Understanding Extreme Value Theory and Application to Fine Dust Level in South Korea

Kihyun Han

1. Introduction

The purpose of this research is to 'comprehend Extreme Value Theory and apply it to fine dust level.' In order to accomplish the purpose, we follow these steps; first, build background knowledge concerning Extreme Value Theory (EVT), then apply EVT to the regional fine dust level data, which is dependent and non-stationary with respect to time.

We referred to two books to learn the theory. The first one is "An Introduction to Statistical Modelling of Extreme Values" (I) by Stuart Coles, and we regard this as the main textbook. When the proof was abridged or omitted, we referred to "Extreme Value Theory: An Introduction" (II) by Laurens de Haan to achieve the mathematical rigor.

EVT is a theory about the distribution of extreme data that deviates from the typical average behavior in a given probability distribution. This theory can be applied to a variety of fields that admits a concept of risk for extreme values. Its major topic is to observe the distribution of maxima and minima of data, and it deals with how often the particular value can be attained – for instance, once in ten observations on average. Since the theory basically requires the extrapolation in the data, which entails several error terms, it exhibits how to understand the error with the concept of confidence intervals.

These days, fine dust level is highly relevant to the daily lives of a number of people in South Korea and attracts considerable attention in environmental sciences. Fine dusts are classified, with respect to their particle size, as PM10, having the diameter of less or equal to 10 μm, and PM2.5, having the diameter of less or equal to 2.5 μm – also known as ultrafine dust. We discover the periodicity and the linearity by fitting the daily maximal fine dust level, measured at all observatories in South Korea, and we analyze the nationwide rank of fine dust level – the highest and the lowest provinces – by applying EVT.

2. Theoretical background

(1) Basic Extreme Value Theory

In general, we let M_n be the maximum of the given one-dimensional data $X_1, X_2, \cdots, X_n$ following the same distribution function F. Then the distribution function of M_n is F^n;

$$P(M_n \leq z) = P(X_1 \leq z, \cdots, X_n \leq z) = P(X_1 \leq z)^n.$$

The value above will obviously converge to 0 if z is less than a specific value, and to 1, otherwise. We say that the distribution of M_n degenerates. Now, we consider the modification $M_n^* = \frac{M_n - b_n}{a_n}$, having non-degenerate distribution, for some sequences a_n and b_n. The Theorem 3.1.1 of (I) determines the distribution of M_n^* uniquely for sufficiently large n.

Theorem 3.1.1

If there exist some sequences $\{a_n > 0\}$ *and* $\{b_n\}$ *satisfying*

$$P\left(M_n^* = \frac{M_n - b_n}{a_n} \leq z\right) = F^n(a_n x + b_n) \to G(z) \ \ as \ \ n \to \infty,$$

then G *is an element of the Generalized Extreme Value (GEV) family of distributions, which is defined by*

$$G(z) = exp\left(-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right)$$

where $-\infty < \mu, \xi < \infty$, $\sigma > 0$, *and* z *satisfies* $1 + \xi(z-\mu)/\sigma > 0$. *If* $\xi = 0$, *we understand the value as the limit.*

In (I), the proof of the theorem states, without further explanation, that G is max-stable if and only if G follows the GEV family, and it briefly exhibits the outline of the rest. The theorem corresponds to the Theorem 1.1.3 of (II), which shows the proof using $U(x) = \left(\frac{1}{1-F(x)}\right)^{-1}$. If we let $E(x) = \frac{U(nx) - U(n)}{a_n}$, then $H(x) = E(e^x)$ would be the exponential function, so G has to follow the GEV family.

We denote the return level z_p by the maximum that returns for the probability p, and we let the return period be $1/p$. Then,

$$z_p = G^{-1}(1-p) = \mu - \frac{\sigma}{\xi}\left(1 - y_p^{-\xi}\right)$$

where $y_p = -bg\ (1-p)$. Also, the return level plot is defined by the plot of the return period and the return level lied in x and y axes, respectively. In practice, this plot, together with p-p plot, q-q plot and the likelihood plot, is implemented in the statistical analysis in order to check the goodness of fit.

