

Article

A Novel Tree Ensemble Model to Approximate the Generalized Extreme Value Distribution Parameters of the PM_{2.5} Maxima in the Mexico City Metropolitan Area

Alejandro Ivan Aguirre-Salado ^{1,*}, Sonia Venancio-Guzmán ¹, Carlos Arturo Aguirre-Salado ² 
and Alicia Santiago-Santos ¹

¹ Institute of Physics and Mathematics, Universidad Tecnológica de la Mixteca, Huajuapán de León, Oaxaca C.P. 69000, Mexico; venanciosonia99@gs.utm.mx (S.V.-G.); alicia@mixteco.utm.mx (A.S.-S.)

² Faculty of Engineering, Universidad Autónoma de San Luis Potosí, San Luis Potosí C.P. 78280, Mexico; carlos.aguirre@uaslp.mx

* Correspondence: aleaguirre@mixteco.utm.mx; Tel.: +55-953-532-0214

Abstract: We introduce a novel spatial model based on the distribution of generalized extreme values (GEVs) and tree ensemble models to analyze the maximum concentrations levels of particulate matter with a diameter of less than 2.5 microns (PM_{2.5}) in the Mexico City metropolitan area during the period 2003–2021. Spatial trends were modeled through a decision tree in the context of a non-stationary GEV model. We used a tree ensemble model as a predictor of GEV parameters to approximate nonlinear trends. The decision tree was built by using a greedy stagewise approach, the objective function of which was the log-likelihood. We verified the validity of our model by means of the likelihood and Akaike's information criterion (AIC). The maps of the generalized extreme value parameters on the spatial plane show the existence of differentiated local trends in the extreme values of PM_{2.5} in the study area. The results indicated strong evidence of an increase in the west–east direction of the study area. A spatial map of risk with maximum concentration levels of PM_{2.5} in a period of 25 years was built.

Keywords: tree ensemble; extreme value theory; greedy stagewise; nonstationary; PM_{2.5}; CDMX

MSC: 62P12



Citation: Aguirre-Salado, A.I.; Venancio-Guzmán, S.; Aguirre-Salado, C.A.; Santiago-Santos, A. A Novel Tree Ensemble Model to Approximate the Generalized Extreme Value Distribution Parameters of the PM_{2.5} Maxima in the Mexico City Metropolitan Area. *Mathematics* **2022**, *10*, 2056. <https://doi.org/10.3390/math10122056>

Academic Editor: Natalia Markovich

Received: 13 May 2022

Accepted: 11 June 2022

Published: 14 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Particulate matter with a diameter of less than 2.5 microns is an air pollutant with potentially negative effects on humans. Called particle pollution also, particulate matter is mainly composed by sulfates, nitrates, and carbon. Sulfates constitute 25% to 55% of the total composition of PM_{2.5}, and together with nitrates, are the result of the transformation of sulfur dioxide emissions from power plants and industrial facilities and nitrogen oxide emissions from cars, trucks, and power plants. Carbon is released from emissions from cars, trucks, industrial facilities, forest fires, etc. Both ammonium sulfate and ammonium nitrate present in the atmosphere are formed from sources such as fertilizers and animal feeding operations. Particulate matter can be in solid or liquid form in particles such as dust, dirt, soot, or smoke, and even some of these can change from one form to other [1].

The negative impact of PM_{2.5} on human health has been established in a growing number of studies [2,3]. In the human respiratory system, scientists have found a significant correlation between PM_{2.5} and respiratory morbidity and mortality [4]. This negative effect on health on humans varies depending on the concentration of PM in the air and the susceptibility of the population, those who are elderly, pregnant women, adolescents, infants, and patients with cardiopulmonary problems being the most vulnerable [5–7]. Regarding the effect of PM_{2.5} concentration, the results vary slightly; in the case of lung

cancer, a study of the American Cancer Society based on a population of 500,000 adults reported an 8% increase in mortality per $10 \mu\text{g}/\text{m}^3$ increase in PM_{2.5} [8], while another study tracked 1.2 million American adults and found that the mortality of lung cancer increased by 15–27% [9]. Overall mortality and mortality of cardiopulmonary diseases are increased by 4% and 6%, respectively, for every $10 \mu\text{g}/\text{m}^3$ of PM_{2.5} increase, after ruling out occupation, smoking, diet, and other risk factors [8]. According to Zanobetti et al. [10], the increased rate in emergency hospital admissions for a $10 \mu\text{g}/\text{m}^3$ increase in 2-day averaged PM_{2.5} concentration is 1.89% for cardiac causes, 2.25% for myocardial infarction, 1.85% for congestive heart failure, 2.74% for diabetes, and 2.07% for respiratory admissions.

The exponential growth of cities has been followed by an increase in carbon emissions and, in general, an increase in air pollution by particulate matter. Such growth has occurred in a non-uniform way mainly in the large cities of the world, generating areas with different population densities within the same city. This unequal distribution of the population within the same region is reflected in a similar distribution of particulate matter (PM), derived from the positive correlation between pollution and different anthropomorphic activities. An example of this case is observed in the metropolitan area of Mexico City, which is one of the largest urban areas in the world and also a region frequently affected by air pollution. In order to monitor atmospheric concentrations of polluting gases, several monitoring stations have been established. All of these stations gather observations every hour of the day on various types of contaminants. However, despite the fact that this region is one of the most developed in the country, the number of monitoring stations is still small compared to the extensive area it covers, so extrapolating to regions without monitoring is a constant challenge. Hinojosa-Baliño [11] carried out a spatial analysis of the distribution of PM_{2.5} air pollution in Mexico city using a land-use regression model, in which two regions with high concentrations and two with low concentrations were found. Although the analysis allowed visualizing the distribution of PM_{2.5} concentrations, their results did not allow making inferences about future risks. Moreover, an analysis of extreme values to determine future risks of extreme concentrations of particulate matter was performed by Aguirre-Salado et al. [12]. This study was carried out for particles of $10 \mu\text{m}$ or less in diameter, modeling the parameters of a distribution of extreme values by smoothing with radial base functions.

