Communications for Statistical Applications and Methods
2022, Vol. 29, No. 3, 319–331

https://doi.org/10.29220/CSAM.2022.29.3.319
Print ISSN 2287-7843 / Online ISSN 2383-4757

# Prediction of extreme $PM_{2.5}$ concentrations via extreme quantile regression

SangHyuk Lee[a], Seoncheol Park[b], Yaeji Lim[1,a]

[a]Department of Statistics, Chung-Ang University, Korea
[b]Department of Information Statistics, Chungbuk National University, Korea

**Abstract**

In this paper, we develop a new statistical model to forecast the $PM_{2.5}$ level in Seoul, South Korea. The proposed model is based on the extreme quantile regression model with lasso penalty. Various meteorological variables and air pollution variables are considered as predictors in the regression model, and the lasso quantile regression performs variable selection and solves the multicollinearity problem. The final prediction model is obtained by combining various extreme lasso quantile regression estimators and we construct a binary classifier based on the model. Prediction performance is evaluated through the statistical measures of the performance of a binary classification test. We observe that the proposed method works better compared to the other classification methods, and predicts 'very bad' cases of the $PM_{2.5}$ level well.

Keywords: $PM_{2.5}$, prediction, classification, quantile regression, extreme value theory

## 1. Introduction

Particulate matter of less than 2.5 $\mu m$($PM_{2.5}$) is a critical issues in modern society, and the World Health Organization WHO (2018) has estimated that around 7 million people die every year from exposure to fine particles in polluted air. Numerous studies have been conducted to show the associations between exposure to ambient $PM_{2.5}$ and adverse health effects. Pui *et al.* (2014) reviewed various aspects of $PM_{2.5}$, including its measurement, source apportionment, visibility, and health effects, and mitigation, and Burnett *et al.* (2014) developed a fine particulate mass-based relative risk function. In addition, Song *et al.* (2017) estimated the health burden attributable to $PM_{2.5}$ based on three years of observed data.

South Korea is one of the worst countries in terms of severe air pollution, and many studies have focused on the $PM_{2.5}$ in South Korea. Choi *et al.* (2012) examined the characteristics, sources, and distributions of $PM_{2.5}$ and carbonaceous species in Incheon, South Korea, and Ryou *et al.* (2018) summarized the findings of PM source apportionment studies on South Korea. More recently, Bae *et al.* (2020) estimated long-term foreign contributions to the $PM_{2.5}$ concentrations in South Korea with a set of air quality simulations.

Various statistical models have been applied to the data to predict $PM_{2.5}$ level. Ordieres *et al.* (2005) compared three different topologies of neural networks for predicting average $PM_{2.5}$ concentrations: multilayer perceptron (MLP), radial basis function (RBF) and square MLP. Dong *et al.* (2009)

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2020 and the National Research Foundation of Korea (NRF) funded by the Korean government (NRF-2021R1A2B5B01001790, NRF-2021R1F1A1064096).

[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: yaeji.lim@gmail.com

Published 31 May 2022 / journal homepage: http://csam.or.kr

© 2022 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

proposed a model based on hidden semi-Markov models for high $PM_{2.5}$ concentration value prediction, and Qiao *et al.* (2019) developed a model based on wavelet transform–stacked autoencoder–long short term memory (LSTM).

These models perform well in the relatively usual event, but they work poorly to predict extremely high levels of $PM_{2.5}$. Since extreme events are rarely occurred, few data are used to learn the extreme patterns of the $PM_{2.5}$ level. Therefore, conventional statistical models does not perform well in predicting extreme events, if there is no assumption for such events in the models.

Various models have been developed to predict extreme values to overcome this issue. D'Amico *et al.* (2015) advanced the generalized Pareto distribution to model the probability distribution function's tail to predict wind speed in Alaska. Quintela-del-Rı and Francisco-Fernández (2011) proposed a nonparametric functional data analysis to estimate ozone data in the UK, and Schaumburg (2012) combined nonparametric quantile regression with extreme value theory. However, few studies have been conducted to predict extreme PM2.5 levels (Qin *et al.*, 2015).

In this paper, we adapt the three-stage model proposed by Wang and Li (2013) to predict extreme concentrations of $PM_{2.5}$ in Seoul, South Korea. The three-stage model extends the model by Wang *et al.* (2012a) that relaxed the assumptions of linear quantile functions of $Y$ and tail equivalency across covariates, $\boldsymbol{x}$. Wang and Li (2013) integrated quantile regression and extreme value theory by estimating intermediate conditional quantiles using quantile regression and extrapolating these estimates to tails based on the extreme value theory.

In this paper, we modify the three-stage model with lasso regression (Tibshirani, 1996) to improve the prediction performance of the extreme concentration of $PM_{2.5}$. Lasso regression performs variable selection, and provides a sparse solution. Therefore, even when numerous predictors are included in the initial model stage, penalized regression methods, such as lasso, provide a parsimonious linear model. There are various penalized quantile regression models (Wu and Liu, 2009; Alhamzawi *et al.*, 2012; Wang *et al.*, 2012b), and we consider the lasso quantile regression in this study. We further combine the three-stage models with the lasso penalty obtained from the various extreme quantile values, and generate a single prediction value. Finally, a binary classifier based on the proposed model is constructed.

The rest of the paper is organized as follows. The data description and exploratory data analysis are provided in Section 2. In Section 3, we propose a binary classifier based on a three-stage model with lasso regression, and an algorithm for implementation is presented in Section 4. In Section 5, we apply the proposed method to the $PM_{2.5}$ in Seoul, South Korea, and validate the results. Finally, the concluding remarks are presented in Section 6.