The shape parameter ξ determines the shape of a GEV distribution. If $\xi > 0$, $G(x) = 0$ for $x < \mu - \sigma/\xi$, and if $\xi < 0$, $G(x) = 0$ for $x > \mu - \sigma/\xi$. According to Smith (1985), the maximum likelihood estimators of μ, σ, ξ have the asymptotic properties for $\xi > 0.5$, whereas they lose the properties for $-1 < \xi < -0.5$ and being not obtainable for $\xi < -1$.

The maximum likelihood estimators (MLE) maximize the log-likelihood function. The log-likelihood function for $Z_1, Z_2, \cdots, Z_m$ of independent and identically distributed as GEV is given as

$$\mathrm{l}(\mu, \sigma, \xi) = -\mathrm{m}\log(\sigma) - (1 + 1/\xi)\sum_{i=1}^{m} bg\ \left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m} bg\ \left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi}$$

where $1 + \xi\left(\frac{z_i - \mu}{\sigma}\right) > 0$ for any $i = 1,2,\cdots,m$. Note that, in general, it is not shown to explicitly find the solution of the score function, the derivative of the equation above, to get the MLE. Thus, the MLE can be approximated as the solution of the score function obtained from the approximation method.

We use two methods for estimating the return level z_p. First, the delta method makes use of the MLE of (μ, σ, ξ). The value of estimator $\widehat{z_p}$ and its variance is given by

$$\widehat{z_p} = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left(1 - y_p^{-\hat{\xi}}\right),\ Var\left(\widehat{z_p}\right) \approx \widehat{\nabla z_p^T} V \widehat{\nabla z_p}$$

where $y_p = -\log(1-p)$. Here, V is the variance matrix of $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$ and

$$\widehat{\nabla z_p} = \left[\partial z_p/\partial\mu, \partial z_p/\partial\sigma, \partial z_p/\partial\xi\right] = \left[1,\ -\ \xi^{-1}\left(1 - y_p^{-\xi}\right),\ \sigma\xi^{-2}\left(1 - y_p^{-\xi}\right) - \sigma\xi^{-1}y_p^{-\xi}\, bg\ \ y_p\right]$$

Is the evaluation at $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$.

The second method is to use the profile likelihood. Note that z_p can be written as a function of (μ, σ, ξ). If we fix z_p as a specific value, then μ can be represented as a function of (σ, ξ) as the following;

$$\mu(\sigma,\xi) = z_p + \frac{\sigma}{\xi}\left[1 - \left(-bg\ (1-p)\right)^{-\xi}\right].$$

By obtaining σ and ξ that maximizes the likelihood and evaluating the likelihood function, we may attain the maximum profile likelihood for a fixed z_p. Then we set z_p and the maximum profile likelihood as, respectively, x and y variable and plot a graph, the x value that maximizes the function would be the estimator $\widehat{z_p}$ attained from the profile likelihood method. Moreover, since the deviance statistic follows $\chi^2(2)$ distribution, we may compute the variance by modifying the maximum profile equation. Since $\widehat{z_p}$ and its variance are different from those attained with delta method, we need to opt for a better estimator under circumstances. For instance, the return level indicates the degree that is repeated in a certain time interval, thus having a higher significance if its value is large. Also, the estimator with a smaller confidence interval is more effective.

(2) Threshold Models

For a specific threshold u, the probability of a random variable X exceeds $u + y$ is given as

$$P(X > u + y | X > u) = \frac{1 - F(u+y)}{1 - F(u)}$$

The following theorem demonstrates that the distribution above can be approximated as the Generalized Pareto distribution for sufficiently large u.

Theorem 4.1.

If X_i are mutually independent and having the identical distribution function F and there exists a function G with

$$P(M_n \le z) \approx G(z) = exp\ \left(-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right),$$

then for sufficiently large u, $P(X > u + y | X > u)$ follows the distribution function

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}$$

where $\tilde{\sigma} = \sigma + \xi(u - \mu)$ and y is defined on $y > 0$ and $1 + \frac{\xi y}{\tilde{\sigma}} > 0$.