A wide variety of methods have been used to analyze the spatial distribution of concentrations of PM_{2.5}. Short-term forecasting has been evaluated using regression models [11], time series models [13], random forest [14], support vector machines [15], and neural networks [16], among others. The theory of extreme values has been used to assess long-term risks. In particular, the analysis of extreme values with non-stationary trends has adjusted approximately well the nonlinear behavior observed at the extremes. Although the theory of extreme values is based on the limit distribution of the maximum of a random sample, known as the extreme value distribution or GEV distribution, several approaches have been proposed to obtain adjustments that adapt adequately to particular cases of observed phenomena. Most of these focus their efforts on approximating the trend by modeling the location parameter of the GEV distribution. Aguirre-Salado et al. [12] carried out a study on the spatial distribution of PM₁₀ in which the effect of the trend was modeled through the location parameter by means of a radial basis smoothing function. Other studies have proposed simultaneously modeling the shape parameter by a sine function [17], linear functions [18], splines [19], etc. In the case of the scale parameter, ref. [20] proposed additive models to approximate the logarithm of the scale parameter.

Regression and classification trees are popular and efficient machine learning algorithms introduced by [21]. The algorithm classifies the space of independent variables in a tree and leaf structure by optimizing some objective function, such as a likelihood function or a loss function. The strategy of building trees can be performed in many ways, some of which are based on entropy and information theory; however, in small samples, a greedy algorithm can sequentially review all possible trees to determine the optimal one. The model can be regularized by adding a penalty term in the target function [22] and

adjusted in a general gradient descent “boosting” framework [23]. This feature enables the algorithm to be highly parallelizable and suitable for big data analysis [22].

The study aims to extend the use of regression trees to the case of extreme value analysis. In extreme value regression, the use of regression trees in the optimization of the likelihood of the generalized extreme value distribution is slightly more complex, because this distribution has three parameters, instead of one, as with regression with normal and binomial distributions. This last feature creates the possibility of using different approaches for its implementation. Therefore, in order to build a parsimonious model, we used the same tree structure for the three parameters with their respective weights.

2. Materials and Methods

2.1. Study Area

The Mexico City metropolitan area (MCMA) is located in the central region of Mexico and is formed by 59 municipalities of the state of Mexico and one municipality of the state of Hidalgo. The basin extends over 9560 km² and is inhabited by approximately 25.4 million people. The average elevation is 2240 m, and it is surrounded by mountains to the east, south, and west. The study area and the primary sampling sites located in Acolman (ACO), Ajusco (AJU), Ajusco Medio (AJM), Benito Juárez (BJU), Camarones (CAM), Centro de Ciencias de la Atmósfera (CCA), Coyoacán (COY), Gustavo A. Madero (GAM), Hospital General de México (HGM), Investigaciones Nucleares (INN), Merced (MER), Miguel Hidalgo (MGH), Montecillo (MON), Milpa Alta (MPA), Nezahualcóyotl (NEZ), Pedregal (PED), La Perla (PER), San Agustín (SAG), Santa FE (SFE), San Juan Aragón (SJA), Tlalnepantla (TLA), UAM Xochimilco (UAX), UAM Iztapalapa (UIZ), Xalostoc (XAL), FES Aragón (FAR), and Santiago Acahualtepec (SAC) are shown in Figure 1.

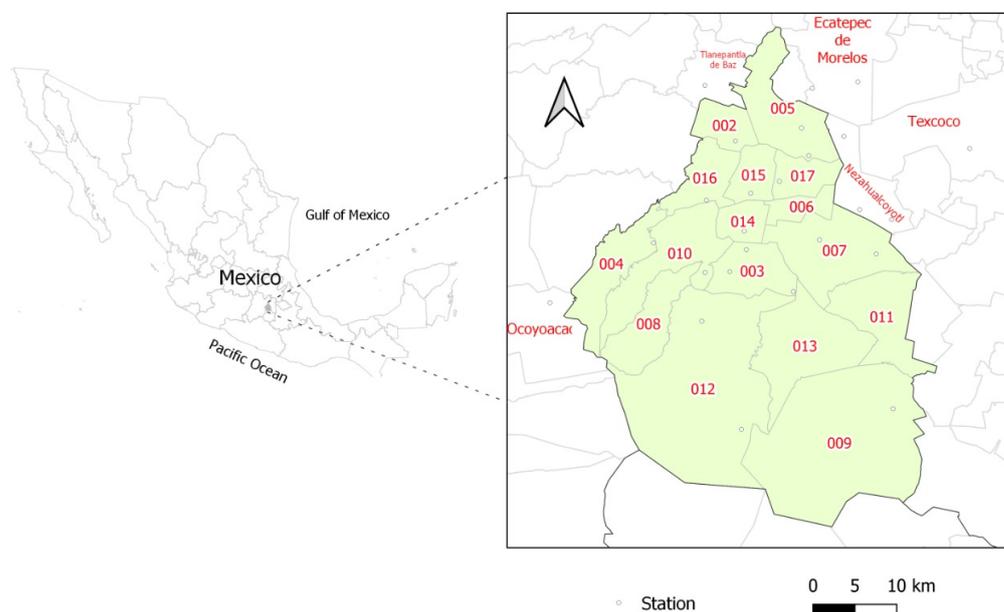


Figure 1. (Left): Mexico at the national level. (Right): Mexico city with Alcaldías. 002: Azcapotzalco, 003: Coyoacán, 004: Cuajimalpa de Morelos, 005: Gustavo A. Madero, 006: Iztacalco, 007: Iztapalapa, 008: La Magdalena Contreras, 009: Milpa Alta, 010: Álvaro Obregón, 011: Tláhuac, 012: Tlalpan, 013: Xochimilco, 014: Benito Juárez, 015: Cuauhtémoc, 016: Miguel Hidalgo, and 017: Venustiano Carranza.

