Communications for Statistical Applications and Methods
2023, Vol. 30, No. 3, 259–272

https://doi.org/10.29220/CSAM.2023.30.3.259
Print ISSN 2287-7843 / Online ISSN 2383-4757

Prediction of sharp change of particulate matter in Seoul via quantile mapping

Jeongeun Lee[a], Seoncheol Park[1,bc]

[a]Department of Information Statistics, Chungbuk National University, Korea
[b]Department of Mathematics, Hanyang University, Korea
[c]Research Institute for Natural Sciences, Hanyang University, Korea

Abstract

In this paper, we suggest a new method for the prediction of sharp changes in particulate matter (PM_{10}) using quantile mapping. To predict the current PM_{10} density in Seoul, we consider PM_{10} and precipitation in Baengnyeong and Ganghwa monitoring stations observed a few hours before. For the PM_{10} distribution estimation, we use the extreme value mixture model, which is a combination of conventional probability distributions and the generalized Pareto distribution. Furthermore, we also consider a quantile generalized additive model (QGAM) for the relationship modeling between precipitation and PM_{10}. To prove the validity of our proposed model, we conducted a simulation study and showed that the proposed method gives lower mean absolute differences. Real data analysis shows that the proposed method could give a more accurate prediction when there are sharp changes in PM_{10} in Seoul.

Keywords: particulate matter, mixture model, generalized additive model, quantile mapping

1. Introduction

In recent years, particulate matter has emerged as a serious atmospheric problem. In Korea, there is a phenomenon of high concentration of particulate matter (PM_{10}), which mainly occurs in spring along with yellow sand. It is known that particulate matter not only contaminates the air, but also affects the human respiratory system, causing coughing, asthma, seizures, and even death, adversely affecting humans (Raaschou-Nielsen *et al*., 2013). To minimize the damage to public health due to air pollution, the weather forecast reports the particulate matter concentration values for 19 regions across Korea along with the weather. Table 1 shows the PM_{10} density classification table used for weather reporting. In Korea, yellow dust (or Asian dust) originating from the deserts of Mongolia in northern China is not a new issue and has been an issue for decades in the springtime. In addition, after the preliminary forecast for particulate matter in August 2013 and the official forecast in February 2014, public awareness of particulate matter and the ultra-particulate matter has increased (Kang and Kim, 2014).

Meanwhile, the causes of particulate matter can be divided into natural causes and artificial causes (fuel combustion, dust from construction sites and roads, etc.). The ministry of environment is making an effort to reduce the amount of ultra-particulate matter by enacting article 18 of the [special act on particulate matter reduction and management] (emergency reduction measures for high concentration particulate matter).

[1] Corresponding author: Department of Mathematics, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Korea. E-mail: pscstat@hanyang.ac.kr

Published 31 May 2023 / journal homepage: http://csam.or.kr

Table 1: PM_{10} density classification according to the ministry of environment Korea

PM_{10} density ($\mu g/m^3$)	≤30	30<PM_{10} ≤80	80<PM_{10} ≤150	PM_{10} > 150
Status	Good	Normal	Bad	Very bad

Table 2: Summary statistics of regional data

Station name (Station number)	Baengnyeong (102)	Seoul (108)	Ganghwa (201)
1st quantile	19.00	25.0	23.00
Median	29.00	39.0	36.00
Mean	38.35	45.1	36.00
3rd quantile	46.00	56.0	43.65
Missing value	7.25%	5.64%	6.71%
Top 5%	97	98	102
Top 1%	165	160	162

The Korea Environment Institute showed that when the daily precipitation increases, the atmospheric cleaning effect of the precipitation occurs, and the occurrence of the high-concentration PM_{10} phenomenon is significantly reduced through a scatter plot of the daily precipitation and the average daily PM_{10} concentration in Seoul (Korea Environment Institute, 2017). Guo *et al*. (2016) investigated the association between rainfall and air quality using a distributed lag non-linear model, which also showed a cleaning effect, resulting in a continuous reduction in particle contamination. Therefore, we consider precipitation as a meteorological variable that affects particulate matter and focus on modeling the change in the concentration of particulate matter according to precipitation in this study.

In the meantime, there have been several studies to predict high-concentration particulate matter in Seoul. Lee *et al*. (2011) investigated the causes of high-concentration particulate matter below PM_{10} and desirable weather conditions in Seoul, Korea based on rear trajectory analysis and cluster analysis in connection with the systematic PM_{10} path. Hur *et al*. (2016) developed a neural network model based on synoptic patterns in several meteorological fields such as geopotential height, air temperature, relative humidity, and wind to provide a statistical reference for the prediction of PM_{10} grades in Seoul.

Since the concentration of particulate matter is greatly affected by the presence or absence of precipitation, the ultimate goal is to implement conditional extreme PM_{10} modeling by studying extreme particulate matter patterns that can consider the influence of weather variables. For example, we want to check whether PM_{10} in the Baengnyeong has a significant impact on Seoul in a few hours. Conditional extreme modeling will be a very challenging problem, and the successful completion of this study will be of great help to citizens who need high-concentration particulate matter forecasts.

Considering high-concentration particulate matter can be interpreted as being interested in the behavioral patterns of observation values corresponding to the end of the particulate matter concentration distribution, and quantile regression (QR) proposed by Koenker and Bassett (1978) as a methodology for modeling can be used. Since the existing quantile regression analysis methodologies assume a linear relationship between the independent variable and the dependent variable, here, we intend to use the quantile generalized additive models (QGAM) of Fasiolo *et al*. (2020) which can consider nonlinear relationships.