(proof) If $F^n(z) \approx exp\ \left(-\left[1 + \xi\left(\frac{\backslash z-\mu}{\sigma}\right)\right]^{-1/\xi}\right)$, we may apply the following asymptotic property for sufficiently large u;

$$n\big(1-F(u)\big) \approx -n\log F(u) = -\log F^n(u) \approx -\log G(u) = \left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}.$$

We apply this to $u+y$ and obtain that

$$n\big(1-F(u+y)\big) \approx \left[1+\xi\left(\frac{u+y-\mu}{\sigma}\right)\right]^{-1/\xi}.$$

Thus, $P(X>u+y|X>u)$ is approximated as a ratio of two values above, which is

$$H(y) = 1-\left(1+\frac{\xi y}{\sigma+\xi(u-\mu)}\right)^{-1/\xi},$$

hence getting Theorem 4.1. by applying $\tilde{\sigma} = \sigma+\xi(u-\mu)$.

The analysis of extreme values in the distribution depends on the threshold value u, so we need to find the optimal u. If a random variable X is larger than u, then the value of $E(X-u|X>u)$ is $\frac{\widetilde{\sigma_u}}{1-\xi}$. Since $\widetilde{\sigma_u}$ is a linear function of u, we may rewrite the value as

$$E(X-u|X>u) = \frac{\widetilde{\sigma_{u_0}}+\xi u}{1-\xi}.$$

for sufficiently large u_0. If we denote the observations that exceeded u as $x_{(1)},\cdots,x_{(n_u)}$, then $E(X-u|X>u)$ can be approximated, in a natural manner, to $\frac{1}{n_u}\sum_{i=1}^{n_u}\big(x_{(i)}-u\big)$. Hence, the mean residual life plot, which is given by the following,

$$\left\{\left(u, \frac{1}{n_u}\sum_{i=1}^{n_u}\big(x_{(i)}-u\big)\right) : u < x_{max}\right\}$$

determines the appropriate threshold u.

In practice, the plot above retains the linearity when u gets large and is no longer linear due to lack of data if u exceeds a specific value. However, we need to take sufficiently large u that the distribution follows Generalized Pareto under $X>u$. Hence, we ought to opt for the optimal threshold u with both having a consistent confidence interval for $E(X-u|X>u) = \frac{\widetilde{\sigma_u}}{1-\xi}$ and taking high u with stable σ^*, the MLE of $\widetilde{\sigma_u}-\xi u$, which is constant if the data follows the Generalized Pareto distribution.

(3) Extremes of Dependent or Non-stationary Sequences

The theories above hold only if each of the observations $X_1, X_2, \cdots$ are mutually independent. From now on, we assume that the independence of the elements of the sequence of random variables is no longer satisfied. To begin with, we may guarantee the independence of random variables that are sufficiently distant if they are stationary and satisfy $D(u_n)$ condition.

Definition 5.1.

If a sequence of stationary random variables $X_1, X_2, \cdots$ holds the following condition, we say that it satisfies $D(u_n)$ condition.

For any $i_1 < \cdots < i_p < j_1 < \cdots < j_q$ with $j_1 - i_p > l$, if we let $A = \{X_{i_1} \le u_n, \cdots, X_{i_p} \le u_n\}$ and $B = \{X_{j_1} \le u_n, \cdots, X_{j_q} \le u_n\}$, then $|P(A \cap B) - P(A)P(B)| \le \alpha(n, l)$ where $\alpha(n, l_n) \to 0$ for some l_n with $\frac{l_n}{n} \to 0$.

Theorem 5.1.

For a sequence of stationary random variables $X_1, X_2, \cdots$, let $M_n = max\{X_1, \cdots, X_n\}$. If there exist a positive real sequence a_n and a real sequence b_n with

$$P\big((M_n - b_n)\backslash a_n\big) \le z) \to G(z) \text{ as } n \to \infty$$

for any real number z for some non-degenerate function G, and $u_n = a_n z + b_n$ satisfies $D(u_n)$, then G is an element of GEV family.