## 2. Data

Table 1 lists the variables used in this study. The meteorological variables, hourly precipitation, temperature, wind speed, and humidity, were collected from the South Korea Meteorological Administration. Several studies indicate that the concentration level of $PM_{2.5}$ is affected by the meteorological factors. For example, Zhang *et al.* (2015) confirmed the critical role of meteorological parameters in air pollution formation, and Zhang *et al.* (2018) studied the influences of critical meteorological parameters, such as wind and precipitation, on PM concentrations. Therefore, we also included these variables in this study.

Air pollutant data were obtained from AIR KOREA, affiliated with the Korea Environment Corporation. The measurements include the hourly $PM_{2.5}$, $PM_{10}$, $SO_2$, $NO_2$, CO, and $O_3$ content. Stracquadanio *et al.* (2007) studied the correlations between $PM_{2.5}$ and gases (benzene, $O_3$, $SO_2$, $NO_2$, and

Table 1: Variables

| Data | Variable | Interval | Source |
|---|---|---|---|
| Meteorological data | Precipitation<br>Temperature<br>Wind Speed<br>Humidity | Hourly | South Korea Meteorological Administration<br>https://www.kma.go.kr/ |
| Air pollution data | $PM_{2.5}$<br>$PM_{10}$<br>$SO_2$<br>$NO_2$<br>CO<br>$O_3$ | Hourly | AIR KOREA<br>https://www.airkorea.or.kr/ |

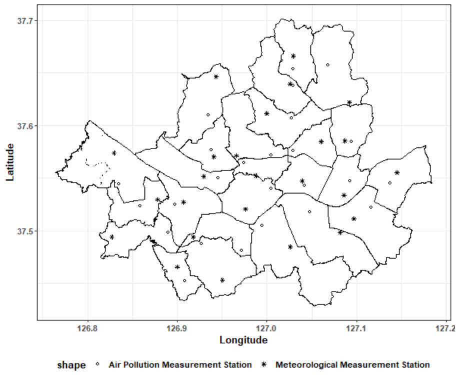
37.7
37.6
37.5
Latitude
126.8
126.9
127.0
127.1
127.2
Longitude
shape ○ Air Pollution Measurement Station * Meteorological Measurement Station

Figure 1: *Variables in 25 districts in Seoul, South Korea, measured at the marked stations.*

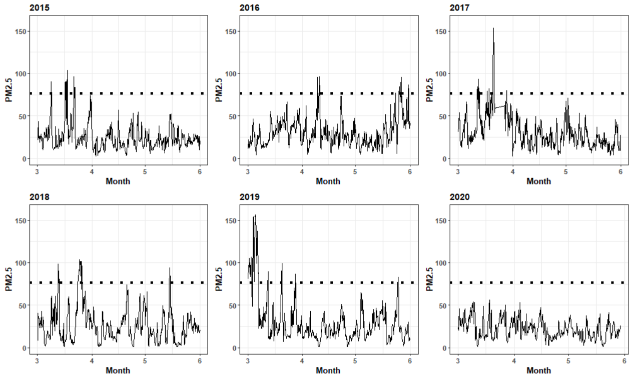

Figure 2: *Four-hour average values of $PM_{2.5}$ from spring 2015 to spring 2020 in the Gangnam district. Dashed horizontal lines indicate 76 $\mu g/m^3$, which is a threshold value rated as 'very bad'.*

CO). Song *et al.* (2015) used the generalized additive model to determine the statistical relationships between $PM_{2.5}$ concentrations and other pollutants, including $SO_2$, $NO_2$, CO, and $O_3$. Based on these studies, we also consider these pollutant variables to be covariates.

The variables in Table 1 are measured in 25 districts in Seoul, South Korea (Figure 1). Differences existed in places where the meteorological variables and air pollutant variables were measured, but the distance between the stations is relatively short.

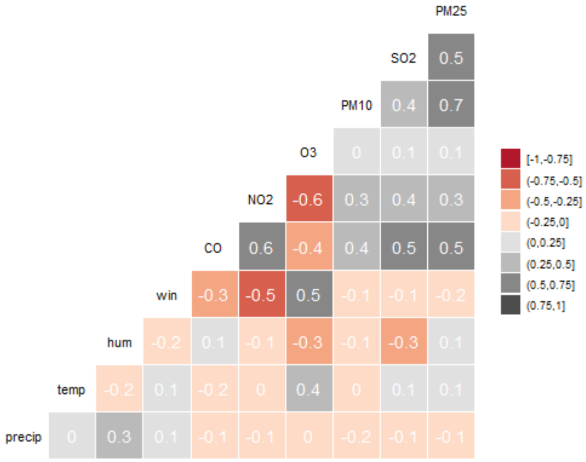

Figure 3: *Correlation matrix between the variables in the Gangnam district.*

Table 2: Category table for $PM_{2.5}$ levels in South Korea

| Daily $PM_{2.5}$ level | Category | | | |
|---|---|---|---|---|
| | Good | Normal | Bad | Very bad |
| ($\mu g/m^3$) | 0 ~ 15 | 16 ~ 35 | 36 ~ 75 | 76 ~ |

All variables were obtained from January 1, 2015, to May 29, 2020, but in this study, we consider only the spring season (March, April, and May) every year, when the extremely high levels of $PM_{2.5}$ are observed. In addition, for the fast computation, we used four-hour average values for all variables. Therefore, the number of time points is 3298 for each district. For example, the four hour average values of $PM_{2.5}$ in the Gangnam district are presented in Figure 2.

Missing values occur in each measurements, thus, we imputed the missing points as follows. First, linear imputation was applied if length of sequential missing points was less or equal to four. Second, if the length of sequential missing points was greater than four and less than or equal to six, the sequence was imputed by the value of the nearest region within 5 km. Finally, the rows were omitted from the data if the previous steps did not impute the points.