2.2. Methodology

The GEV Distribution

Let Y_1, \dots, Y_n be a random sample and $M_n = \max(Y_1, \dots, Y_n)$, according to the order statistics results, the probability density function of M_n degenerates its mass of probabilities

into a point as the sample size increases. We can still use the density function of M_n through $G_n = (M_n - a_n)/b_n$, by using a sequence of constants $\{b_n > 0\}$ and $\{a_n\}$, which allow stabilizing the distribution of M_n . Important results about the asymptotic distribution of G_n have been developed and combined to achieve the limit distribution known as the generalized extreme value (GEV) distribution, shown in (1). Moreover, similar to the generalized linear models, we can associate covariates with the population parameters of the GEV distribution by means of a link function, to make predictions when having useful additional information [24].

$$G(y) = \begin{cases} \exp\left\{-\left(1 + \kappa\frac{(y-\mu)}{\sigma}\right)^{-\frac{1}{\kappa}}\right\}, & \kappa \neq 0; 1 + \kappa\frac{(y-\mu)}{\sigma} > 0 \\ \exp\left\{-\exp\left(-\frac{(y-\mu)}{\sigma}\right)\right\}, & \kappa = 0 \end{cases} \tag{1}$$

where $-\infty \leq y \leq +\infty, -\infty \leq \mu \leq +\infty, -\infty \leq \kappa \leq +\infty, \sigma > 0$; see [25,26].

2.3. Proposed Approach

In a wide variety of cases, where extreme values are analyzed, complex patterns in the behavior of the data are observed. Consequently, adjusting the parameters of the GEV distribution as if they were from a single identically distributed sample is not consistent with the information observed in the sample, because the GEV distribution is built on the assumption of independence. One possible way to solve this problem is to assume that the sample is not identically distributed in all regions of a spatial area of study and to adjust the parameters similarly to the case of generalized linear models. The novelty of our proposal is in the way of associating the covariates with the parameters of the GEV distribution. Traditionally, a linear predictor of covariates is assigned to the parameters of the distribution. However, in this case, we proposed a decision tree based on the satisfactory results obtained in several applications of machine learning. In addition, we proposed one of the first implementations where decision trees were simultaneously adjusted to more than one parameter. We assumed that observations in the same spatial locality s have an equal shape and scale parameters in the GEV distribution. However, we also assumed that the location parameter spatially varies according to a trend that is modeled by its respective decision tree. The proposed model is as follows:

$$\mu_t = \sum_{k=1}^K u_k(x_t) \tag{2}$$

$$\kappa_t = \sum_{k=1}^K v_{s,k}(x_t) \tag{3}$$

$$\log \sigma_t = \sum_{k=1}^K w_{s,k}(x_t) \tag{4}$$

where $u_{s,k}, v_{s,k}, w_{s,k} \in \mathbf{F}$ and \mathbf{F} is the class of functions of all possible regression trees.

The proposed model ensures that, locally, the sample comes from the same population, where the trend is also allowed to be estimated jointly throughout the entire region. This approach allowed obtaining a regularized model, without the need to include an additional term in the likelihood to regularize the model. An alternative was to leave σ_t and κ_t free, as well as the location parameter μ_t ; however, Yee and Stephenson [20] observed that, allowing the shape parameter to be free causes the estimate to be numerically unstable in parameterized models.

Likelihood Function

Let $\underline{y} = (y_1, \dots, y_n)$ be a sample of n extremes; the likelihood for the non-stationary GEV is defined to be the joint density as a function of the parameters as follows:

$$L(\mu_t, \sigma_t, \kappa_t | \underline{y}) = \prod_{t=1}^n \frac{1}{\sigma_t} \exp \left\{ - \left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t} \right) \right]^{-\frac{1}{\kappa_t}} \right\} \times \left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t} \right) \right]^{-\left(1 + \frac{1}{\kappa_t}\right)}$$

In practice, we estimated the parameters using the log-likelihood function, because the likelihood and log-likelihood have the same critical points; however, the log-likelihood is a numerically simpler function to optimize. The log-likelihood function for the non-stationary GEV is as follows:

$$\ell(\mu_t, \sigma_t, \kappa_t | \underline{y}) = -n \log \sigma_t - \sum_{t=1}^n \left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t} \right) \right]^{-\frac{1}{\kappa_t}} - \sum_{t=1}^n \left(1 + \frac{1}{\kappa_t} \right) \log \left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t} \right) \right]. \tag{5}$$

To simplify the notation we define $\ell_t(\mu_t, \sigma_t, \kappa_t | \underline{y}) = -\log \sigma_t - \left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t} \right) \right]^{-\frac{1}{\kappa_t}} - \left(1 + \frac{1}{\kappa_t} \right) \log \left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t} \right) \right]$ and $\ell_n(\mu_t, \sigma_t, \kappa_t | \underline{y}) = \sum_{t=1}^n \ell_t(\mu_t, \sigma_t, \kappa_t | \underline{y})$, and consequently we can rewrite Equation (5) as:

$$\ell_n(\mu_t, \sigma_t, \kappa_t | \underline{y}) = \sum_{t=1}^n \ell_t(\mu_t, \sigma_t, \kappa_t | \underline{y})$$