If a sequence of random variables is not stationary, we assume that some of GEVD parameters μ, σ, ξ is not time-homogeneous and follows non-constant function of time. For example, when measuring heights at the east coast of Australia, it is assumed that the maxima $\mu(t) = \beta_0 + \beta_1(t) + \beta_2 SOI(t)$ where $SOI(t)$ means the Southern Oscillation Index at time t. Then each t has distinct GEV distribution, so the return level will change with respect to time.

3. Fine dust level data analysis

(1) Methods

Fine dust level data applied in this research was extracted from AIRKOREA of Korean Environment Corporation as the data of six air pollution matters that are listed in environmental standards. AIRKOREA operates four types of observatories – 'metropolitan' measuring urban areas, 'rural' measuring rural areas, 'background' measuring national background level, and 'street' measuring streets with high traffic. Each observatory measures the average fine dust level

every hour and records it hourly. Among 320 observatories that accumulated data for 1004 days (from the 1st quarter of 2016 to the 3rd quarter of 2018), we excluded the 'street' observatories, which is likely to distract the regional air quality data. Moreover, we additionally excluded the observatories if the number of unrecorded days exceeded seven days. Finally, we applied EVT with daily maxima of 24 fine dust levels corresponding to 24 hours with 177 observatories for PM10 and 85 observatories for PM2.5.

The analysis of fine dust level data utilized the 'extRemes' package of R and referred to the package manual by Eric Gilleland. For the GEVD fitting, we adopted the parameters to maximize the likelihood function and used the delta method.

(2) Optimal Model Finding

We note that the unit of fine dust level (PM10, PM2.5) in this research is 10^{-6} g / m^3.

In general, we are aware of the seasonality of GEVD parameters μ, σ, ξ when fitting an environmental data. Also, the fine dust level exhibits a linear decreasing trend, so we may additionally impose a linear trend to μ and σ. Accordingly, we consider the following six models.

Model 1: μ, σ, ξ are all time-homogeneous
Model 2: $\mu(t) = \mu_0 + \mu_1 \cos\left(\frac{2\pi t}{365} - \mu_2\right)$ and σ, ξ are time-homogeneous
Model 3: $\mu(t) = \mu_0 + \mu_1 \cos\left(\frac{2\pi t}{365} - \mu_2\right), \sigma(t) = \sigma_0 + \sigma_1 \cos\left(\frac{2\pi t}{365} - \sigma_2\right)$ and ξ is time-homogeneous
Model 4: $\mu(t) = \mu_0 + \mu_1 \cos\left(\frac{2\pi t}{365} - \mu_2\right), \sigma(t) = \sigma_0 + \sigma_1 \cos\left(\frac{2\pi t}{365} - \sigma_2\right), \xi(t) = \xi_0 + \xi_1 \cos\left(\frac{2\pi t}{365} - \xi_2\right)$
Model 5: $\mu(t) = \mu_0 + \mu_1 t + \mu_2 \cos\left(\frac{2\pi t}{365} - \mu_3\right), \sigma(t) = \sigma_0 + \sigma_1 \cos\left(\frac{2\pi t}{365} - \sigma_2\right)$ and ξ is time-homogeneous
Model 6: $\mu(t) = \mu_0 + \mu_1 t + \mu_2 \cos\left(\frac{2\pi t}{365} - \mu_3\right), \sigma(t) = \sigma_0 + \sigma_1 t + \sigma_2 \cos\left(\frac{2\pi t}{365} - \sigma_3\right)$ and ξ is time-homogeneous
(Here, t means the number of days from January 1st, 2016)

Model 1 does not consider the seasonality and linearity, so it has a problem of inaccurate fitting with a smaller number of parameters. Generally, the higher likelihood of a model indicates that the model fits the data more effectively. On the contrary, we should also consider that the increased likelihood is caused by the high number of parameters (overfitting). Thus, we compare

AIC (Akaike Information Criterion) – the minus log-likelihood with an additional penalty term of the number of parameters – of all models. AIC is similar to BIC (Bayesian Information Criterion), but BIC selects the true model whereas AIC finds the optimal model that explains the given data. The value of AIC is obtained by $2k - 2\ln(\hat{L})$ where k is the number of parameters and $\hat{L}$ is the maximum log-likelihood.