To observe the variable correlations, we computed the correlation coefficients. Figure 3 displays the correlation in the Gangnam district, where $PM_{10}$, $SO_2$, and CO have a positive association of more than 0.5 with $PM_{2.5}$, a relatively strong association. However, others have weak relationships with less than 0.5. We also observed relatively strong associations between $O_3$, $NO_2$, CO, and wind speed, implying multicollinearity between variables. For the other districts, we observed a similar variable correlation.

The Ministry of Environment in South Korea categorizes the concentrations of $PM_{2.5}$ levels between 0 and 15 $\mu g/m^3$ as 'good', between 16 and 35 as 'normal', between 36 and 75 as 'bad', and more than 76 as 'very bad' (Table 2). In this study, we focus on the forecasting 'very bad' cases in Seoul that causes severe health effects. The number of time points that categorized based on the $PM_{2.5}$ levels are presented in Table 3 for 25 districts. 'Very bad' cases are rarely observed for all districts, implying that conventional mean based statistical models may not work well in predicting extremely high $PM_{2.5}$ levels.

Table 3: Number of time points (%) in each categorized $PM_{2.5}$ level in 25 districts in Seoul

| District | $PM_{2.5}$ level | | District | $PM_{2.5}$ level | |
|---|---|---|---|---|---|
| | $PM_{2.5} < 76$ | $76 \leq PM_{2.5}$ | | $PM_{2.5} < 76$ | $76 \leq PM_{2.5}$ |
| Gangnam | 3200 (97%) | 98 (3%) | Seodaemun | 3210(97%) | 88(3%) |
| Gangdong | 3229 (98%) | 69 (2%) | Seocho | 3208(97%) | 90(3%) |
| Gangbuk | 3236 (98%) | 62 (2%) | Seongdong | 3191(97%) | 107(3%) |
| Gangseo | 3229 (98%) | 69 (2%) | Seongbuk | 3251(99%) | 47(1%) |
| Gwanak | 3194 (97%) | 104 (3%) | Songpa | 3239(98%) | 59(2%) |
| Gwangjin | 3203 (97%) | 95 (3%) | Yangcheon | 3217(98%) | 81(2%) |
| Guro | 3207 (97%) | 91 (3%) | Yeongdeungpo | 3183(97%) | 115(3%) |
| Geumcheon | 3225 (98%) | 73 (2%) | Yongsan | 3213(97%) | 85(3%) |
| Nowon | 3224 (98%) | 74 (2%) | Eunpyeong | 3232(98%) | 66(2%) |
| Dobong | 3244 (98%) | 54 (2%) | Jongno | 3215(97%) | 83(3%) |
| Dongdaemun | 3221 (98%) | 77 (2%) | Jung | 3243(98%) | 55(2%) |
| Dongjak | 3211 (97%) | 87 (3%) | Jungnang | 3211(97%) | 87(3%) |
| Mapo | 3185 (97%) | 113 (3%) | | | |

## 3. Methodology

The proposed method is based on the three-stage model by Wang and Li (2013) with some modifications. We consider the response variable $Y$ and covariate vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ with the $X_1 = 1$, and assume that we observe a random sample $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ of the random vector $(Y, \boldsymbol{X})$. The goal is to estimate the extremely high conditional quantiles for $\tau_n \to 1$ as $n \to \infty$,

$$Q_Y(\tau_n|\boldsymbol{x}) := \inf\{y : F_Y(y|\boldsymbol{x}) \geq \tau_n\}, \tag{3.1}$$

where $F_Y(\cdot|\boldsymbol{x})$ is the conditional cumulative distribution function of $Y$.

Similar to Wang and Li (2013), we assume that $F_Y(\cdot|\boldsymbol{x})$ is in the maximum domain of attraction of an extreme value distribution $G_{\gamma(\boldsymbol{x})}$, where $\gamma(\boldsymbol{x}) > 0$ is the extreme value index. For the random sample $Z_1, \ldots, Z_n$ from $F_Y(\cdot|\boldsymbol{x})$, constants $a_n(\boldsymbol{x}) > 0$ and $b_n(\boldsymbol{x}) \in \mathbb{R}$ exist such that,

$$P\left(\frac{Z_{(n)} - b_n(\boldsymbol{x})}{a_n(\boldsymbol{x})} \leq y\right) \to G_{\gamma(x)}(y) = \exp\left\{-(1 + \gamma(\boldsymbol{x})y)^{-\frac{1}{\gamma(x)}}\right\}, \quad \text{as } n \to \infty, \tag{3.2}$$

where $Z_{(n)}$ is the largest order statistic of the samples.

Although the three-stage model uses conventional quantile regression as a base model, we propose regularized quantile regression using the lasso penalty to overcome the multicollinearity problem and improve the prediction performance.

The entire procedure consists of four steps. First, the power transformation parameter $\lambda$ of response variable $Y$ is estimated. Secondly, the conditional intermediate quantiles are fitted to the transformed response variable. Then, the extreme quantile is estimated by extrapolating the intermediate quantile estimates and we ensemble the estimates using a simple average. Finally, the result is divided according to the threshold into two groups: ‘not very bad’ and ‘very bad’. The following section details the description of the procedure.

