respectively. The names and descriptions of the variables are given in Table 1.

For categorical variables wind.direction, time, and month, we re-categorized them to reduce the number of categories. For wind.direction, the original 16 categories were merged into 8 categories - E (East), W (West), S (South), N (North), NE (Northeast), NW (Northwest), SE (Southeast), SW (Southwest). For time variable, its original 24 categories, from 0 to 23, were converted into 4 time zones - dawn (0–5), morning (6–11), afternoon (12–17), night (18–23), since the variation of fine dust concentration in each time zone is not large according to the Meteorological Administration of Korea.

To see the seasonal effect on PM10, the daily average PM10 concentration in 2019 was plotted (Figure 4). As shown in Figure 4, PM10 was highest from November to March, followed by April and May, and generally low from June to October. Therefore, we converted the variable month to season which has 3 categories; spring (Apr-May), summer-fall (Jun-Oct), winter (Nov-Mar).

Table 2: Basic statistics of predictor variables according to fPM10

		fPM10				
		Total	Good	Normal	Bad	Assoc
		(N = 26244)	(N = 11969)	(N = 12644)	(N = 1631)	*p*-value*
Variable		Mean ± SD or N (%)				
cPM10		38.44 ± 26.62	44.69 ± 24.93	27.56 ± 18.66	69.78 ± 43.35	<0.001
O3		0.0223 ± 0.0174	0.0235 ± 0.0191	0.0215 ± 0.0153	0.0188 ± 0.0172	<0.001
NO2		0.0312 ± 0.0152	0.0327 ± 0.0153	0.028 ± 0.014	0.0438 ± 0.0148	<0.001
CO		0.5163 ± 0.2119	0.5538 ± 0.2137	0.4452 ± 0.1652	0.7463 ± 0.2591	<0.001
SO2		0.0033 ± 0.0010	0.0035 ± 0.0011	0.003 ± 8e-04	0.0040 ± 0.0013	<0.001
Temperature		13.47 ± 10.69	11.16 ± 10.60	16.59 ± 10.36	8.44 ± 6.59	<0.001
Wind speed		2.0624 ± 1.1494	2.0979 ± 1.1501	2.0435 ± 1.1523	1.9256 ± 1.1097	<0.001
Humidity		59.82 ± 20.33	56.00 ± 20.26	64.42 ± 19.39	55.67 ± 20.57	<0.001
Season	spring	4386 (16.71 %)	1512 (12.63 %)	2631 (20.81 %)	243 (14.9 %)	<0.001
	summer-fall	11001 (41.92 %)	7533 (62.94 %)	3355 (26.53 %)	113 (6.93 %)	
	winter	10857 (41.37 %)	2924 (24.43 %)	6658 (52.66 %)	1275 (78.17 %)	
Time	dawn	6534 (24.9 %)	3450 (28.82 %)	2784 (22.02 %)	300 (18.39 %)	<0.001
	morning	6570 (25.03 %)	2991 (24.99 %)	3175 (25.11 %)	404 (24.77 %)	
	afternoon	6570 (25.03 %)	2627 (21.95 %)	3426 (27.1 %)	517 (31.7 %)	
	night	6570 (25.03 %)	2901 (24.24 %)	3259 (25.78 %)	410 (25.14 %)	
Wind direction	E	1300 (4.95 %)	864 (7.22 %)	368 (2.91 %)	68 (4.17 %)	<0.001
	W	3996 (15.23 %)	1570 (13.12 %)	2146 (16.97 %)	280 (17.17 %)	
	S	549 (2.09 %)	236 (1.97 %)	255 (2.02 %)	58 (3.56 %)	
	N	939 (3.58 %)	574 (4.8 %)	326 (2.58 %)	39 (2.39 %)	
	SE	1454 (5.54 %)	882 (7.37 %)	509 (4.03 %)	63 (3.86 %)	
	SW	4211 (16.05 %)	1682 (14.05 %)	2144 (16.96 %)	385 (23.61 %)	
	NE	6156 (23.46 %)	3222 (26.92 %)	2554 (20.2 %)	380 (23.3 %)	
	NW	7639 (29.11 %)	2939 (24.56 %)	4342 (34.34 %)	358 (21.95 %)	

*Assoc *p*-value is the *p*-value from one-way ANOVA or chi-square test betwen PM10 and each predictor variable.