Comparing the AIC of GEVD fitting with 177 observatories having PM10 data, we evaluate that Model 6 has the minimal AIC in 81 observatories, which is the largest number, thus the optimal model that accounts for PM10 level. Also, in a similar manner, we specify Model 6 to be optimal for PM2.5 as well since it has the minimal AIC in 45 observatories among 85.

Table 1 The model having smallest AIC of PM10 data for each observatory

Model index	1	2	3	4	5	6
Observatories	0	0	2	48	46	81

Table 2 The model having smallest AIC of PM2.5 data for each observatory

Model index	1	2	3	4	5	6
Observatories	0	0	2	48	46	81

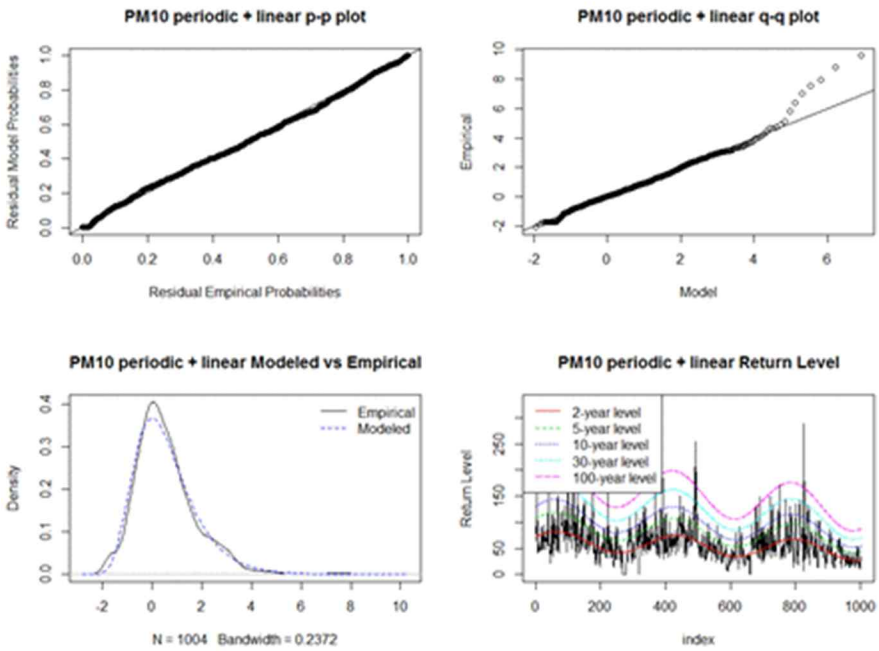


Figure 1 PM10 fitting result of Observatory 111123 (Jongno-gu, Seoul)

Figure 1 shows the p-p plot, q-q plot, density plot, and the return level (from the upper left) of Model 6 fit with PM10 concentration of the 'metropolitan' observatory in Jongno-gu, Seoul (code: 111123). Note that the legend of the return level plot shows the #-day level, not #-year level. Even though we may regard the p-p plot as a well fit following the line, the q-q plot indicates that some

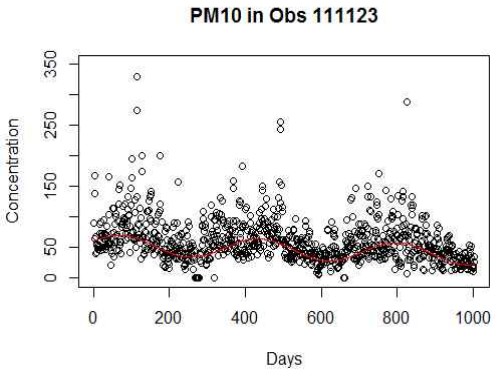
high fine dust level values deviate from the GEVD fitting. These values show the aberrance again in the return level plot (lower right) at some periods as the sharp peaks. Other observatories show similar plots, which we cannot include due to the page limit. The density plot implies inaccurate fitting near the mode. However, it is considered as an inevitable problem in the fitting of 1004 observations that contain high variations.