### 3.1. Power transformation

In quantile regression, conditional quantiles of $Y$ are assumed to be linear in $\boldsymbol{x}$ at the tails. To relax this linearity assumption, Wang and Li (2013) considered the power transformation of $Y$. The power-transformed quantile regression model is defined as follows,

$$Q_{\Lambda_\lambda(Y)}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\theta}(\tau), \qquad \text{for } i = 1, \ldots, n, \tag{3.3}$$

where $\tau \in [1 - \epsilon, 1]$, where $\epsilon$ is a small positive constant, and $\theta(\tau)$ is the $\tau$-th quantile regression coefficient.

$$\Lambda_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(y), & \text{if } \lambda = 0. \end{cases} \tag{3.4}$$

The power transformation parameter $\lambda$ in (3.4) is estimated as follows,

$$\hat{\lambda} = \arg\min_{\lambda} \sum_{i=1}^{n} \left\{R_n\left(x_i, \lambda; \tau\right)^2\right\}, \tag{3.5}$$

where $R_n(t, \lambda; \tau) = 1/n \sum_{i=1}^n \mathbf{I}(\boldsymbol{x}_i \leq t)[\tau - \mathbf{I}\{\Lambda_\lambda(y_i) - \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}}^{\text{LASSO}}(\tau; \lambda) \leq 0\}]$, and $\mathbf{I}$ is the indicator function. The estimated lasso coefficient, $\hat{\boldsymbol{\theta}}^{\text{LASSO}}(\tau; \lambda)$, is computed as follows,

$$\hat{\boldsymbol{\theta}}^{\text{LASSO}}(\tau; \lambda) = \arg\min_{\boldsymbol{f}=(b_1,\ldots,b_p)^T} \sum_{i=1}^{n} \rho_\tau\left(\Lambda_\lambda(y_j) - \boldsymbol{x}_i^T\boldsymbol{f}\right) + \nu \sum_{l=1}^{p} \left|b_j\right|, \tag{3.6}$$

where $\rho_\tau(x) = x\{\tau - \mathbf{I}(x < 0)\}$ is the $\tau^{th}$ quantile loss function, and $\nu$ is a penalty parameter in lasso regression estimated at each $\lambda$ through using the cross-validation method (Tibshirani, 1996; Wu and Liu, 2009). Wang and Li (2013) suggested using the upper quantile level $\tau = 1 - \epsilon$ with small positive constant $\epsilon$, and we set $\tau = 0.95$.

## 3.2. Estimating intermediate quantiles

In this step, we estimate the intermediate quantiles $Q_Y(\tau_j|\boldsymbol{x})$ for $\tau_j = j/(n+1)$, $j = 1, \ldots, m$, with $m = n - [n^\eta]$. Parameter $\eta$ is set as 0.1, as suggested by Wang and Li (2013), and $[x]$ denotes the integer part of $x$.
The estimate of the intermediate quantile is

$$\hat{Q}_Y(\tau_j|\boldsymbol{x}) = \Lambda_{\hat{\lambda}}^{-1}\left\{\boldsymbol{x}^T\hat{\boldsymbol{\theta}}^{\text{LASSO}}\left(\tau_j; \hat{\lambda}\right)\right\}, \tag{3.7}$$

where $\hat{\lambda}$ is a power transformation parameter estimated from (3.5), and

$$\hat{\boldsymbol{\theta}}^{\text{LASSO}}\left(\tau_j; \hat{\lambda}\right) = \arg\min_{\boldsymbol{f}=(b_1,\ldots,b_p)^T} \sum_{i=1}^{n} \rho_{\tau_j}\left(\Lambda_{\hat{\lambda}}(y_i) - \boldsymbol{x}_i^T\boldsymbol{f}\right) + \nu \sum_{l=1}^{p} |b_l|. \tag{3.8}$$

## 3.3. Extrapolating intermediate quantiles to tails

Next, we extrapolated the intermediate quantile estimates to the extreme tails. For $\tau_n \to 1$, we estimated $Q_Y(\tau_n|\boldsymbol{x})$ as follows,

$$\hat{Q}_Y(\tau_n|\boldsymbol{x}) = \left(\frac{1 - \tau_{n-k}}{1 - \tau_n}\right)^{\hat{\gamma}_k(\boldsymbol{x})} \hat{Q}_Y(\tau_{n-k}|\boldsymbol{x}), \tag{3.9}$$

where $k = k_n \to \infty$ and $k/n \to 0$.
The extreme value index $\hat{\gamma}_k(\boldsymbol{x})$ is estimated using the followings,

$$\hat{\gamma}_k(\boldsymbol{x}) = \frac{1}{k - [n^\eta] + 1} \sum_{j=[n^\eta]}^{k} \log \frac{\hat{Q}_Y(\tau_{n-j}|\boldsymbol{x})}{\hat{Q}_Y(\tau_{n-k}|\boldsymbol{x})}. \tag{3.10}$$

The selection of $k$ is a crucial part of the three-stage model, and we applied a selection procedure, a modified version of Section 3.3 of Wang and Li (2013), by adding a lasso penalty term.
Then, we ensemble the extreme quantile results, $\hat{Q}_Y(\tau_n|\boldsymbol{x})$, for $\tau_n \in \mathbb{E}$, using a simple average,

$$\hat{Q}_{\text{ensemble}} := \frac{1}{|\mathbb{E}|} \sum_{\tau_n \in \mathbb{E}} \hat{Q}_Y(\tau_n|\boldsymbol{x}), \tag{3.11}$$

where $\mathbb{E}$ contains extreme quantile levels, and we set $\mathbb{E} = \{0.950, 0.955, \ldots, 0.990, 0.995, 0.999\}$. $|\mathbb{E}|$ denotes the number of elements in the set, which is equal to 11.