After these pre-processing of data, the final data consisted of $n = 26,244$ observations on $p = 12$ variables (1 target variable and 11 predictor variables).

3.3. Discretization

Since Bayesian networks for mixed cases containing both continuous and discrete varaibles are very complex and it is difficult to find softwares that can be easily applied, we applied a discrete Bayesian network in this paper. To apply the discrete Bayesian network, the continuous variables fPM10, cPM10, O3, NO2, CO, SO2, temperature, wind.speed, and humidity were discretized as follows. For fPM10, cPM10, O3, and NO2, we categorised them as (good, normal, bad, very bad) following the standards given by Air Korea, and then merged ‘bad’ and ‘very bad’ into ‘bad’ since ‘very bad’ category had very small number of observations. CO, temperature, wind.speed, and humidity were discretized using quartiles Q1, Q2, Q3 as thresholds so that the probability of each category was similar. For SO2, since more than 25% of observed values of SO2 were 0.003 making Q1 = Q2 = 0.003, SO2 was descritized into three categories using Q1 and Q3 as thresholds. Details on discretization

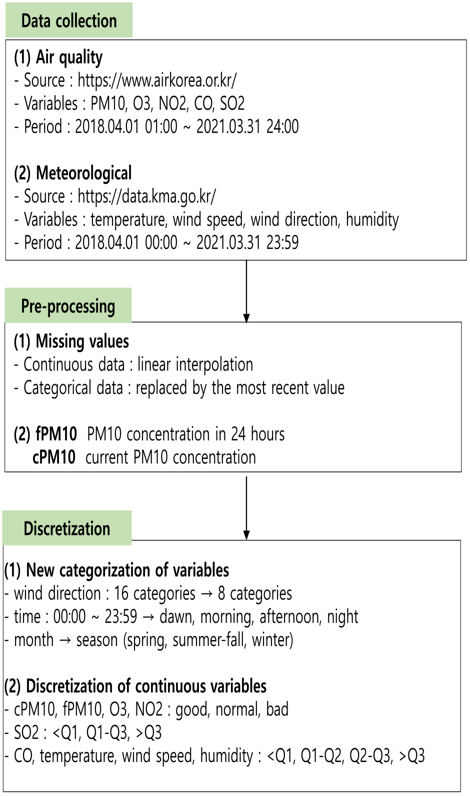
Data collection
(1) Air quality
- Source : https://www.airkorea.or.kr/
- Variables : PM10, O3, NO2, CO, SO2
- Period : 2018.04.01 01:00 ~ 2021.03.31 24:00
(2) Meteorological
- Source : https://data.kma.go.kr/
- Variables : temperature, wind speed, wind direction, humidity
- Period : 2018.04.01 00:00 ~ 2021.03.31 23:59
Pre-processing
(1) Missing values
- Continuous data : linear interpolation
- Categorical data : replaced by the most recent value
(2) fPM10 PM10 concentration in 24 hours
cPM10 current PM10 concentration
Discretization
(1) New categorization of variables
- wind direction : 16 categories → 8 categories
- time : 00:00 ~ 23:59 → dawn, morning, afternoon, night
- month → season (spring, summer-fall, winter)
(2) Discretization of continuous variables
- cPM10, fPM10, O3, NO2 : good, normal, bad
- SO2 : <Q1, Q1-Q3, >Q3
- CO, temperature, wind speed, humidity : <Q1, Q1-Q2, Q2-Q3, >Q3

Figure 5: *Data preparation process.*

criteria for continuous variables are given in Table 1.