A mixture model refers to a probability model for representing several subgroups in the entire population to assume that they exist. In extreme value theory, we are generally interested in singular values of the population, i.e. the values that correspond to the tails of the distribution. To analyze this, an appropriate threshold is set, and values below the threshold are routine distributions such as

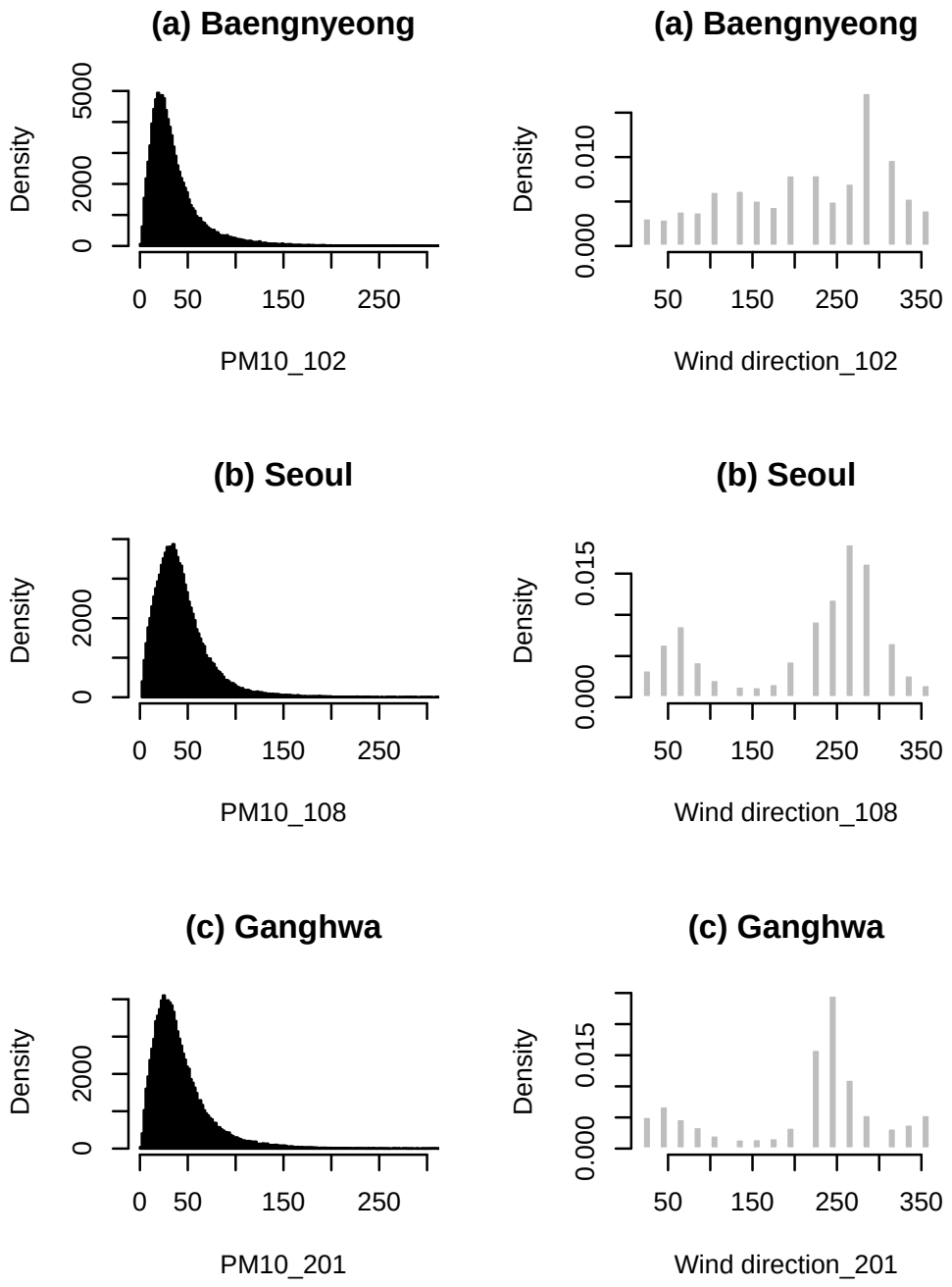

Figure 1: *PM_{10} and wind direction histogram of Baengnyeong, Seoul, and Ganghwa. PM_{10} follows a distribution with a heavy tail on the right and the wind directions are around 250–290 degrees are the most.*

Figure 2: *Location plots of Baengnyeong (Red), Seoul (Green), and Ganghwa (Blue) monitoring stations. The yellow arrow indicates a 280-degree direction based on the Seoul Metropolitan Government's monitoring station.*

normal or gamma. The values above the threshold are also modeled as extreme value distributions such as generalized Pareto.

2. Data

The data used for the analysis are hourly particulate matter data collected from the meteorological data open portal of the Korea meteorological administration. There are three observed points: Baengnyeong, Seoul, and Ganghwa. Baengnyeong is labeled 102, Seoul is 108, and Ganghwa is 201.

Table 2 is a summary of statistics for regional data from July 1, 2008, to June 30, 2020. The average concentration of particulate matter was higher in Seoul, an urban area, than in the other two rural areas. However, it could be confirmed that the top 5% and 1% particulate matter concentrations did not have a relatively large variation depending on the location. Considering that the total amount of data is large and the missing values are random, it was judged that the missing values account for a small proportion, so the analysis was conducted without filling in with other values such as the mean and median.

PM_{10} observed in spring in Korea also follows a distribution with a heavy tail on the right due to the presence of yellow sand. Therefore, it is desirable to use a mixture model when analyzing the concentration of particulate matter in Korea.

To examine the additional relationship between these three points, the wind direction information for the springtime at Baengnyeong, Seoul, and Ganghwa is plotted as a histogram, as shown in Figure 1. As for the mode of wind direction for each location, it can be seen that Baengnyeong is 290 degrees, Seoul is 270 degrees, and Ganghwa is 250 degrees, so the wind directions are around 250–290 degrees are most.

In this regard, if you look at the map of the location of the weather station used in the study printed in Figure 2, it can be seen that the Ganghwa and Baengnyeong weather stations are located between 280 and 285 degrees from the Seoul weather station. Based on Figures 1 and 2, it can be assumed that the events observed at the weather station in the west were similarly observed at the weather station in the east a few hours later in Korea's spring weather phenomena, including particulate matter and precipitation.