### 3.4. Binary classification procedure

In the last step, the estimated extreme quantile, $\hat{Q}_{\text{ensemble}}$ is classified into two classes based on the threshold value. This study focuses on the prediction of 'very bad' cases of $PM_{2.5}$; thus, we set the threshold as $76\mu g/m^3$, which is defined in Table 2.

$$\hat{Y} = \begin{cases} 1, & \text{if } \hat{Q}_{\text{ensemble}} \geq 76, \\ 0, & \text{if } \hat{Q}_{\text{ensemble}} < 76. \end{cases} \tag{3.12}$$

## 4. Algorithm

We observed that $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$, where $y_i$ is $PM_{2.5}$, and $\boldsymbol{x}_i = (x_{1,i}, \ldots, x_{9,i})^T$ is a covariate vector on the $i$th time. We used nine variables as covariates: $PM_{10}$, $SO_2$, $NO_2$, CO, $O_3$, precipitation, temperature, wind speed, and humidity.
Note that, we have used predicted explanatory variables. Therefore, proposed model can be written as,

$$Y_{i+1} = f(\hat{\boldsymbol{x}}_{i+1|i}), \quad \text{where } \hat{\boldsymbol{x}}_{i+1|i} = g(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i),$$

where $f(\cdot)$ is proposed three-stage model and $g(\cdot)$ is the forecast model based on the exponential smoothing (ETS) algorithm (Hyndman *et al.*, 2008) computed as follows,

$$\hat{\boldsymbol{x}}_{i+1|i} = \alpha \boldsymbol{x}_i + (1 - \alpha)\hat{\boldsymbol{x}}_{i|i-1},$$

where $\alpha$ is the smoothing parameter estimated by maximizing the likelihood.

The number of time points in the data set is 3,298. We have few 'very bad' cases in the overall data; thus, a small test set contains few 'very bad' cases. Therefore, to validate the performance in predicting extreme cases, we have considered splitting the data into training and test sets at a ratio of 1:2. In this study, $n_1$ = 1,099 observations from March 1, 2015, to March 31, 2017, were used for the training data, and the test data are $n_2$ = 2,199 observations obtained from April 1, 2017, to May 29, 2020.

We constructed the proposed model using training data, and validated the model using test data. The training data are denoted as $\{(y_i^{\text{tr}}, \boldsymbol{x}_i^{\text{tr}}), i = 1, \ldots, n_1\}$, and the test data are $\{(y_i^{\text{te}}, \boldsymbol{x}_i^{\text{te}}), i = 1, \ldots, n_2\}$.

For each district in Seoul, we ran the following algorithm.

**Algorithm 1:** Three-stage model with Lasso ensemble

```
for District in {Gangnam,...,Jungnang} do
    for $\tau_e$ in $\mathbb{E}$={0.950, 0.955, ..., 0.990, 0.995, 0.999} do
        [Step 1]
        Find the optimal power transformation parameter $\hat{\lambda}$ from $\Lambda \in \{-1.5, -1.4, \ldots, 1.4, 1.5\}$ using the training data. In this step, a penalty parameter $\nu$ in the lasso regression at each $\lambda$ is estimated using the 10-fold cross-validation technique.
        [Step 2]
        Find the optimal $\hat{k}$ is computed from the equally spaced $\mathbb{K} = [10, 20, \ldots, 110]$ as follow,
            $\hat{k} = \arg\min_{k \in \mathbb{K}} \sum_{i=1}^{n_1} \{\hat{\gamma}_k(\boldsymbol{x}_i^{tr}) - \hat{\gamma}_k^*(\boldsymbol{x}_i^{tr})\}^2,$
        where $\hat{\gamma}_k^*(\boldsymbol{x}) = M_k^{(1)} + 1 - 1/2(1 - (M_k^{(1)})^2/M_k^{(2)})^{-1}$ with
            $M_k^{(l)} = \frac{1}{k - [n_1{}^{\eta}] + 1} \sum_{j=1}^{k-[n_1{}^{\eta}]+1} \left\{\log \frac{\boldsymbol{x}^T \hat{\boldsymbol{\theta}}^{\text{LASSO}}(\tau_j; \hat{\lambda})}{\boldsymbol{x}^T \hat{\boldsymbol{\theta}}^{\text{LASSO}}(\tau_1; \lambda)}\right\}^l, \quad \text{for} \quad l = 1, 2.$
        Here, $\eta = 0.1$.
        [Step 3]
        Let $u := \hat{k} - [n_1{}^{\eta}] + 1$. Then, consider $u$ quantile values, $\tau_1 < \cdots < \tau_u$, equally spaced between $[1 - \hat{k}/(n_1 + 1),\ (n_1 - [n_1^{\eta}])/n_1 + 1]$. Compute
            $\hat{\boldsymbol{\theta}}^{\text{LASSO}}(\tau; \hat{\lambda}) = \arg\min_{\boldsymbol{f}} \sum_{i=1}^{n_1} \rho_\tau(\Lambda_{\hat{\lambda}}(y_i^{\text{tr}}) - \boldsymbol{x}_i^{\text{tr}T} \boldsymbol{f}) + \nu \sum_{l=1}^{p} |b_l|,$
        where the optimal penalty parameter $\nu$ is computed using cross-validation.
        [Step 4]
        For $j = 1, \ldots, u$, estimate intermediate quantiles as
            $\hat{Q}_{Y^{\text{te}}}(\tau_j|\hat{\boldsymbol{x}}^{te}) = \Lambda_{\hat{\lambda}}^{-1}\{\hat{\boldsymbol{x}}^{\text{te}T} \hat{\boldsymbol{\theta}}^{\text{LASSO}}(\tau_j; \hat{\lambda})\},$
        where $\hat{\boldsymbol{x}}^{te}$ is predicted by ETS algorithm, $\hat{\boldsymbol{x}}^{te} = g(\boldsymbol{x}^{\text{tr}})$.
        [Step 5]
        Extrapolate the intermediate quantiles to the extreme tail,
            $\hat{Q}_{Y^{\text{te}}}(\tau_n|\hat{\boldsymbol{x}}^{\text{te}}) = \left(\frac{1 - \tau_1}{1 - \tau_n}\right)^{\hat{\gamma}_{\hat{k}}(\hat{\boldsymbol{x}}^{te})} \hat{Q}_{Y^{\text{te}}}(\tau_1|\hat{\boldsymbol{x}}^{\text{te}}),$
        where
            $\hat{\gamma}_{\hat{k}}(\hat{\boldsymbol{x}}^{te}) = \frac{1}{u} \sum_{j=1}^{u} \log \frac{\hat{Q}_{Y^{\text{te}}}(\tau_j|\hat{\boldsymbol{x}}^{te})}{\hat{Q}_{Y^{\text{te}}}(\tau_1|\hat{\boldsymbol{x}}^{\text{te}})}.$
    end
    Ensemble the extreme quantile results of the test sample,
        $\hat{Q}_{\text{ensemble}} := \frac{1}{|\mathbb{E}|} \sum_{\tau_n \in \mathbb{E}} \hat{Q}_Y(\tau_n|\hat{\boldsymbol{x}}^{\text{te}}).$
    The final binary prediction value for the test data is following,
        $\hat{Y}^{te} = \begin{cases} 1, & \text{if} \quad \hat{Q}_{\text{ensemble}} \geq 76, \\ 0, & \text{if} \quad \hat{Q}_{\text{ensemble}} < 76. \end{cases}$
end
```