Among the final 26,244 observations, fPM10 was ‘good’ in 11,969 cases (45.69%), ‘normal’ in 12,644 cases (48.18%), and ‘bad’ in 1,631 cases (6.21%). Table 2 summarizes basic statistics of each predictor variable by fPM10 status. According to the status of fPM10, the mean ± standard deviation (SD) is given for each continuous variable, and the number of observations and the percentage (%) for each category is given for each discrete variable. The last column of Table 2 shows the p-values from an association tests between fPM10 and predictor variables. One-way ANOVA tests were done for continuous variables and chi-square tests were done for discrete variables. All p-values were smaller than 0.001, showing that there was a significant association between each predictor variable and fPM10.

As the average concentrations of the air quality variables cPM10, O3, NO2, CO, and SO2 increased, fPM10 tended to increase. This implies that future PM10 concentration were positively correlated with current air quality variables. fPM10 concentrations were generally low in summer-fall and high in spring and winter, This phenominon was also reported in Oh and Park (2022) and it may be because of migrating cyclones and dry land surface in spring, and of increased fuel consumption in winter. In addition, the concentration of fPM10 was high when the wind direction was Northwest, Southwest, or West since transboundary air pollutants from China has significant effect on fine dust

Table 3: Classification accuracy and AUC of Bayesian networks

	Classifier		
	NB	TAN	GBN
Accuracy (%)	63.59 ± 0.95	66.65 ± 0.86	65.92 ± 0.75
AUC	0.7577 ± 0.056	0.7748 ± 0.0483	0.7567 ± 0.0518

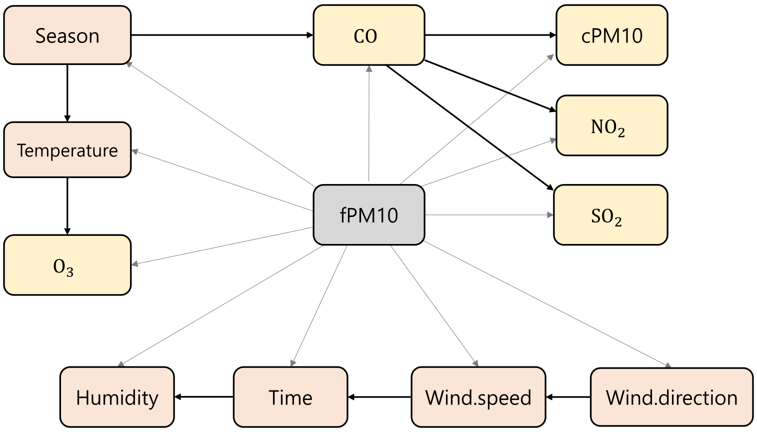

Figure 6: *Bayesian network prediction model for fPM10.*

in Korea (Sohn and Kim, 2015; Kim, 2019). On the other hand, the effect of wind speed on fPM10 was relatively weak in terms of Cramér's V (= 0.0413) which is a measure of association between two categorical variables.

Figure 5 summarizes the whole process of data collection, pre-processing, and discretization.

4. Bayesian network prediction model

4.1. Model selection

We considered the three Bayesian network structures introduced in Section 2 - NB, TAN, GBN - as candidate DAG structures in predicting fPM10. For each of the Bayesian network structure, we repeated 10-fold cross validation 10 times and calculated accuracy and AUC. Accuracy is the relative proportion of observations in the entire data whose predicted and actual values are the same, and AUC represents the area under the receiver operating characteristic (ROC) curve. Table 3 presents the mean and SD of accuracy and AUC for each Bayesian network structure. It can be seen that TAN had the best predictive performance in terms of accuracy and AUC. Thus, TAN was selected as the structure of the Bayesian network.

The final Bayesian network model was obtained from model averaging 100 TAN models from the 10 repetitions of 10-fold cross validation. Figure 6 shows the DAG structure of the final TAN prediction model.