Therefore, the gradient of the likelihood is given by:

$$\begin{aligned} \frac{\partial \ell_n(\mu_t, \sigma_t, \kappa_t | \underline{y})}{\partial \mu_t} &= \sum_{t=1}^n \frac{\ell_t(\mu_t, \sigma_t, \kappa_t | \underline{y})}{\partial \mu_t} \\ \frac{\partial \ell_n(\mu_t, \sigma_t, \kappa_t | \underline{y})}{\partial \log \sigma_t} &= \sum_{t=1}^n \frac{\ell_t(\mu_t, \sigma_t, \kappa_t | \underline{y})}{\partial \log \sigma_t} \\ \frac{\partial \ell_n(\mu_t, \sigma_t, \kappa_t | \underline{y})}{\partial \kappa_t} &= \sum_{t=1}^n \frac{\ell_t(\mu_t, \sigma_t, \kappa_t | \underline{y})}{\partial \kappa_t} \end{aligned}$$

where

$$\begin{aligned} \frac{\ell_t(\mu_t, \sigma_t, \kappa_t | \underline{y})}{\partial \mu_t} &= -\frac{1}{\sigma_t \left[1 + \kappa_t \left(\frac{y - \mu_t}{\sigma_t} \right) \right]^{(1/\kappa_t)+1}} + \frac{\kappa_t (1 + (1/\kappa_t))}{\sigma_t \left(1 + \kappa_t \left(\frac{y - \mu_t}{\sigma_t} \right) \right)} \\ \frac{\ell_t(\mu_t, \sigma_t, \kappa_t | \underline{y})}{\partial \log \sigma_t} &= -1 - \frac{\left(\frac{y - \mu_t}{\sigma_t} \right)}{\left(1 + \kappa_t \left(\frac{y - \mu_t}{\sigma_t} \right) \right)^{(1/\kappa_t)+1}} + \frac{(1 + (1/\kappa_t)) \left(\kappa_t \left(\frac{y - \mu_t}{\sigma_t} \right) \right)}{1 + \kappa_t \left(\frac{y - \mu_t}{\sigma_t} \right)} \\ \frac{\ell_t(\mu_t, \sigma_t, \kappa_t | \underline{y})}{\partial \kappa_t} &= \frac{-\left((1/\kappa_t) + 1 \right) \left(\frac{y - \mu_t}{\sigma_t} \right)}{1 + \kappa_t \left(\frac{y - \mu_t}{\sigma_t} \right)} + \frac{\log \left(1 + \kappa_t \left(\frac{y - \mu_t}{\sigma_t} \right) \right)}{\kappa_t^2} \\ &\quad + \frac{\left(\frac{y - \mu_t}{\sigma_t} \right)}{\kappa_t \left(1 + \left(\kappa_t \left(\frac{y - \mu_t}{\sigma_t} \right) \right) / \left(\sigma_t \right) \right)^{(1/\kappa_t)+1}} + \frac{\log \left(1 + \left(\kappa_t \left(\frac{y - \mu_t}{\sigma_t} \right) \right) / \left(\sigma_t \right) \right)}{\kappa_t^2 \left(1 + \left(\kappa_t \left(\frac{y - \mu_t}{\sigma_t} \right) \right) / \left(\sigma_t \right) \right)^{(1/\kappa_t)}} \end{aligned}$$

The optimization strategy within each tree is by gradient descent. According to Equations (3) and (4), both the parameters of the shape and scale in all observations are equal at the same monitoring station. However, Equation (2) shows that we have an equal value of the location parameter for all observations that are on the same leaf of a tree. We used these assumptions based on the fact that the observations obtained at the same monitoring stations come from a GEV distribution with the same shape and scale parameter.

Without loss of generality, we wrote $\ell_n(\mu_t, \sigma_t, \kappa_t | \underline{y}) = \phi(F(\mathbf{x})) = \sum_{t=1}^n \ell(y_t, F(\underline{x}_t))$, and following the numerical optimization procedure, we took the solution to be

$$F(\underline{x}_t) = \sum_{k=1}^K f_k(\underline{x}_t)$$

where

$$f(\underline{x}_t) = \begin{pmatrix} u_k(\underline{x}_t) \\ v_{s,k}(\underline{x}_t) \\ w_{s,k}(\underline{x}_t) \end{pmatrix}$$

$u_0(\underline{x}_t), v_{s,0}(\underline{x}_t); w_{s,0}(\underline{x}_t)$ is an initial guess, and $\{u_k(\underline{x}_t)\}_1^K, \{v_{s,k}(\underline{x}_t)\}_1^K$ and $\{w_{s,k}(\underline{x}_t)\}_1^K$ are incremental functions (“steps” or “boosts”) defined by the optimization method.

For the steepest descent,

$$f_k(\underline{x}_t) = -\rho_k g_k(\underline{x}_t)$$

with

$$g_k(\underline{x}_t) = \left[\frac{\partial \ell(y_t, F(\underline{x}_t))}{\partial F(\underline{x}_t)} \right]_{F(\mathbf{x})=F_{k-1}(\mathbf{x})}$$

Each of the K-simpler functions $f_k(\underline{x}_t)$ can be alternatively obtained by using a “greedy stagewise” approach, for $m = 1, 2, \dots, M$:

$$f_k(\underline{x}_t) = \arg \max_{f_k} \sum_{t=1}^n \ell(y_t, F_{k-1}(\underline{x}_t) + f_k(\underline{x}_t))$$

The above algorithm may be numerically efficient; this may not be optimal, because the addition of functions is sequential. To understand this, consider the multiple linear regression model. In these models, we cannot obtain an optimal global solution by adding the variables one by one, but by optimizing the entire set of variables simultaneously, the optimal can be obtained. Similar to this situation, the optimal solution, although in some situations infeasible, is one that satisfies the following condition:

$$f_k(\underline{x}_t) = \arg \max_{f_k} \sum_{t=1}^n \ell \left(y_t, \sum_{k=1}^K f_k(\underline{x}_t) \right)$$

The structure proposed by the model allowed improving the adjustment as the number of leaves increases without incurring over-fitting. A model with few leaves has the advantage of including the spatial relationships of nearby monitoring stations in a parsimonious model; however, it may not be optimal when using a “greedy stagewise” approach. Increasing the number of “boost” steps, improves the model without incurring an over-fit; therefore, a simulation study was not required for the proposed model.