# 5. Application

## 5.1. Evaluation metrics

The results were summarized using the sensitivity, specificity, positive predictive value, negative predictive value, and the F-score, which are statistical performance measures for a binary classification

Table 4: Confusion matrix

| | | Prediction | |
|---|---|---|---|
| | | Positive (‘Very bad’ ) | Negative (‘Not very bad’ ) |
| Actual | Positive (‘Very bad’ ) | True positive (TP) | False negative (FN) |
| | Negative (‘Not very bad’ ) | False positive (FP) | True negative (TN) |

test. Based on the confusion matrix in Table 4, the following measures are defined,

- Sensitivity measures the proportion of correctly identified positives (the proportion of ‘very bad’ time points that are correctly identified as ‘very bad’),

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

- Specificity measures the proportion of correctly identified negatives (the proportion of ‘not very bad’ time points that are correctly identified as ‘not very bad’),

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

- The positive predictive value (PPV) is calculated as follows,

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

- The negative predictive value (NPV) is calculated as follows,

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}.$$

- The F-score measures overall accuracy is calculated as follows,

$$\text{F-score} = \frac{\left(1 + \beta^2\right)\text{TP}}{(1 + \beta^2)\,\text{TP} + \beta^2\text{FN} + \text{FP}} = \frac{\left(1 + \beta^2\right) \times \text{PPV} \times \text{Sensitivity}}{(\beta^2 \times \text{PPV}) + \text{Sensitivity}}.$$

  The F-score is distributed from 0 to 1 and controlled by $\beta$, and it is chosen such that sensitivity is considered $\beta$ times as important as the PPV. Considering our purpose is to accurately predict a ‘very bad’ event and the rareness of the event, we set $\beta = 2$ (Sasaki, 2007).

## 5.2. Comparison methods

Three conventional classification models are applied to evaluate the relative performance of the proposed model. As a representative non-parametric ensemble algorithm, the random forest (RF) is known to be robust to outliers and performs well in many classification problems. The MLP and LSTM models are artificial neural networks which are widely used in various applications. Significantly, variants of the LSTM have been used in predicting of $PM_{2.5}$ because it is appropriate for time series data.

The details of each model are summurized in Table 5. The hyperparameters are selected by trial and error. For example, we considered 400, 600, and 800 number of trees, and choose the 800 as the optimal value.

The commonly used threshold value which is used to classify the observation as positive is 0.5. However, the value may be inappropriate, considering the extremeness of the ‘very bad’ events in the data (Zou *et al.*, 2016). Therefore, we considered threshold values from 0.01 to 1 in the test data and choose the result that provides the best F-score (Lakshmi and Prasad, 2014).

Table 5: Hyper parameters and architectures of comparison methods

| | Hyper parameter | Value | | |
|---|---|---|---|---|
| RF | n_estimators | 800 | | |
| | max features | 3 | | |
| MLP | **Hyper parameter** | **Value** | | |
| | Optimizer | Adam | | |
| | Batch_size | 16 | | |
| | Loss function | binary cross entropy | | |
| | Learning rate | 0.001 | | |
| | **Layer type** | **Input** | **Output** | **Activation** |
| | Dense layer | 9 | 64 | ReLU |
| | Dropout layer | 64 | 64 | · |
| | Dense layer | 64 | 1 | Sigmoid |
| LSTM | **Hyper parameter** | **Value** | | |
| | Optimizer | Adam | | |
| | Batch_size | 16 | | |
| | Loss function | binary cross entropy | | |
| | Learning rate | 0.001 | | |
| | Input window | 6 | | |
| | **Layer type** | **Input** | **Output** | **Activation** |
| | LSTM layer | (6,9) | 64 | tanh |
| | Dense layer | 64 | 1 | Sigmoid |

Table 6: Performance table for Guro, Yangcheon, and Yeongdeungpo (Bold indicates best performance)

| District | Method | F-score | PPV | NPV | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Guro | TSLE | **0.768** | **0.463** | **0.997** | **0.920** | **0.962** |
| | RF | 0.545 | 0.241 | 0.992 | 0.797 | 0.912 |
| | MLP | 0.284 | 0.098 | 0.981 | 0.541 | 0.826 |
| | LSTM | 0.331 | 0.137 | 0.981 | 0.514 | 0.887 |
| Yangcheon | TSLE | **0.584** | **0.692** | **0.984** | 0.562 | **0.991** |
| | RF | 0.341 | 0.532 | 0.974 | 0.312 | 0.990 |
| | MLP | 0.122 | 0.028 | 0.869 | **0.712** | 0.072 |
| | LSTM | 0.279 | 0.110 | 0.976 | 0.450 | 0.863 |
| Yeongdeungpo | TSLE | **0.727** | **0.568** | 0.990 | 0.781 | **0.973** |
| | RF | 0.561 | 0.356 | 0.984 | 0.656 | 0.946 |
| | MLP | 0.464 | 0.161 | **0.993** | **0.875** | 0.791 |
| | LSTM | 0.498 | 0.253 | 0.983 | 0.656 | 0.911 |