Table 4: Mutual information (MI) scores between PM10 and predictor variables

Variable	MI
cPM10	0.1013
Season	0.0887
CO	0.0615
Temperature	0.0535
SO2	0.0370
NO2	0.0261
Humidity	0.0187
Wind.direction	0.0187
Time	0.0047
O3	0.0041
Wind.speed	0.0000

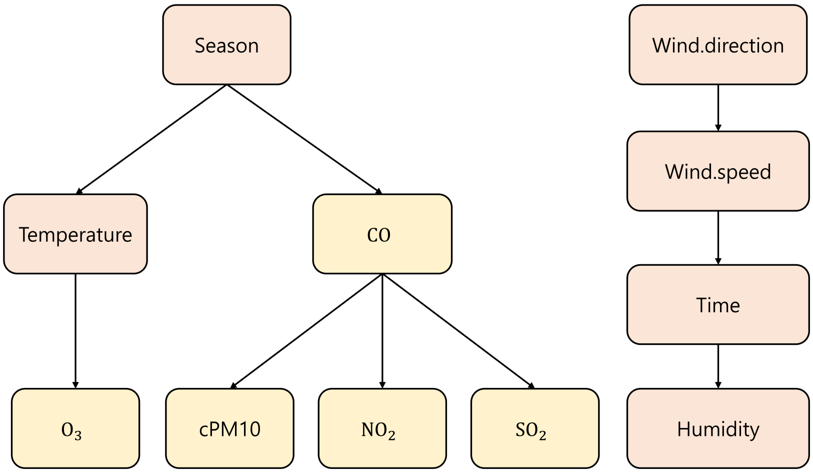

Figure 7: *Relationship between predictor variables in Bayesian network prediction model.*

4.2. Mutual information

We calculated the mutual information (MI) scores between the response variable fPM10 and the predictor variables. MI score is a quantitative measure of the degree of interaction between each node and its parent node in a network. In other words, MI(X, Y) measures the amount of information that predictor variable (X) provides about the target (response) variable (Y). It can be computed from the marginal distributions $P(X = x)$, $P(Y = y)$ and the joint distribution $P(X = x, Y = y)$ of two variables using the formula (4.1) (Cover, 1999),

$$\mathrm{MI}(X, Y) = \sum_{x} \sum_{y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{4.1}$$

Table 4 shows the MI scores between fPM10 and each predictor, computed using the mutinformation function of the R infotheo package.

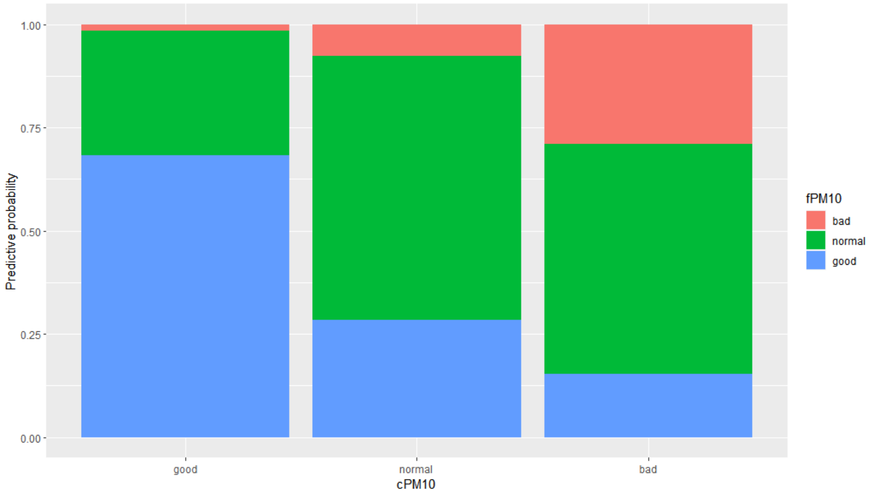

Figure 8: *Predictive probabilities of fPM10 given cPM10.*

The top three predictors with high MI scores are cPM10, season, CO. The variable cPM10 has the largest MI score, showing strong temporal effect on PM10 concentration. The variable season has the second largest MI score. This is in good agreement with the significant seasonal effect shown in Figure 4. Also, the percentages of the levels of 'good' or 'normal' of PM10 concentration in our data is 25 $\sim$ 35% in spring and winter while it is about 68% in summer-fall. The variable CO has the third largest MI score. Previous studies supporting this result are as follows. According to Park *et al*. (2011), PM10/CO (the slope of the regression line of PM10 against CO for 24 hours) and the amount of change in fine dust had a correlation of about 0.31, and PM10/CO was an important factor on the fine dust concentration in Seoul metropolitan area.