Table 3: AIC and BIC values for extreme value mixture model candidates for the PM_{10} model in air quality stations

	Gamma Baengnyeong	Weibull Baengnyeong	Gamma Seoul	Weibull Seoul	Gamma Ganghwa	Weibull Ganghwa
AIC	231639.8	232230.4	231280.5	231370.0	232992.6	233590.6
BIC	231672.2	232262.8	231313.0	231402.4	233025.0	233623.0

Therefore, Baengnyeong and Ganghwa monitoring stations can be used as an indicator of particulate matter concentration for several hours ago in predicting particulate matter in Seoul and its adjacent areas, which have a population of more than 20 million in metropolitan areas. This effect suggests that it is likely to be used as a more efficient predictive indicator in a situation in which the PM_{10} concentration in Seoul changes rapidly. For example, in a situation where the PM_{10} concentration is low and suddenly rises, rather than in a situation in which the PM_{10} concentration in Seoul does not change significantly. Particulate matter data have autocorrelation. For example, if we calculate the autocorrelation coefficient of the Seoul PM_{10} time series, the autocorrelation is quite high even before 6 hours. In other words, if there is no sudden change in the particulate matter concentration, it means the index in Seoul PM_{10} a few hours ago can be a good indicator for predicting the current PM_{10} concentration in Seoul.

However, we want to prove that this conjecture is correct under the assumption that the PM_{10} concentration in Baengnyeong can be a preliminary indicator if the particulate matter concentration in Seoul is very low and suddenly rises vertically after a few hours. Therefore, in this study, we will examine under what circumstances can we currently predict particulate matter in Seoul using data on the concentration and precipitation of particulate matter several hours ago observed at Baengnyeong and Ganghwa meteorological observatory.

3. Methodology

Our method can be broadly divided into two parts. First, the observation data for both Baengnyeong and Seoul are divided into two groups: (1) data for a while with zero precipitation, and (2) data for a while with non-zero precipitation. And when we collect data according to the presence or absence of precipitation for each place, we find the probability distribution that best describes this data.

3.1. Quantile mapping

Quantile mapping was suggested by climatology articles, such as Gudmundsson *et al*. (2012). The main idea of quantile mapping came from the generalization of the probability integral transform, which considered the transformation between a random variable from any continuous distribution and a standard uniform distribution. Suppose that $PM_{10}(B,t)$ ($PM_{10}(S,t)$) is the PM_{10} concentration distribution function of Baeknyeong (Seoul) station, t-hours before. When $t = 0$, it refers to the current PM_{10} density. According to Gudmundsson *et al*. (2012), the main goal of the quantile mapping is finding a transformation function f,

$$\mathrm{PM}_{10}(B,t) = f\left(\mathrm{PM}_{10}(S,t)\right).$$

Suppose that we know distributions of $PM_{10}(B,t)$ and $PM_{10}(S,t)$, denote them as $F_{B,t}$ and $F_{S,t}$, which have each inverse function $F_{B,t}^{-1}$ and $F_{S,t}^{-1}$. Then, the transformation f is defined as

$$\mathrm{PM}_{10}(B,t) = F_{B,t}^{-1}\left(F_{S,0}\left(\mathrm{PM}_{10}(S,0)\right)\right). \tag{3.1}$$

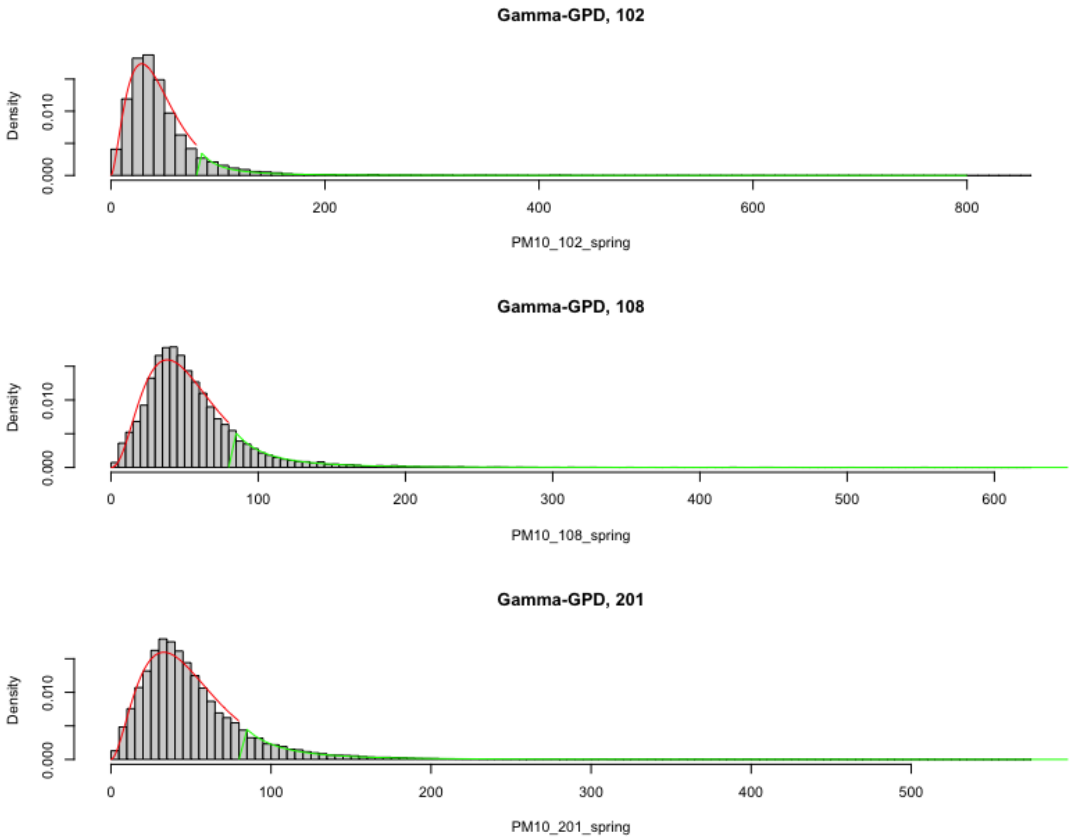

Figure 3: *The extreme mixture modeling of data on Baengnyeong, Seoul, and Ganghwa in spring. Gray bars represent histograms of the empirical data at each station, red lines are the estimated PDF of the bulk model, and green lines are the estimated PDF of the tail model.*

For the relationship of PM_{10} concentrations between Ganghwa and Seoul, we can derive a similar equation. In this paper, based on the idea of quantile mapping, our proposed main idea is that when the number of data is large, it is plausible to change $F_{B,t}$ and $F_{S,t}$ to their estimate based on the data, $\hat{F}_{B,t}$ and $\hat{F}_{S,t}$, respectively.