2.4. Data Analysis

The extreme values were obtained using the block maxima approach [27]. The extreme values were obtained from a large block with 720 observations, with hourly measured data. Therefore, we have a random sample of monthly block maxima that is roughly independent. We assumed that the value of the location parameter was spatially homogeneous between nearby stations, while the scale and shape parameters are specific to each monitoring station. Weight estimation for leaves of the estimated decision trees was carried out sequentially, taking the coordinates of each station and calculating the resulting log-likelihood using a greedy algorithm that reviewed all monitoring stations. The graphical overview of the global procedure is shown in Figure 2. Although this approach was considered numerically efficient, an optimal search scheme would consist of forming all possible combinations of trees. Therefore, we performed the optimization by using the gradient descent method, with step sizes of 0.05, 0.001, and 0.0001 for the parameters of location, scale, and shape,

respectively. We verified that, when using these configurations, the algorithm always converged to a critical point. The homogeneity of maxima in each monitoring station justified that each of these had its own parameter of scale and shape, leading to a simple and parsimonious model. This feature allowed the model to remain regularized without the need to add extra terms in the likelihood function. For simplicity, we adjusted the parameters on each leaf of a tree, instead of using additive functions on each leaf. We observed that, although this approach was slower, it also was equivalent to the other.

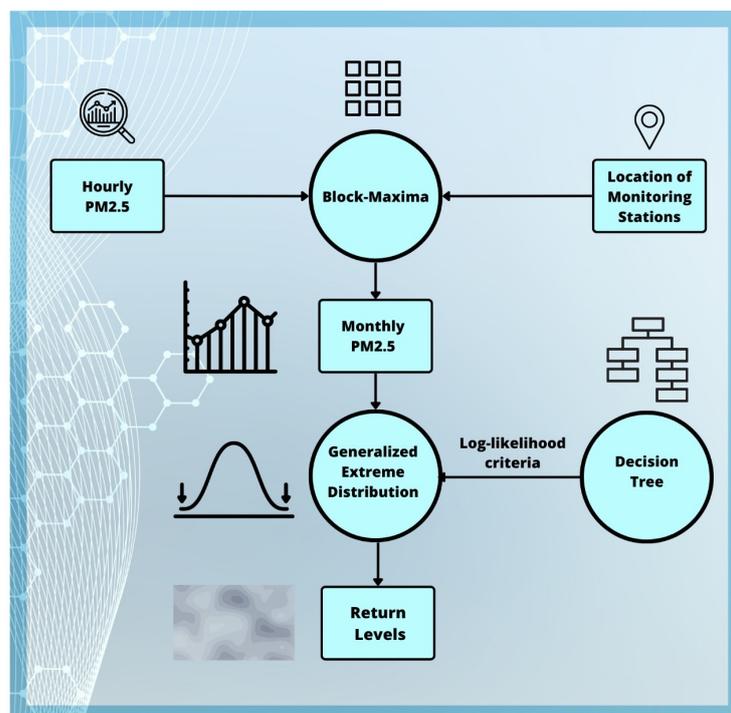


Figure 2. Graphical overview of the study structure.

2.5. Data Collection

The data used corresponded to 2683 observations of monthly block maxima of $PM_{2.5}$, between 2 August 2003 and 11 September 2021, obtained at 26 fixed monitoring stations of the Sistema de Monitoreo Atmosférico (SIMAT) through the Red Automática de Monitoreo Atmosférico (RAMA), a network established by the Mexico City Ministry of Environment (Secretaría del Medio Ambiente (SEDEMA)), responsible for data gathering and reporting air quality levels. SIMAT has a total of 69 stations, of which only 26 stations measure $PM_{2.5}$.

3. Results and Discussion

We present a descriptive summary of the data in Table 1 in which we observe that each monitoring station has different parametric characteristics. We highlight the “PER” monitoring station with high concentrations of $PM_{2.5}$ measured, in contrast with the “AJM” station, which has a lower average concentration, as well as the “SAC”, “SFE”, and “PED” stations. We can also observe the existence of highly atypical values in most monitoring stations. The “SAC” station has an atypical value greater than $600 \mu\text{g}/\text{m}^3$, which is more than double the magnitude of the next highest value in the same monitoring station. However, the largest atypical value is at the “XAL” station, with a concentration of $988 \mu\text{g}/\text{m}^3$. We ordered the stations in the box-plot according to their geographical proximity and observed that the next stations share similar distributional characteristics. This justifies the choice of the proposed model, in which the location parameter is shared between nearby stations.

Table 1. Descriptive summary information on the PM2.5 maxima in the Mexico City metropolitan area.

Block ID	Key	Long(W)	Lat(N)	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	ACO	−98.9°	19.6°	35	62.8	87	100.5	117.8	281
14	AJU	−99.2°	19.2°	34	47	62	80.9	100	302
19	CCA	−99.2°	19.3°	30	48	56	65.3	72	302
16	INN	−99.4°	19.3°	20	42.2	51	58.8	59.5	246
18	AJM	−99.2°	19.3°	44	57	66	69.9	77.8	127
24	MGH	−99.2°	19.4°	51	64.5	71	86.1	90	267
25	BJU	−99.2°	19.4°	41	59	67	83.4	83	690
23	MER	−99.1°	19.4°	31	72.8	84.5	94.5	101	428
8	TLA	−99.2°	19.5°	33	73.5	86	93	104	294
9	FAR	−99°	19.5°	34	47.5	59	68.1	72.5	236
22	HGM	−99.2°	19.4°	36	74	90	97.2	107	346
20	PED	−99.2°	19.3°	41	59	69	73.8	78.5	179
26	COY	−99.2°	19.4°	21	72	85	95.6	105	544
3	SAC	−99°	19.3°	50	67	77	84.4	98	211
15	MPA	−99°	19.2°	46	57	65.5	84.8	103.5	211
11	SJA	−99.1°	19.5°	34	65.2	81.5	95.7	110	333
21	CAM	−99.2°	19.5°	43	71	83	95.7	99.2	777
2	MON	−98.9°	19.5°	29	50	65	73.3	83	227
6	UAX	−99.1°	19.3°	40	55	66.5	78.5	86.8	209
17	SFE	−99.3°	19.4°	28	56.5	70	72.9	81	179
5	PER	−99°	19.4°	53	87	125	167.2	200	681
10	GAM	−99.1°	19.5°	48	66	75	86.4	89	359
12	SAG	−99°	19.5°	36	65	77	98.4	101.8	698
13	XAL	−99.1°	19.5°	58	84	101	125.3	129	988
4	NEZ	−99°	19.4°	39	63.2	81	100.9	114	393
7	UIZ	−99.1°	19.4°	44	73	88	101.4	110.2	429