It should be noted however that MI score is a measure of a pairwise association between a predictor variable and the target variable, and it does not reflect the effect of the predictor variable when other variables are controlled for. When all the predictor variables are considered together, the order of strength of the effect of variables may differ from the order of MI scores.

4.3. Interpretation of the network

Figure 7 visualizes a tree structure of the relationship between predictor variables in the final Bayesian network prediction model. From the network structure given in Figure 7, we may get useful information on causal relationships from the DAG structure. For instance, Figure 7 shows the well known causal effects of season on temperature, and of temperature on O3 (Yang, 2019; Korean Ministry of Environment, 2017). It also shows the effect of season on CO, and of CO on cPM10.

The two variables directly connected in Figure 7 have an interactive effect on the target variable fPM10, when other variables are adjusted for. For instance, cPM10 and CO have an interactive effect on fPM10. To see this more clearly, we plotted a bar plot of the predictive probability of fPM10 given cPM10 (Figure 8). As expected, as the concentration of cPM10 increases the probabilities of 'normal' and 'bad' for fPM10 increase. We also plotted a bar plot of the predictive probability of fPM10 given cPM10 and CO (Figure 9). It is clearly seen that, the effect of CO on fPM10 varies wildly depending on the level of cPM10 (shown at the top of the figure). This implies a significant interactive effect of cPM10 and CO on fPM10.

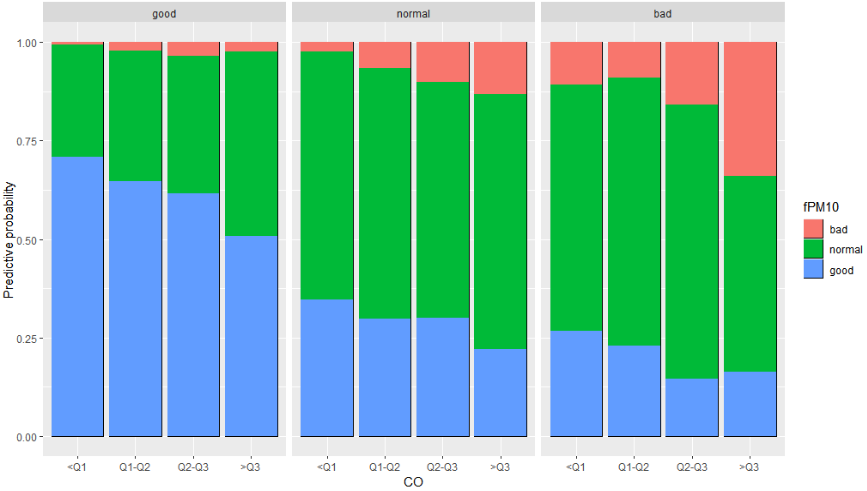

Figure 9: *Predictive probabilities of fPM10 given cPM10 and CO.*

5. Summary and discussion

Fine dust is a very serious environmental problem in Seoul, Korea, a metropolitan city with a population of about 10million. Studies have revealed that fine dust is related to various diseases. Therefore, it is important to predict the concentration of fine dust to minimize exposure to fine dust, and to identify factors affecting fine dust to reduce fine dust. To this end, we developed a PM10 prediction model using a Bayesian network that was trained from air quality measurement data and meteorological data observed in Seoul between 2018 and 2021. A Bayesian network is a DAG-based probabilistic model, and has the advantage of good interpretability because it is possible to identify relationship between variables and discover causative factors using the generated DAG. In this study, we considered three commonly used Bayesian network structures - NB, TAN, GBN - and selected TAN as our prediction model structure based on prediction performance. The final Bayesian network prediction model was obtained from model averaging of TAN models from cross-validations.