3.1.1. Extreme value mixture models

When it comes to using quantile mapping for the data analysis, it is crucial how to estimate $F_{\mathrm{B,t}}$ and F_{S}. In this paper, we estimate these cumulative functions using extreme value mixture models. The mixture model referred to here divides the data into non-extreme (called bulk model) and extreme (called tail model) parts to find and model the probability distribution that best describes the probability density for each part.

Let h(·) be the pdf of the bulk model and g(·) be the pdf of the tail model. And let the parameter vectors corresponding to the bulk model and the tail model be $\boldsymbol{\Theta}_b$ and $\boldsymbol{\Theta}_u$, respectively. Then the

Table 4: Parameter estimation results for the gamma-GPD model

	Baengnyeong 102	Seoul 108	Ganghwa 201
Gamma shape	2.7305	3.5044	2.8989
Gamma scale	16.6383	15.2893	17.3669
GPD shape	23.4946	24.8987	25.8975
GPD scale	0.5629	0.3444	0.3530

aforementioned model can be written as

$$F(x \mid \boldsymbol{\Theta}_b, \boldsymbol{\Theta}_u) = \begin{cases} H(x \mid \boldsymbol{\Theta}_b), & \text{for } x \le u, \\ H(u \mid \boldsymbol{\Theta}_b) + [1 - H(u \mid \boldsymbol{\Theta}_b)]\, G(x \mid \boldsymbol{\Theta}_u), & \text{for } x > u. \end{cases} \tag{3.2}$$

At this time, in Equation (3.2), $H(\cdot)$ is the cumulative distribution function (CDF) of the bulk model, and $G(\cdot)$ is the CDF of the tail model. For example, if the bulk model is a gamma distribution and the tail model is a generalized Pareto distribution (GPD), $\boldsymbol{\Theta}_b$ is the gamma distribution parameters, and $\boldsymbol{\Theta}_u$ is the GPD parameters (including u). The probability density function (PDF) version of the Equation (3.2) is:

$$f(x \mid \boldsymbol{\Theta}_b, \boldsymbol{\Theta}_u) = \begin{cases} h(x \mid \boldsymbol{\Theta}_b), & \text{for } x \le u, \\ [1 - H(u \mid \boldsymbol{\Theta}_b)]\, g(x \mid \boldsymbol{\Theta}_u), & \text{for } x > u, \end{cases} \tag{3.3}$$

where h is the corresponding PDF of the bulk model. In this paper, we consider two popular extreme value mixture models: (i) gamma-GPD and (ii) Weibull-GPD. The PDF and CDF of the gamma distribution are:

$$f(x \mid \boldsymbol{\Theta}_b) = \frac{1}{\Gamma(k)\,\lambda^k} x^{k-1} e^{-\frac{x}{\lambda}}, \quad F(x \mid \boldsymbol{\Theta}_b) = \frac{1}{\Gamma(k)} \gamma\left(k, \frac{x}{\lambda}\right), \ x \ge 0, \tag{3.4}$$

where $\boldsymbol{\Theta}_b = (k, \lambda)$, $k > 0$, and $\lambda > 0$ are shape and scale parameters of the gamma distribution, respectively. $\Gamma(\gamma)$ is the (incomplete) gamma function. On the other hand, the PDF and CDF of the Weibull distribution are:

$$f(x \mid \boldsymbol{\Theta}_b) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, \quad F(x \mid \boldsymbol{\Theta}_b) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k}, \ x \ge 0, \tag{3.5}$$

where $\boldsymbol{\Theta}_b = (k, \lambda)$, $k > 0$, and $\lambda > 0$ are shape and scale parameters of the Weibull distribution, respectively. Furthermore, the PDF and CDF of the generalized Pareto distribution are:

$$f(x \mid \boldsymbol{\Theta}_u) = \frac{1}{\sigma}\left(1 + \xi\frac{x-u}{\sigma}\right)^{-\frac{1}{\xi}+1}, \quad F(x \mid \boldsymbol{\Theta}_u) = 1 - \left(1 + \xi\frac{x-u}{\sigma}\right)^{-\frac{1}{\xi}}, \ x \ge u, \tag{3.6}$$

where $\boldsymbol{\Theta}_u = (\xi, \sigma)$, ξ, and $\sigma > 0$ are shape and scale parameters of the generalized Pareto distribution, respectively. We suggest that u in Equations (3.2) and (3.3) can be selected as 150 (bad-very bad boundary) or 80 (normal-bad boundary), which has a special meaning as a threshold. Of course, u

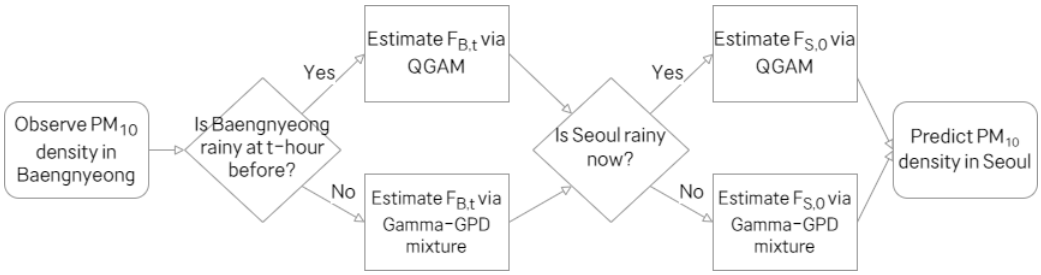

Figure 4: *The diagram of the prediction algorithm.*

can also be estimated, but it is also complicated. Since we focused on extreme particulate matter, we selected "$u = 80$ (normal-bad boundary)" and proceeded with mixture modeling.