The box-plot diagram in Figure 3 shows that nearby monitoring stations share similar characteristics, especially those related to the upper quantiles of the distribution. Therefore, these properties in the observations mean that the estimation of the parameters by means of decision trees can be considered an appropriate choice for the estimation of the parameters. Moreover, these features, mainly those based on the magnitude of the quantiles, could be used to propose heuristics in the process of building the decision trees, in order to reach global solutions.

The greedy stagewise Algorithm 1 was implemented using the statistical software R 4.1.2. This algorithm was based on the log-likelihood and was used to build both the tree, as well as all the other algorithms used in this research. We denote by symbol I the initial set of instances, i.e., the set of all geographical localities where the monitoring stations were located. Therefore, an element in I is a bivariate vector containing the latitude and longitude of some monitoring station. We also denote by J the set of unit vectors in R^2 and operation \cdot as the usual inner product in R^2 . We propose to split the tree by taking an element in I and an element in J and evaluated the log-Likelihood of the resulting tree after splitting, then we retained the candidate with the highest log-Likelihood. This stage was the most computationally intense, mainly in the first divisions of the tree where each branch has relatively many observations. The algorithm was repeated the number of times defined in the *depth* variable.

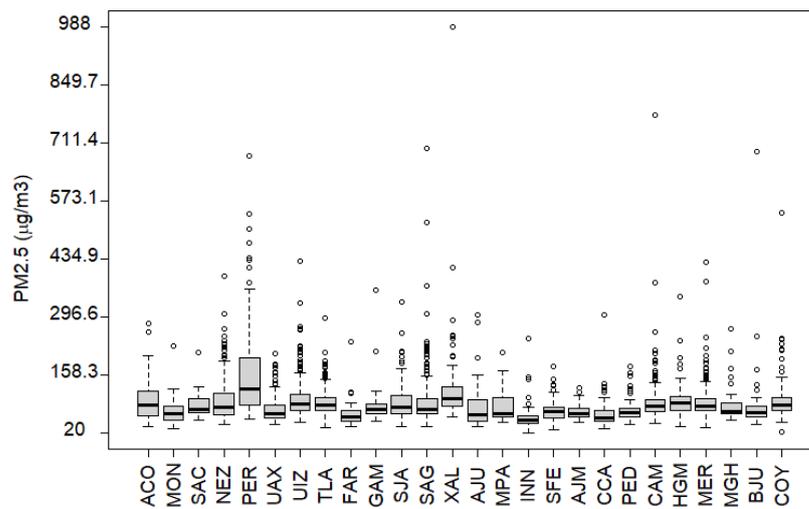


Figure 3. Box-plots of the PM_{2.5} maxima at 26 monitoring stations in the Mexico City metropolitan area.

Algorithm 1 Greedy stagewise algorithm for split finding

Input: I , initial instance set

- 1: **for** i in I
- 2: **for** j in J
- 3: $L \leftarrow \max_F \sum_{k \in \{k | k \cdot j < i \cdot j\}} \ell(y_k, F(x_k))$
- 4: $R \leftarrow \max_F \sum_{k \in \{k | k \cdot j > i \cdot j\}} \ell(y_k, F(x_k))$
- 5: score $\leftarrow L + R$
- 6: **end for**
- 7: **end for**

Output: Split with max score

We verified the validity of our model by means of the log-Likelihood and the Akaike’s information criterion (AIC). Because the representation of the model in (4) induces a parametric model on each leaf, we calculated the AIC of the global tree by adding the contributions of the AIC on each leaf. The log-likelihood of the stationary extreme value model was $-13,061.27$, whereas the log-likelihood of the final model obtained by using the decision tree was $-12,844.21$, which means that the final model has properly learned the distribution of the data. Moreover, the AIC for the stationary model is $26,128.54$, while the AIC of the proposed model is $25,844.42$. In addition, in Figure 4, we present the quantile–quantile plot, which demonstrates the goodness of fit of our model. Therefore, the validity of our model was verified.

Several algorithms can be proposed in order to obtain the rules that split the tree in an optimal way. Chen and Guestrin [22], in single-parameter models, proposed an exact greedy algorithm for split finding and an approximate algorithm for split finding based on percentiles. We used a similar approach in a multi-parametric model, using each component of the parameter vector simultaneously as homogeneous candidates for split finding. This setting decreased the processing time; however, it also decreased the solution space. Therefore, with this configuration, we obtained a balance between the processing time and the number of trees visited.

The adjusted decision tree is shown in Figure 5. We observed that the likelihood in the root node of the tree was equal to the likelihood of the stationary model. At the end of the 10th depth, branching the tree improved the log-likelihood to $-12,844.21$, which is enough to ensure that the model has improved statistically. In order to divide the tree and add more branches, we chose the node that maximized the sum of the log-likelihood of the resulting

highest values for the location parameter are located near the meridian -99.2 W and parallel 19.5 N. In Figure 6b, we show a map with the estimates for the scale parameter. We can observe that the characteristics of the map are similar to the location parameter. Although, these have notable differences in the northeast region, where the scale parameter tends to increase rather than decrease, in contrast to the location parameter. A similar situation occurred between the parallels 19.2 N and 19.4 N and east of the meridian -99.3 W. A totally different behavior is observed with the shape parameter. The decision tree adjusted for the shape parameter is shown in Figure 6c. This figure shows that, in general, the parameter is positive, with an increasing trend in the southwest to northeast direction, concentrating the highest values in the central and northeastern region of the study area.