The proposed Bayesian network prediction model visualized the relationship between variables in a simple and interpretable way. In terms of MI scores, current PM10 concentration, season, and CO level were highly effective variables in 24 hours ahead prediction of PM10 concetration. The model also showed that there were significant interactive effects among predictor variables in prediction of PM10 concentration.

There exist other studies on predicting PM10 concentration and identifying significant factors. However, many of the previous studies are based on linear or logistic regression models (see Park *et al*., 2017; Kim and Song, 2017; An and Lim, 2020; Son and Kim, 2020; Won and Na, 2021) which assume specific forms of association between the target veriable and predictor variables. For example, the logistic regression model assumes a linear effect of predictor variables on the log odds of the probabilities of the target variable. The Bayesian network model, on the other hand, gives a great flexibility since it does not assume any specific form of predictor effects and it incorporates interactive effects in a natural way (Oh *et al*., 2022).

In this study, we considered three structures of Bayesian network (NB, TAN, GBN) and applied a data-driven methodology to construct a Bayesian network model. Incorporating expert knowledge in selection of predictor variables and/or in construction of a DAG structure may improve the prediction performance and give more practically meaningful model.

We used air quality data observed at one location (Jongno-gu, Seoul), hence care needs to be taken to generalize the results of this studye to other regions of Seoul. Extending the method proposed in this study to handle data from multiple observatories and take into account of regional variables such as traffic volume, number of factories in the region is our future research interest.

References

An S and Lim Y (2020). Forecasting daily PMnullnull concentration in Seoul Jong-no district by using various statistical techniques, *The Korean Data & Information Science Society*, **31**, 187–198.

Ang SL, Ong HC, and Low HC (2016). Classification using the general Bayesian network, *SCIENCE & TECHNOLOGY*, **24**, 205–209.

Cha J and Kim J (2018). Development of data mining algorithm for implementation of fine dust numerical prediction model, *Journal of the Korea Institute of Information and Communication Engineering*, **22**, 595–601.

Cho KW, Jung YJ, Kang CG, and Oh CH (2019). Conformity assessment of machine learning algorithm for particulate matter prediction, *Journal of the Korea Institute of Information and Communication Engineering*, **23**, 20–26.

Cover TM (1999). *Elements of Information Theory*, John Wiley & Sons, New York.

Daly R, Shen Q, and Aitken S (2011). Learning Bayesian networks: Approaches and issues, *The Knowledge Engineering Review*, **26**, 99–157.

Han Y, Heo Y, Hong Y, and Kwon SO (2019). Correlation between physical activity and lung function in dusty areas: Results from the chronic obstructive pulmonary disease in dusty areas (CODA) Cohort, *Tuberculosis and Respiratory Diseases*, **82**, 311–312.

Jang YK (2016). Current status and problems of fine dust pollution, *Journal of Environmental Studies*, **58**, 4–13.

Kim BY, Lim YK, and Cha JW (2022). Short-term prediction of particulate matter (PM10 and PM2.5) in Seoul, South Korea using tree-based machine learning algorithms, *Atmospheric Pollution Research*, **13**, 101547.

Kim DS, Jeong J, and Ahn J (2016). Characteristics in atmospheric chemistry between NO, NO2 and O3 at an urban site during MAPS (megacity air pollution study)-Seoul, Korea, *Journal of Korean Society for Atmospheric Environment*, **32**, 422–424.

Kim SY and Song I (2017). National-scale exposure prediction for long-term concentrations of particulate matter and nitrogen dioxide in South Korea, *Environmental Pollution*, **226**, 21–29.

Kim MJ (2019). The effects of transboundary air pollution from China on ambient air quality, *Heliyon*, **5**, e02953.

Korb KB and Nicholson AE (2010). *Bayesian Artificial Intelligence*, CRC Press, Boca Raton, FL.

Lee JS, Jung YJ, and Oh CH (2020). Seasonal correlation analysis between PM10 and meteorological and air pollutant data, In *Proceedings of Korea Institute of information and Communication Engineering 2020 Autumn General Conference Paper*, 248–250.