(3) Parameter Analysis

Model 6 considers the seasonality and the linearity of μ and σ. We observe how the periodic change would be and whether the fine dust level increases or decreases. In order to deal with the average change of the fine dust level, we need to select which value would be considered as the "average." We assume that the average fine dust level at a fixed time is the mode of the GEVD, a point with the highest probability to be sampled, rather than μ. It is well known that the value is given by $\mu + \sigma\frac{((1+\xi)^{-\xi}-1)}{\xi}$.

We analyze the PM10 level data of Jongno-gu, Seoul, which was used above. By showing the GEVD modes from January 1st, 2016 to September 30th, 2018, we obtain that the curve of modes (red curve) has seasonality and linearity among densely marked points. (Figure 2) The curve conforms to the trend of the daily maxima of PM10 level. Hence, we may deduce that excessively high values, which are considered as outliers, do not have any significant influence on fitting the model.

We analyze the average date of the highest fine dust level for each observatory in a year using the seasonal trend of the modes obtained from GEVD fitting. The maximum of modes in PM10 data of Jongno-gu, Seoul is observed at 57th days from January 1st, which is February 27th. By applying the same to all observatories, we obtain Figure 3, which shows that the date that maximizes PM10 level lies between January and April. If we assume that the dates of 177 observatories are independent and identically distributed, we can conclude that the mean of these dates follows the normal distribution by appealing to the central limit theorem. Since the mean and the variance are 55.19 and 178.96, respectively, the 95% confidence interval for the average date of maximal PM10 level is 55.19 ± 26.22 days from January 1st. That is, the highest period of nationwide PM10 level is from early February to late March for the confidence level 95%. A similar analysis shows that the confidence interval for the maximum of nationwide PM2.5 level is 54.8 ± 31.31 days from January 1st, which is from late January to early April.



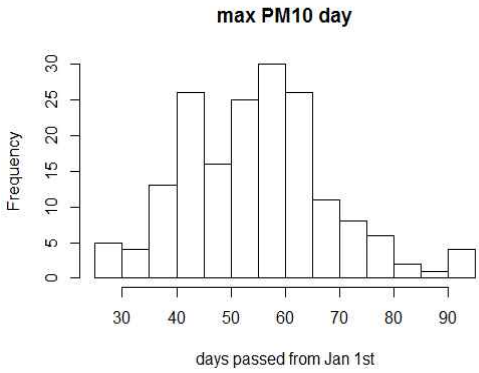
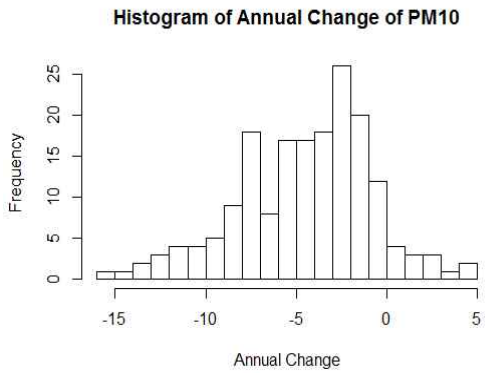


Figure 2 GEVD modes of Jongno-gu, Seoul

Figure 3 the annual ma of PM10 levels

Next, we discuss whether the level increases or decreases with respect to time. The most desirable method is to compare the change of GEVD modes of the fine dust level. Nevertheless, we can check that the annual change of μ in Model 6 – $365\mu_1$ – and the annual change of modes are similar, so we take $365\mu_1$ instead of the modes to analyze the confidence interval. The histogram of the annual change of PM10 level is shown in Figure 4. We can see that the values are largely negative, and by arranging the values in increasing order, we obtain Figure 5. The y-axis indicates the annual change of fine dust level, $365\mu_1$. The curves surrounding the points mean the upper and lower bounds of the 95% confidence interval attained from the GEVD, and there are 140 out of 177 observatories having the 95% confidence interval below 0. In other words, for the confidence level 95%, we can conclude that, in 78.5% of all observatories, the PM10 level decreases over time. Similarly, the PM2.5 level decreased for 95% confidence level in 67 out of 85 observatories, which account for 78.8%.