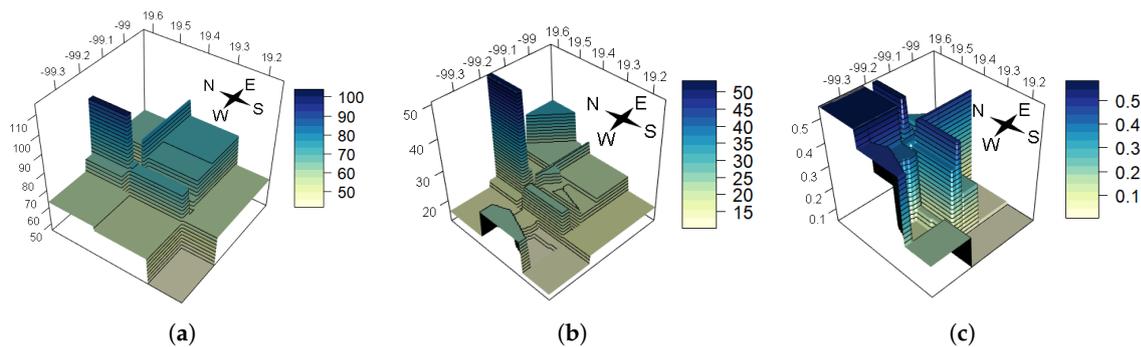


Figure 6. (a) Three-dimensional representation of the adjusted decision tree of the location parameter, (b) scale parameter, and (c) shape parameter. The X and Y axes are in geographical coordinates (decimal degrees). Z is the calculated value of the corresponding parameter (location, scale, and shape), for each geographical position.

The results in Figure 6 also show the existence of geographical zones with distributions of heavy tails in areas surrounding the coordinates -99 W and 19.4 N. These findings are similar to those found by Hinojosa-Baliño et al. [11] on daily $PM_{2.5}$ concentrations in the same regions of Mexico City. These results show that, in general, in the eastern region of the study area, the highest values of $PM_{2.5}$ concentrations were observed. Our results also coincided with the results obtained by their research in regions of the southern and southwest part of the study area. The results are similar in both investigations, showing a positive correlation between the location parameter of the GEV distribution and the daily $PM_{2.5}$ observations analyzed by them.

The map of 25-year return levels is shown in Figure 7. Return levels Z_p are concentration levels whose values are expected to be exceeded once every $1/p$ years. Return level estimates at monitoring stations were obtained using the model (2), and we extrapolated these values on the map using the inverse distance weighting algorithm, with the aim of smoothing the map of return levels. The highest return levels were expected in areas near the PER station. The lowest return levels were expected in areas near the AJM station. We observed that return levels tend to increase in the east–west direction and decrease again after the MON station. The return map also showed the characteristic of the model of grouping nearby stations according to their location parameter, which leads us to a more homogeneous map and a model with fewer parameters. An important feature of the map was the smoothness of the estimates from one monitoring station to another, in addition to the stability of the estimates. Indeed, the models that associate each observation with a single distribution tend to over-adjust the data, causing unrealistic estimates.

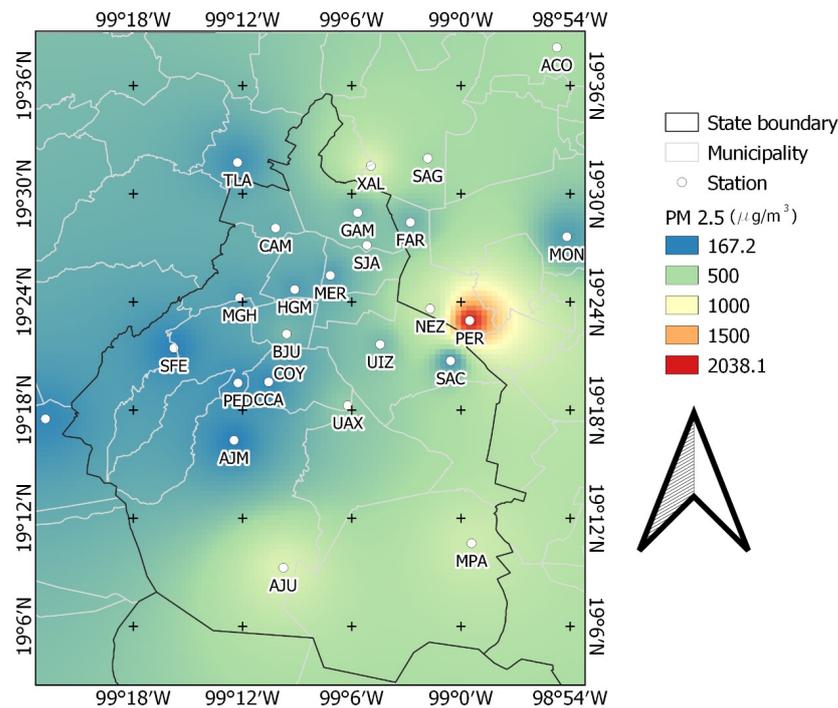


Figure 7. Spatial distribution of PM2.5 for a return period of 25 years for the study region.