In this paper, two famous extreme value mixture models were considered based on the data analysis results: (i) gamma-GPD and (ii) Weibull-GPD. For the extreme value mixture model selection, we calculated AIC and BIC evaluation scales for the gamma-GPD and Weibull-GPD of mixture models. In this case, the nllh (negative log-likelihood) value was obtained during the model fitting process, so the calculation was carried out as in Equation (3.7), and the results are shown in Table 3. Gamma-GPD showed slightly lower values than Weibull-GPD in all three regions of Baengnyeong, Seoul, and Ganghwa. Therefore, we selected gamma-GPD as the final mixture model, and Figure 3 shows the result of the extreme mixture modeling of data on the three regions. Also, Table 4 is the estimated result of gamma-GPD fitted to the data.

$$\text{AIC} = 2k + 2\text{nllh}, \quad \text{and} \quad \text{BIC} = k\ln(n) + 2\text{nllh}. \tag{3.7}$$

3.1.2. Quantile generalized additive model

Since the concentration of particulate matter is expected to change according to the amount of precipitation in a situation where the amount of precipitation is not 0, we expect that more precise modeling can be performed by performing quantile regression analysis by setting the explanatory variable X as the precipitation amount and the response variable Y as the particulate matter concentration. Here, we assume a nonlinear relationship between precipitation and particulate matter concentration and then use the quantile generalized additive model of Fasiolo *et al*. (2020), which can be modeled more flexibly than the conventional quantile regression analysis.

The quantile generalized additive model equation is that:

$$q_\tau\left(\sqrt{\text{PM}_{10,i}} \mid \text{Precipitation}_i\right) = f\left(\sqrt{\text{Precipitation}_i}\right) + \varepsilon_i, \tag{3.8}$$

where $q_\tau(y|x)$ is a τ^{th} conditional quantile function of y given x, and f is a sufficiently flexible function to represent nonlinear behavior of the data, ε_i an independent and identically distributed (i.i.d.) Gaussian error of the i^{th} data. The variable transformation was applied to reduce the range of the x-axis. Log transformation is generally used a lot, but we used square root transformation because there are times when PM_{10} and precipitation are zero in our data. In addition, when analyzing the results, the square was taken again so that there was no error in the calculation.

Quantile generalized additive modeling for this data analysis was summarized as (1) Set a sequence of τ vector, τ, which is from 0.01 to 0.99, by changing 0.01, that is, $\tau = 0.01, 0.02, \ldots, 0.99$. (2) Fit a sequence of quantile generalized additive model, we use `mqgam` function in R `qgam` package. (Fasiolo *et al*., 2020) For estimation, we use a shrinkage version of the cubic regression splines.

3.1.3. Prediction algorithm

We have Baengnyeong data from t-time ago, and we want to predict the concentration of particulate matter in Seoul at present through this. Although we don't know the concentration of particulate matter in Seoul, it is assumed that we know the amount of precipitation and whether it rains at the time. Then the whole algorithm is done in the following form and is visually identifiable in Figure 4.

1. Observe the PM_{10} concentration of Baengnyeong before t hours.
2. Check whether precipitation exists in the data of Baengnyeong before t-time.
   (a) When it is raining in the Baengnyeong area (when there is not much extreme particulate matter), the relative degree of quantiles is measured by performing a quantile generalized additive model between precipitation and particulate matter.
   (b) When it is not raining in the Baengnyeong area (when there is a lot of extreme particulate matter), in this case, the empirical quantile of particulate matter observed in the Baengnyeong area is calculated using the appropriate result with the gamma-GPD model mentioned above.
3. Check whether precipitation exists in the data of the current Seoul.
   (a) When it is raining in the Seoul area (when there is not much extreme particulate matter), the relative degree of quantiles is measured by performing a quantile generalized additive model between precipitation and particulate matter.
   (b) When it is not raining in the Seoul area (when there is a lot of extreme particulate matter), the empirical quantile of particulate matter observed in the Seoul area is calculated using the appropriate estimation result using the gamma-GPD model we mentioned above.
4. Predict the PM_{10} concentration in Seoul under the assumption that the calculated quantiles for each station will remain the same in the Seoul area after a few hours.

4. Simulation study

In this section, we compare the effect of the proposed algorithm via a simulation study. The simulation study aims to check that our algorithm could produce reasonable predictions under similar simulated datasets. In this study, we generate a set of artificial time series. In addition, we assume that there are no rainfall effects on the simulated dataset for simplicity.

We briefly introduce the basic simulated setting. We generate two time series of 500 observations at each iteration, called A and B. Figure 5 shows the scatter plot of one simulated data. In this setting, signal A has a similar role to Baengnyeong station, and signal B has a similar role to Seoul station. Therefore, in the simulation study, we could measure how the information of station A can be affected by the precise prediction of observations in B.

Simulated data are constructed by reflecting similarities of the real dataset. Let $X(A,t)$ be the simulated value at location A and time t. Then, $X(A,t)$ are generated by following equation:

$$X(A,t) = a \times [c + m(t) + \varepsilon(A,t)], \tag{4.1}$$

where $a = 2.5$ is an inflation factor, and $c = 20$ is a constant to generate similar data to real PM_{10}, respectively. $\varepsilon(A,t)$ are generated in two ways: (i) independent and identically distributed (i.i.d.) normal