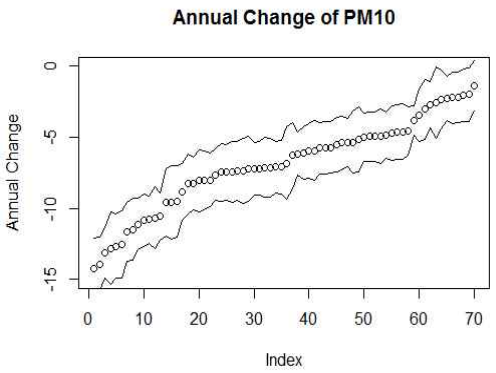


Figure 4 Annual change of PM10 level

Figure 5 Confidence intervals for annual change

(3) Regional Comparison of Fine dust level

We compare the 30 days-return levels of fine dust levels for each region. Note that the 30 days-return level indicates the fine dust level appearing once a month on average. For convenience, we used Model 1, assuming all parameters are time-homogeneous, in searching observatories of the highest and the lowest value and Model 6 in ranking the provinces of South Korea.

First, we discover the highest and the lowest tem observatories. For PM10, the top is Pyeongtaek Port, located in Pyeongtaek-si, Gyeonggi-do, and the 30-days return level is 218.15. There are similarities between the highest 10 observatories that most are located near facilities that emit air pollutants such as factories. The lower bounds of 95% confidence interval of these observatories exceed 150, which means 'Very Unhealthy' according to the classification. The highest level of PM2.5 is observed at Wonsi-dong, Danwon-gu, Ansan-si, Gyeonggi-do, and the value is 127.01. Also, the 95% confidence intervals for the highest 10 observatories exceed 75, which means 'Very Unhealthy' for PM2.5. Therefore, the 30 days-return levels of the highest 10 present 'Very Unhealthy' more than once a month for 95% confidence level.

Table 3 The highest 10 observatories for 30 days-return level of PM10

Rank	Obs. code	Place (Address)	PM10 level	95% Confidence Interval	
1	131343	Pyeongtaek Port, Poseung-eup, Pyeongtaek-si, Gyeonggi-do	218.1572	206.4939	229.8205
2	633311	Maepo-eup, Danyang-gun, Chungcheongbuk-do	214.1353	197.5512	230.7194
3	131232	Sihwa Industrial Complex, Jeongwang-dong, Siheung-si, Gyeonggi-do	197.9961	186.4365	209.5557
4	131551	Namyang-dong, Hwaseong-si, Gyeonggi-do	197.2783	186.0032	208.5535
5	131473	Tongjin-eup, Gimpo-si, Gyeonggi-do	194.8178	185.1088	204.5268
6	131341	Bijeon 1-dong, Pyeongtaek-si, Gyeonggi-do	194.3013	185.3234	203.2792
7	131195	Bugok-dong, Sangnok-gu, Ansan-si, Gyeonggi-do	188.3267	177.8063	198.847
8	131471	Sau-dong, Gimpo-si, Gyeonggi-do	186.1911	177.7178	194.6643
9	238371	Onsan-eup, Ulju-gun, Ulsan	185.3346	176.7786	193.8906
10	131441	Seolseong-myeon, Icheon-si, Gyeonggi-do	183.5511	174.7932	192.309

Conversely, the lowest PM10 observatories are obtained as the following Table 4. The lowest 30 days-return level is attained at Munsu-dong, Yeosu-si, Jeollanam-do, and the 95% confidence interval is from 102.85 to 114.49. We observe that Yeosu-si and Gimhae-si take two seats in the lowest ten. The 95% confidence intervals for the lowest ten observatories lie between 100 and 130, which means 'Unhealthy'. That is, these places experience 'Unhealthy' level once a month on average, and the 'Very Unhealthy' level requires more than a month.