Lim JM (2019). An estimation model of fine dust concentration using meteorological environment data and machine learning, *Journal of Information Technology Services*, **18**, 173–186.

Madden MG (2009). On the classification performance of TAN and general Bayesian networks, *Knowledge-Based Systems*, **22**, 489–495.

Oh M and Park CK (2022). Regional source apportionment of PM2.5 in Seoul using Bayesian multivariate receptor model, *Journal of Applied Statistics*, **49**, 738–751.

Oh R, Lee HK, Pak YK, and Oh MS (2022). An interactive online app for predicting diabetes via machine learing from environment-polluting chemical exposure data, *International Journal of Environmental Research and Public Health*, **19**, 1–15.

Park AK, Heo JB, and Kim H (2011). Analyses of factors that affect PM10 level of Seoul focusing on meteorological factors and long range transferred carbon monooxide, *Particle and Aerosol Research*, **7**, 60–68.

Park E, Chang HJ, and Nam HS (2018). A Bayesian network model for predicting post-stroke outcomes with available risk factors, *Frontiers in Neurology*, **9**, 1–11.

Park JH, Yoo SJ, Kim KJ, Gu YH, Lee KH, and Son UH (2017). PM10 density forecast model using long short term memory, In *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, Milan, Italy, 576–581.

Pearl J (2011). Bayesian networks, U of California, Los Angeles, department of statistics papers, Available from: https://escholarship.org/content/qt53n4f34m/qt53n4f34m.pdf

Roh SY (2017). A study on interpolation for the discrete time series of Seoul prime office price, *Journal of Real Estate Analysis*, **3**, 3–4.

Scutari M, Silander T, and Ness R (2022). bnlearn: Bayesian network structure learning, parameter learning and inference, *R package version 4.8.1*.

Sesen MB, Nicholson AE, Banares-Alcantara R, Kadir T, and Brady M (2013). Bayesian networks for clinical decision support in lung cancer care, *PLoS ONE*, **8**, e82349.

Shin DC (2007). Health effects of ambient particulate matter, *Journal of the Korean Medical Association*, **50**, 178.

Shin MK, Lee CD, Ha HS, Choe CS, and Kim YH (2007). The influence of meteorological factors on PM10 concertration in Incheon, *Journal of Korean Society for Atmospheric Environment*, **23**, 322–323.

Sohn KT and Kim D (2015). Development of statistical forecast model for PM10 concentration over Seoul, *Journal of the Korean Data & Information Science Society*, **26**, 291–293.

Son S and Kim J (2020). Evaluation and predicting PM 10 concentration using multiple linear regression and machine learning, *Korean Journal of Remote Sensing*, **36**, 1711–1720.

Witten IH, Frank E, and Hal MA (2011). Data mining: practical machine learning tools and technique, *ACM SIGSOFT Software Engineering Notes*, **36**, 51–52.

Won J and Na J (2021). Prediction of PM 10 in Seoul using penalized regression, *The Korean Data & Information Science Society*, **32**, 631–640.

Yang WH (2019). Changes in air pollutant concentrations due to climate change and the health effect of exposure to particulate matter, *Health and Welfare Forum*, **269**, 21–22.

Yang J and Zhang BT (2000). Analysis of Web Customers Using Bayesian Belief Networks, In *Proceedings of Korean Institute of Intelligent Systems 2000 Autumn Conference Paper*, Jeonju, 387–388.

Yang G, Lee H, and Lee G (2020). A hybrid deep learning model to forecast particulate matter concentration levels in Seoul, South Korea, *Atmosphere*, **11**, 348, Available from: https://doi.org/10.3390/atmos110403489

Korean Ministry of Environment (2014). Knowing the fine dust to protect our health, Available from: https://www.me.go.kr/home/web/board/read.do?menuId=10181&boardMasterId=54&boardCategoryId=&boardId=342997

Korean Ministry of Environment (2017). Ozone, know and prepare properly, Available from: https://www.me.go.kr/home/web/board/read.do?menuId=10181&boardMasterId=54&boardId=790790

Received April 10, 2023; Revised June 24, 2023; Accepted July 21, 2023