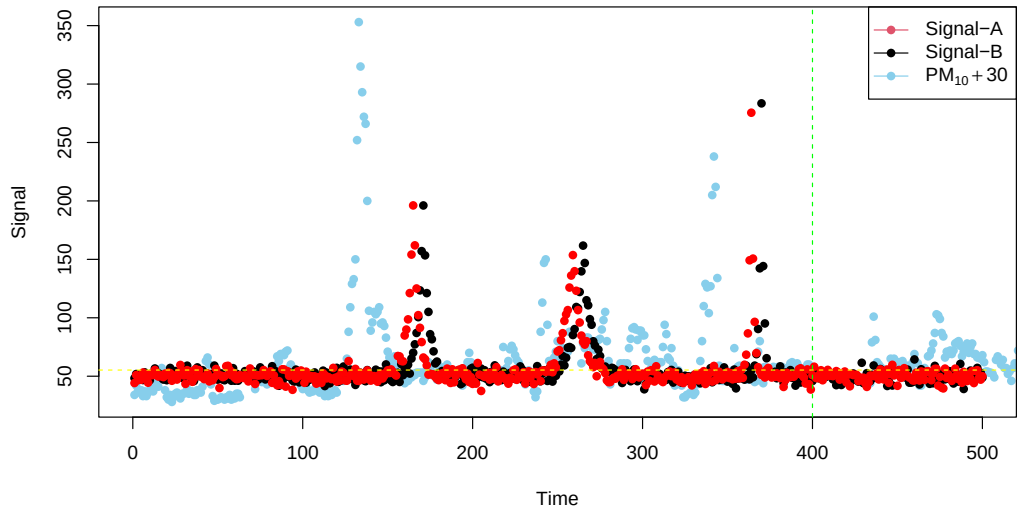

Figure 5: *Time series of the simulated data. Our goal is to make a prediction model of a black signal using $t-6$ hours before signals of red and black. Skyblue dots are the actual time series of Baeknyeong station from April 1, 2018, to April 22, 2018. For comparison, we added 30 for each real data observation. The green dashed line shows indicators to divide the whole times series into train and test sets. The yellow dashed line means the threshold for the construction of a gamma-GPD model.*

random variables with constant variance, i.e., $\varepsilon(A,t) \sim \mathcal{N}(0,\sigma^2)$ or (ii) AR(1) model with parameter 0.5. $m(t)$ is a combination of extreme value signals. Especially, $m(t)$ is generated by Laplace densities:

$$m(t) = \sum_{i=1}^{n_m} m_i(t), \quad m_i(t) = 100\sqrt{f_L(t|,\mu_i, 8b_i)}, \tag{4.2}$$

where $f_L(t|,\mu_i,8b_i)$ is a density function of Laplace distribution with location parameter μ_i and scale parameter $8b_i$. $\mu_1,\mu_2,\ldots,\mu_{n_m}$ are generated by discrete uniform distribution from $[1,500]$ with satisfying $\min_{i\neq j}|\mu_i-\mu_j| \geq 50$. b_i are independently generated from a beta distribution with parameters $(2,5)$. Let $X(B,t)$ be the simulated value at location B and time t. $X(B,t)$ are generated in the same way:

$$X(B,t) = a \times [c + m(t+6) + \varepsilon(B,t)]. \tag{4.3}$$

Although our simulation data generation setting is not intuitive, simulated signals are similar to the original PM_{10} data, shown in Figure 5.

For comparison, we used the conventional vector autoregressive (vector AR) model. Suppose that we have two time series $\{X(A,t)\}$ and $\{X(B,t)\}$. From Woodward *et al.* (2022), the VAR(p) is defined by followings:

$$\begin{pmatrix} X(A,t-6) \\ X(B,t-6) \end{pmatrix} = \begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix} + \begin{pmatrix} \phi_{AA(1)} & \phi_{AB(1)} \\ \phi_{BA(1)} & \phi_{BB(1)} \end{pmatrix} \begin{pmatrix} X(A,t-7) \\ X(B,t-7) \end{pmatrix} + \cdots + \begin{pmatrix} \phi_{AA(p)} & \phi_{AB(p)} \\ \phi_{BA(p)} & \phi_{BB(p)} \end{pmatrix} \begin{pmatrix} X(A,t-6-p) \\ X(B,t-6-p) \end{pmatrix} + \begin{pmatrix} e(A,t-6) \\ e(B,t-6) \end{pmatrix}, \tag{4.4}$$

Table 5: Simulation study results

Data type	Mean (Abs. Diff.) (Std. Dev.)	$n_m = 3$				$n_m = 5$			
		$\sigma = 1.5$	$\sigma = 2$	AR, $\sigma = 1.5$	AR, $\sigma = 2$	$\sigma = 1.5$	$\sigma = 2$	AR, $\sigma = 1.5$	AR, $\sigma = 2$
Whole	Proposed	4.400 (0.2437)	5.7573 (0.2657)	5.0227 (0.3025)	6.6600 (0.5056)	4.3409 (0.2426)	5.7328 (0.2349)	5.0236 (0.3243)	6.6005 (0.3333)
	VAR	5.4883 (1.2144)	6.6399 (1.5398)	13.0132 (2.8471)	14.4584 (2.9995)	5.8915 (1.6927)	7.1600 (1.8666)	14.1104 (3.7483)	14.0732 (2.5950)
Extreme values	Proposed	5.6121 (1.0072)	7.2143 (0.9208)	6.5190 (1.0947)	8.8583 (2.2644)	5.0317 (0.9753)	6.4694 (0.8959)	5.8563 (1.2412)	7.6711 (1.2622)
	VAR	11.5747 (12.0698)	13.5523 (9.3516)	32.4455 (11.5835)	39.5570 (70.9901)	14.5413 (12.0444)	21.3550 (25.4140)	39.5938 (25.4946)	39.1815 (19.4816)
Non-Extreme values	Proposed	4.0940 (0.1920)	5.3891 (0.2437)	4.6439 (0.2602)	6.1833 (0.3914)	4.1659 (0.1907)	5.5471 (0.2387)	4.8143 (0.2619)	6.3345 (0.3396)
	VAR	3.4989 (0.3410)	4.4424 (0.3844)	5.7262 (2.0488)	7.633 (2.0708)	3.7844 (0.5623)	4.8470 (0.4651)	7.1897 (2.0994)	8.1849 (2.0752)

where β_A and β_B are intercepts, ϕ is a time-invariant constant, and e denotes an error term. Our goal is to make a 6-time ahead prediction $\hat{X}(B,t)$. Since it barely works in i.i.d. error setting, we used a modified version of VAR model with additional informaton of $X(B,t),\ldots,X(B,t-5)$ for the prediction in i.i.d. error setting:

$$\begin{pmatrix} X(A,t-6) \\ X(B,t) \end{pmatrix} = \begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix} + \begin{pmatrix} \phi_{AA(1)} & \phi_{AB(1)} \\ \phi_{BA(1)} & \phi_{BB(1)} \end{pmatrix} \begin{pmatrix} X(A,t-7) \\ X(B,t-1) \end{pmatrix} + \cdots + \begin{pmatrix} \phi_{AA(p)} & \phi_{AB(p)} \\ \phi_{BA(p)} & \phi_{BB(p)} \end{pmatrix} \begin{pmatrix} X(A,t-6-p) \\ X(B,t-p) \end{pmatrix} + \begin{pmatrix} e(A,t-6) \\ e(B,t) \end{pmatrix}. \tag{4.5}$$