Table 4 The lowest 10 observatories for 30 days-return level of PM10

Rank	Obs. code	Place (Address)	PM10 level	95% Confidence Interval	
1	336125	Munsu-dong, Yeosu-si, Jeollanam-do	108.6732	102.8537	114.4927
2	238181	Dongsang-dong, Gimhae-si, Gyeongsangnam-do	109.4717	105.0491	113.8943
3	336128	Deokchung-dong, Yeosu-si, Jeollanam-do	111.3851	106.4701	116.3001
4	238127	Sinjeong-dong, Nam-gu, Ulsan	115.8922	109.9313	121.8531
5	336133	Suncheonman bay, Suncheon-si, Jeollanam-do	115.9907	110.0761	121.9053
6	422201	Hyeonpung-myeon, Dalseong-gun, Daegu	116.0205	109.8042	122.2368
7	221251	Bugok-dong, Geumjeong-gu, Busan	116.4539	110.8913	122.0165
8	339121	Donghong-dong, Seogwipo-si, Jeju-do	117.5551	111.6098	123.5003
9	238183	Jangyu-dong, Gimhae-si, Gyeongsangnam-do	117.8819	111.1741	124.5898
10	132401	Bangsan-myeon, Yanggu-gun, Gangwon-do	118.8108	110.2888	127.3327

We classify the observatories with the first digit of their codes and obtain eight sections that divide South Korea. The provinces that each number represents are listed as the following. We already derived that PM10 levels of Model 6 are maximized in late February and minimized in late August. Therefore, we considered late February, late May, late August, and late November to obtain the 30 days-return level of PM10. Then we compared the quartiles since the number of observatories is different with each section. We get that the 'Southeast', 'Southwest', and 'East' sections, which are described as 2, 3, and 4, have relatively small PM10 levels, and the 'Capital', 'West', and 'Northwest' described as 1, 7, and 8, have relatively high levels. As an exception, we obtain that the 'Northeast' has smaller PM10 only in late August.

1(Capital): Seoul, Gyeeonggi-do

2(Southeast): Busan, Gyeongsangnam-do, Ulsan

3(Southwest): Jeju-do, Gwangju, Jeollanam-do

4(East): Daegu, Gyeongsangbuk-do

5(Center): Chungcheongbuk-do, Chungcheongnam-do, Daejeon, Sejong

6(Northeast): Gangwon-do, Chuncheongbuk-do (partly)

7(West): Jeollabuk-do

8(Northwest): Incheon, Gyeonggi-do (partly)

Figure 6 Map of South Korea

Table 5 PM10 level ranks in late February / late May / late August / late November

Level	Q1	Q2	Q3
Low	3	4	2
	4	3	4
	2	2	3
	5	5	5
	6	6	6
	8	7	7
High	1	1	1
	7	8	8

Level	Q1	Q2	Q3
Low	3	4	2
	4	3	4
	2	2	3
	5	5	5
	6	6	6
	8	7	7
High	1	1	1
	7	8	8

Level	Q1	Q2	Q3
Low	4	6	6
	6	4	3
	3	3	4
	2	2	2
	5	5	5
	7	7	7
High	1	8	8
	8	1	1

Level	Q1	Q2	Q3
Low	3	4	3
	2	3	2
	4	2	4
	5	5	6
	1	6	5
	6	1	7
High	8	7	1
	7	8	8

[References]

[1] Stuart Coles, An Introduction to Statistical Modeling of Extreme Values

[2] Laurens de Haan, Extreme Value Theory: An Introduction

[3] Eric Gilleland, Package 'extRemes', R document,
https://cran.r-project.org/web/packages/extRemes/extRemes.pdf

[4] AIRKOREA fine dust level data,

https://www.airkorea.or.kr/web/last_amb_hour_data?pMENU_NO=123

[5] Map of South Korea, https://maps-southkorea.com/south-korea-regions-map