## 5.3. Results

The results from three districts, Guro, Yangcheon, and Yeongdeungpo, are presented in Table 6. The proposed three-stage model with lasso ensemble (TSLE) method provides the highest PPVs and comparable sensitivity values, and works best according to the F-score. Although the sensitivity values for the MLP are higher than those for the other models on for Yangcheon and Yeongdeungpo, the MLP has the lowest PPVs, which indicates that a false alarm frequently occurs. This outcome implies that the proposed method has relatively balanced performance for predicting 'very bad' events. However, all models perform well in terms of the NPV and specificity, because of the extremely imbalanced data.

In Figure 4, we plot the F-scores from the three-stage model with lasso regression (TSL) before ensemble. The specific quantile result may provide the best result but there is inconsistent. For example, TSL prediction when $\tau_n = 0.990$ has the highest F-score in Guro, whereas the highest value occurs when $\tau_n = 0.999$ in Yangcheon. Therefore, the ensemble technique solves this selection

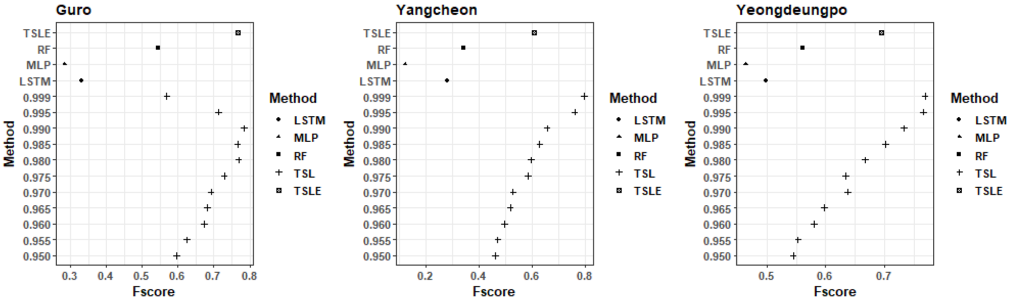

Figure 4: *F-score values for three districts.*

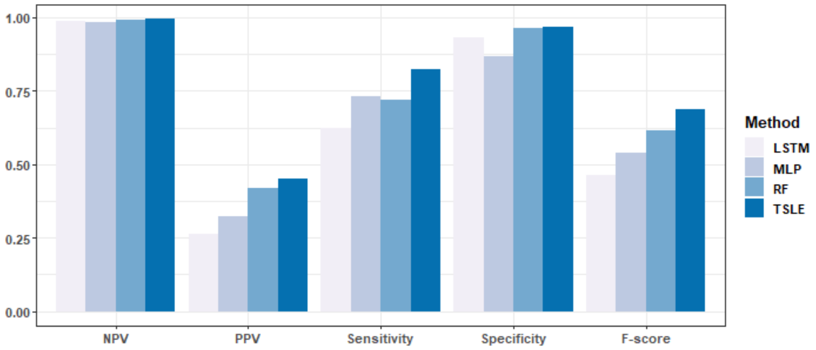

Figure 5: *Average measures for 25 districts.*

Table 7: Proportion of selection for each variable in the three-stage lasso models

| Variable | Proportion of selection |
|---|---|
| Precipitation | 0.664 |
| Temperature | 0.976 |
| Humidity | 0.980 |
| Wind Speed | 0.672 |
| CO | 0.343 |
| $NO_2$ | 0.014 |
| $O_3$ | 0.031 |
| $PM_{10}$ | 1 |
| $SO_2$ | 0 |

problem, and offers superior performance overall.

We obtained similar results for all 25 districts (not shown), and we plotted the average measures of all districts in Figure 5. Overall, the proposed TSLE method works best for all five measures.

As the lasso regression performs the variable selection, some variables are not selected in the regression model. Therefore, we present the proportion of the selection for each variable in Table 7. Temperature, humidity, and $PM_{10}$ were selected in most models, and $NO_2$, $O_3$ and $SO_2$ were rarely selected.

## 6. Conclusion

In this paper, we consider the prediction of extreme values of $PM_{2.5}$ in Seoul, South Korea. Compared to the conventional mean-based models, the proposed method is based on the quantile regression with the extreme value theory. Therefore, the proposed model predicts the extremely high values of $PM_{2.5}$ especially well. Moreover, we added the lasso penalty term to the quantile model, so it performs variable selections. Based on the statistical measures of the performance of a binary classification test, such as the sensitivity, PPV, and F-score, the proposed method works well for 25 districts in Seoul, South Korea.

We expected that the proposed model performance would improve by adding meteorological variables in China, which significantly affect the atmospheric conditions in South Korea. In addition, the data file and R code for implementation are provided at https://github.com/SaeSimcheon/ Extreme-PM2.5-prediction.

## Funding

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2020 and the National Research Foundation of Korea (NRF) funded by the Korean government (NRF-2021R1A2B5B01001790, NRF-2021R1F1A1064096).

## References

Alhamzawi R, Yu K, and Benoit DF (2012). Bayesian adaptive Lasso quantile regression, *Statistical Modelling*, **12**, 279–297.