In this setting, we just do 1-time ahead prediction $\hat{X}(B,t+1)$. For more information about the vector AR model, see Woodward *et al*. (2022). In VAR(p) model, an appropriate selection of time lag p is such a difficult task. In this paper, we consider all VAR(1), VAR(2), ..., VAR(p) predictions and choose the prediction which gives the most similar prediction value compared to true $X(B,t)$ at each t.

In this simulation study, 400 observations are used for model construction, and the remaining 100 observations are used for model performance evaluation using absolute value difference computation.

$$\text{Mean absolute difference} = \frac{\sum_{t=1}^{n_t} \left| X(B,t) - \hat{X}(B,t) \right|}{n_t}, \tag{4.6}$$

where $\hat{X}(B,t)$ is the estimated value from the proposed or VAR approach, and n_t denotes the number of observations in the test set. In the proposed model, we used we set the threshold $u = 0.8$ empirical quantile of observations to distinguish between gamma and GPD mixture. We also use the threshold u to distinguish values in the test set into two parts: Extreme values (observations above u) and non-extreme values (observations below u).

Table 5 summarizes the simulation study results. From the simulation study results, we found that when the variance of the error term is small or the number of extreme value signals is large, then the proposed method gives a lower absolute difference compared to the VAR approach. When we consider non-extreme values in the test set, the VAR method also gives lower mean absolute difference values. However, for extreme values in the test set, the proposed method give good series of predictions. This result supports that the proposed method is a good way to choose when there are a few sharp changes in the series.

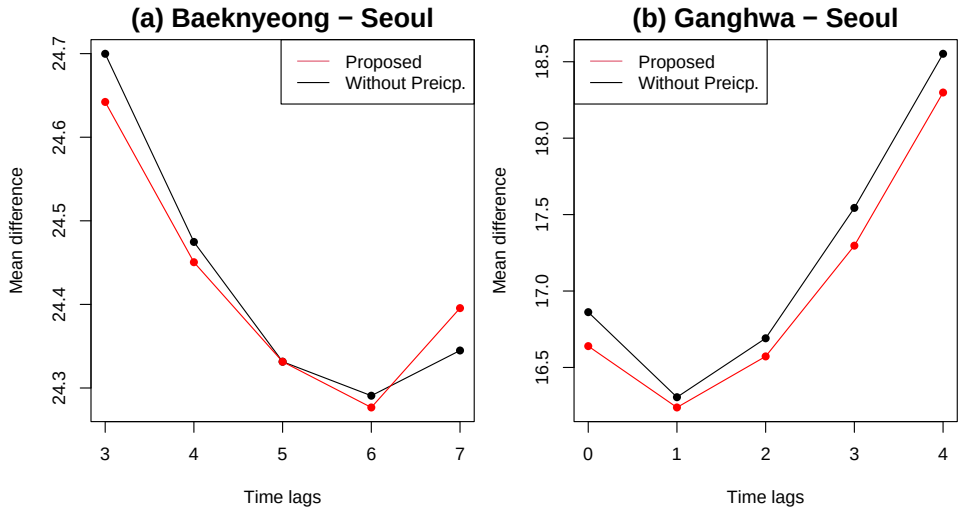

Figure 6: *(a) Prediction error (absolute difference between the predicted value and actual value) according to a time lag of Baengnyeong-Seoul region. (b) Forecast error according to a time lag in the Ganghwa-Seoul region.*

5. Real data analysis

The following results were based on the data of (a) Baengnyeong-Seoul and (b) Ganghwa-Seoul, respectively, and the absolute value of the difference between the predicted and actual values whether the data on particulate matter and precipitation a few hours ago in the front location was the best fit for the prediction of the concentration of particulate matter in the back location.

Figure 6 is the output of prediction error figures according to the time lag change in (a) the Baengnyeong-Seoul region and (b) the Ganghwa-Seoul region. To check whether it is meaningful to make a model separately considering the amount of precipitation, a version without consideration of the presence of precipitation was created. As can be seen from the results, it was the most accurate to predict the current concentration of particulate matter in Seoul using data on particulate matter and precipitation in Baengnyeong 6 hours ago. Similarly, a difference of about 1 hour between Ganghwa and Seoul was the best way to increase the prediction accuracy. On the other hand, in all cases except for the case where the time lag of the Baengnyeong-Seoul region was 7, our model, which was fitted considering the amount of precipitation, presented a result closer to the actual value. Judging from the results, it can be said that it is more reasonable to make a model considering the presence or absence of precipitation.

It is the output of the Figure 7 that the prediction power of (a) Baeknyeong-Seoul (6 hours difference) and (b) Ganghwa-Seoul (1-hour difference). The observed difference shows the difference in particulate matter concentration in Seoul 6 hours before and at the predicted time in (a), and the difference in Seoul particulate matter concentration between 1 hour before and at the predicted time in (b). The red line is the difference between the predicted value and the actual value through our proposed method and represents the average of the difference between the predicted value and the actual value of the proposed method in a situation where the observed difference $\pm 5\mu g/m^3$ calculated and printed. For comparison, we also compute the result of the VAR approach used in Section 4, shown in the green line. If the Seoul particulate matter concentration 6 hours or 1 hour ago is used to predict the Seoul particulate matter concentration at present, the prediction error will be the same as the black

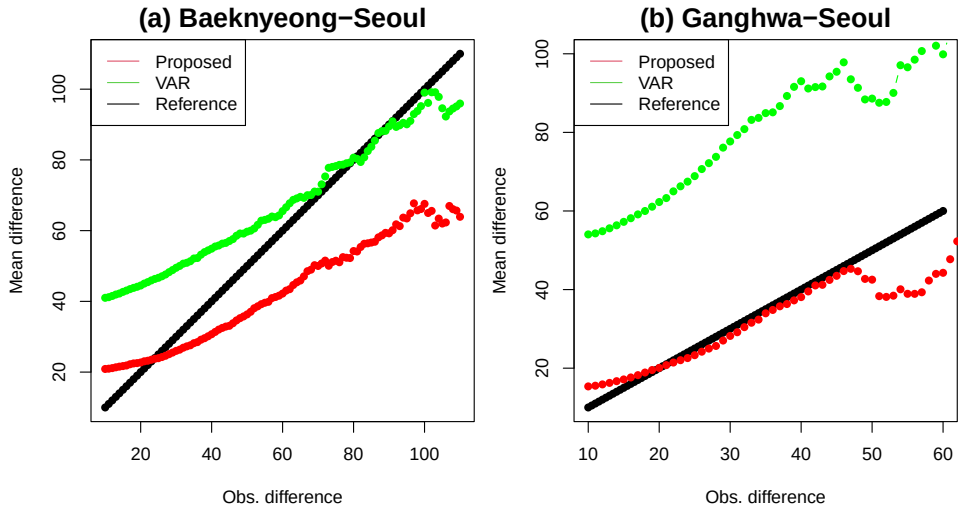

Figure 7: *(a) Prediction error according to the obs mean difference in the Baengnyeong-Seoul region. (b) The figure of prediction error according to obs mean difference of Ganghwa-Seoul region.*

line (reference). When the observed difference is small, that is, when the concentration of particulate matter in Seoul is constant to some extent, the proposed method does not have much benefit, but as this difference increases, the red line in the Figure 7 goes below the reference line. It can be judged that our method is effective as a prediction method when there is a sharp change in the concentration of particulate matter in Seoul.

6. Conclusions and further works

In this paper, we suggested a new and simple algorithm for the prediction of the PM_{10} concentration in Seoul based on the weather and PM_{10} observation result in nearby air condition monitoring stations. The proposed prediction algorithm was based on the application of quantile mapping, which is a generalization of the probability integral transform. To describe the heavy-tail PM_{10} density, we used a gamma-GPD extreme value mixture model. To evaluate the effect of the amount of precipitation on the PM_{10} density, we used the quantile GAM. We showed that the proposed prediction method gives better prediction especially when there are sharp changes of PM_{10} density in Seoul. Through this study, PM_{10} information in the Baengnyeong area can help predict particulate matter in Seoul, and it was possible to measure the effect of precipitation on the concentration of particulate matter. Although the correlation between particulate matter in Baengnyeong 6 hours ago and particulate matter in Seoul was lower than the correlation between particulate matter in Seoul 6 hours ago and particulate matter in Seoul now, this is because the effect is large when there is no rapid change in particulate matter concentration. So it showed that particulate matter from Baengnyeong 6 hours ago can be used as an index for predicting particulate matter concentration in Seoul when there is a rapid change in particulate matter concentration. Likewise, the correlation between particulate matter in Ganghwa an hour ago and particulate matter in Seoul was lower than that of particulate matter in Seoul 1 hour ago, but when there is a sharp change in particulate matter concentration in Seoul, it can be used as an indicator for predicting particulate matter concentration in Seoul.

However, there are some points for a better understanding of the prediction of extreme PM_{10}

in Korea. Since this method only considered PM_{10} densities and the amount of precipitation, other components are also related to the prediction of PM_{10} densities, such as wind speed, urbanization index, etc. So, there are many variations in the data that cannot be explained by this method. In the future, it is necessary to create a more sophisticated high-concentration particulate matter prediction technique in combination with a complex physical model, which we will leave as a further study.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2021R1F1A1064096).

References

Fasiolo M, Wood SN, Zaffran M, Nedellec R, and Goude Y (2020). Fast calibrated additive quantile regression, *Journal of the American Statistical Association*, **116**, 1402–1412.

Gudmundsson L, Bremnes JB, Haugen JE, and Engen-Skaugen T (2012). Technical note: Downscaling RCM precipitation to the station scale using statistical transformations – a comparison of methods, *Journal of the American Statistical Association*, **16**, 3383–3390.

Guo LC, Zhang Y, Lin H, Zeng W, Liu T, Rutherford S, You J, and Ma W (2016). The washout effects of rainfall on atmospheric particulate pollution in two Chinese cities, *Environmental Pollution*, **215**, 195–202.

Hur SK, Oh HR, Ho CH, Kim J, Song CK, Chang LS, and Lee JB (2016). Evaluating the predictability of PM_{10} grades in Seoul, Korea using a neural network model based on synoptic patterns, *Environmental Pollution*, **218**, 1324–1333.

Kang D and Kim JE (2014). Fine, ultrafine, and yellow dust: Emerging health problems in Korea, *Journal of Korean Medical Science*, **29**, 621–622.

Korea Environment Institute (2017). Multi-Faceted analysis of the current state of fine dust concentration, *KEI Focus*, **4**, 6–7. (written in Korean)

Koenker R and Bassett G (1978). Regression quantiles, *Econometrica*, **46**, 33–50.

Lee S, Ho CH, and Choi YS (2011). High-PM_{10} concentration episodes in Seoul, Korea: Background sources and related meteorological conditions, *Atmospheric Environment*, **45**, 7240–7247.

Raaschou-Nielsen O, Andersen ZJ, Beelen R *et al*. (2013). Air pollution and lung cancer incidence in 17 European cohorts: Prospective analyses from the European study of cohorts for air pollution effects (ESCAPE), *The Lancet Oncology*, **14**, 813–822.

Woodward WA, Sadler BP, and Robertson S (2022). *Time Series for Data Science*, Chapman and Hall/CRC, New York.

Received August 11, 2022; Revised December 12, 2022; Accepted March 10, 2023