Bae MA, Kim BU, Kim HC, and Kim ST (2020). A multiscale tiered approach to quantify contributions: A case study of $PM_{2.5}$ in South Korea during 2010-2017, *Atmosphere*, **11**, 141.

Burnett RT, Pope III CA, Ezzati M, *et al.* (2014). An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure, *Environmental Health Perspectives*, **122**, 397–403.

Choi JK, Heo JB, Ban SJ, Yi SM, and Zoh KD (2012). Chemical characteristics of $PM_{2.5}$ aerosol in Incheon Korea, *Atmospheric Environment*, **60**, 583–592.

D'Amico G, Petroni F, and Prattico F (2015). Wind speed prediction for wind farm applications by extreme value theory and copulas, *Journal of Wind Engineering and Industrial Aerodynamics*, **145**, 229–236.

Dong M, Yang D, Kuang Y, He D, Erdal S, and Kenski D (2009). $PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining, *Expert Systems with Applications*, **36**, 9046–9055.

Hyndman R, Koehler AB, Ord JK, and Snyder RD (2008). *Forecasting with Exponential Smoothing: The State Space Approach*, Springer Science & Business Media.

Lakshmi TJ and Prasad Ch SR (2014). A study on classifying imbalanced datasets. In *Proceedings of the 2014 First International Conference On Networks and Soft Computing (ICNSC2014)*, 141–145.

Ordieres JB, Vergara EP, Capuz RS, and Salazar RE (2005). Neural network prediction model for fine particulate matter $PM_{2.5}$ on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua), *Environmental Modelling and Software*, **20**, 547–559.

Pui DYH, Chen S-C, and Zuo Z (2014). $PM_{2.5}$ in China: Measurements, sources, visibility and health effects, and mitigation, *Particuology*, **13**, 1–26.

Qin S, Liu F, Wang C, Song Y, and Qu J (2015). Spatial-temporal analysis and projection of extreme particulate matter ($PM_{10}$ and $PM_{2.5}$) levels using association rules: A case study of the Jing-Jin-Ji region, China, *Atmospheric Environment*, **120**, 339–350.

Qiao W, Tian W, Tian Y, Yang Q, Wang Y, and Zhang J (2019). The forecasting of $PM_{2.5}$ using a hybrid model based on wavelet transform and an improved deep learning algorithm, *IEEE Access*, **7**, 142814–142825.

Quintela-del-Rı A and Francisco-Fernández M (2011). Nonparametric functional data estimation applied to ozone data: Prediction and extreme value analysis, *Chemosphere*, **82**, 800–808.

Ryou HG, Heo JB, and Kim SY (2018). Source apportionment of $PM_1$0 and $PM_{2.5}$ air pollution, and possible impacts of study characteristics in South Korea, *Environmental Pollution*, **240**, 963–972.

Song C, He J, Wu L, *et al.* (2017). Health burden attributable to ambient $PM_{2.5}$ in China, *Environmental Pollution*, **223**, 575–586.

Sasaki Y (2007). The truth of the F-measure, Retrieved May 26th, 2021 from https://www. cs. odu. edu/˜ mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07. pdf

Schaumburg J (2012). Predicting extreme value at risk: Nonparametric quantile regression with refinements from extreme value theory, *Computational Statistics and Data Analysis*, **56**, 4081–4096.

Song YZ, Yang HL, Peng JH, Song YR, Sun Q, and Li Y (2015). Estimating $PM_{2.5}$ Concentrations in Xi'an city using a generalized additive model with multi-source monitoring data, *PLoS One*, **10**, e0142149.

Stracquadanio M, Apollo G, and Trombini C (2007). A Study of $PM_{2.5}$ and $PM_{2.5}$-Associated Polycyclic Aromatic Hydrocarbons at an Urban Site in the Po Valley (Bologna, Italy), *Water, Air, And Soil Pollution*, **179**, 227–237.

Sun Y, Wong AKC, and Kamel MS (2009). Classification of imbalanced data: A review, *International Journal of Pattern Recognition and Artificial Intelligence*, **23**, 687–719.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.

Weissman I (1978). Estimation of parameters and large quantiles based on the k largest observations, *Journal of the American Statistical Association*, **73**, 812–815.

Wu Y and Liu Y (2009). Variable selection in quantile regression, *Statistica Sinica*, 801–817.

Wang HJ, Li D, and He X (2012). Estimation of high conditional quantiles for heavy-tailed distributions, *Journal of the American Statistical Association*, **107**, 1453–1464.

Wang L, Wu Y, and Li R (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension, *Journal of the American Statistical Association*, **107**, 214–222.

Wang HJ and Li D (2013). Estimation of extreme conditional quantiles through power transformation, *Journal of the American Statistical Association*, **108**, 1062–1074.

WHO (2018). 9 out of 10 people worldwide breathe polluted air, but more countries are taking action, Retrieved November 4th, 2021, from https://www.who.int/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action

Zhang H, Wang Y, Hu J, Ying Q, and Hu X-M (2015). Relationships between meteorological parameters and criteria air pollutants in three megacities in China, *Environmental Research*, **140**, 242–254.

Zhang B, Jiao L, Xu G, Zhao S, Tang X, Zhou Y, and Gong C (2018). Influences of wind and precipitation on different-sized particulate matter concentrations ($PM_{2.5}$, $PM_{10}$, $PM_{2.5-10}$), *Meteorology and Atmospheric Physics*, **130**, 383–392.

Zou Q, Xie S, Lin Z, Wu M, and Ju Y (2016). Finding the best classification threshold in imbalanced classification, *Big Data Research*, **5**, 2–8.

*Received October 7, 2021; Revised November 15, 2021; Accepted November 22, 2021*