We compared our results with findings from similar research on *PM10* carried out for the same area of study as the present research. Aguirre-Salado et al. [12] developed a hierarchical model for the spatial analysis of *PM10* pollution extremes in the Mexico City Metropolitan Area. They based their estimates on radial-based smoothing functions and spatially modeled only the location parameter, allowing the scale and shape parameters to be constant. Therefore, the non-stationary extreme values were modeled, obtaining a linear increment pattern in the southeast–northwest direction, as shown schematically in Figure 8. We found a similar pattern with respect to the extreme values of *PM2.5*. However, the trend was slightly modified, resulting in a direction of increase in the west–east. In addition, the proposed decision tree is parametrically simpler and more stable, which was reflected in the adequate convergence obtained in each of the adjustment steps.

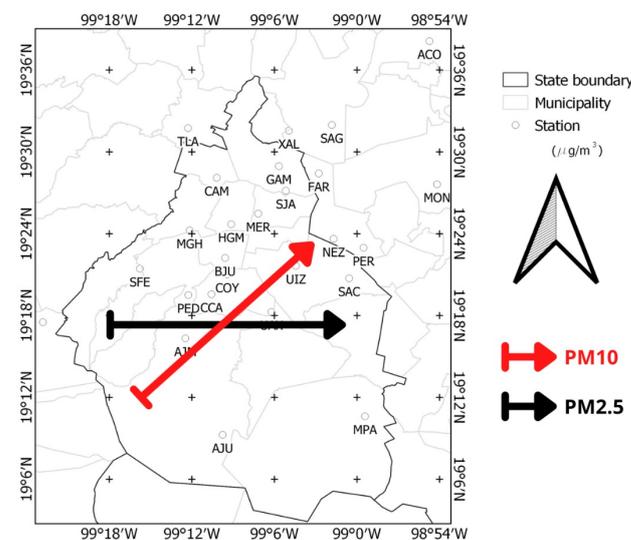


Figure 8. Spatial comparison of increase trends in the study region. *PM10* (red line, Aguirre-Salado et al. [12]) and *PM2.5* (black line).

The advantage of considering observations from the same monitoring station as elements of the same distribution is observed in Figure 7. Previous studies on the distribution of non-stationary extreme values on particulate matter in the metropolitan area of Mexico City did not restrict the model to prevent observations obtained in the same geographic location [12]. Although they assumed a model too flexible to represent the conditions observed in each measurement, this still caused bias when the elements of the distribution tails was considered and resulted in unreal and unstable estimates. Therefore, such a situation generated unstable and non-robust models, and when added or removed, observations produced drastically different estimators.

In contrast, the model proposed in this research allowed considering the observations of the same monitoring station with an identical distribution, leading to robust estimators, which in addition did not present the common problems of the non-convergence of the estimation algorithm.

We observed that the greedy stagewise algorithm and any other greedy algorithm that add incremental functions, such as “steps” or “boosts”, had a high risk of generating decision trees that satisfy conditions of local optima. Further work may involve proposing algorithms, in the context of extreme value theory, using the upper quantiles of the maxima as distances for the clustering of monitoring stations and increasing the number of nodes in groups, not sequentially one by one. Additionally, it should be investigated if proposing the candidate nodes in the leaves of the each candidate decision tree, the gain in the log-likelihood improves. We believe that in this situation, the respective loss of algorithm performance will be followed by a gain in the log-likelihood.

4. Conclusions

In this research work, we proposed a study on the extreme non-stationary values of PM_{2.5} maxima using a tree ensemble model. The model had the advantage of approximating complex nonlinear spatial trends of extreme values, using a decision-tree-based assembly model for the parameters of the GEV distribution that makes use of a simpler K model. The parameters in each leaf were estimated via gradient descent, which has the advantage of being easily implemented, which in each leaf almost surely converges to the optimal solution. Additionally, the estimates were obtained by adjusting to the non-stationary GEV model using a decision tree approach simultaneously for the three parameters of the extreme value distribution, which is a novel way to perform estimation in multivariate parameter models by using decision trees. Our model was validated by comparing the log-likelihood and the AIC of the stationary model with those obtained for the fitted model, resulting in the best values being obtained by the proposed model, showing us the support and validity of our results. We also concluded that an important change to extend the model should consider the construction of the decision tree using conglomerates based on the proximity of the monitoring stations and their upper quantiles; this should help to solve the problem of local minimum solutions that can be obtained by the greedy stagewise approach. Our findings indicated the existence of areas with increased risks of high PM_{2.5} concentrations in the west–east direction of the study area. The results of our work should help administrative authorities improve the policies of prevention and contention of extreme events of PM_{2.5} concentrations.

Author Contributions: Conceptualization, A.I.A.-S. and S.V.-G.; methodology, C.A.A.-S., A.S.-S. and S.V.-G.; software, A.I.A.-S. and C.A.A.-S.; validation, C.A.A.-S., S.V.-G. and A.S.-S.; formal analysis, A.I.A.-S., S.V.-G. and C.A.A.-S.; investigation, A.I.A.-S. and S.V.-G.; resources, S.V.-G.; data curation, A.I.A.-S. and C.A.A.-S.; writing—original draft preparation, A.I.A.-S., A.S.-S., C.A.A.-S. and S.V.-G.; writing—review and editing, A.I.A.-S.; visualization, A.I.A.-S. and C.A.A.-S.; supervision, S.V.-G. and A.S.-S.; project administration, A.I.A.-S., S.V.-G. and A.S.-S. All authors have read and agreed to the published version of the manuscript.

Funding: The authors want to thank CONACYT for the financial support through the scholarship granted to Sonia Venancio-Guzmán for pursuing her doctoral degree.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank the Ministry of the Environment of Mexico City (<http://www.aire.cdmx.gob.mx/default.php?opc='aKBh'>, accessed on 10 June 2022) for providing the data used in this research. Special thanks are also given to two anonymous reviewers who shared insightful observations that deeply improved our work.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike's information criterion
CDMX	Ciudad de México
EVT	Extreme value theory
GEV	Generalized extreme value
RAMA	Red Automática de Monitoreo Atmosférico
SIMAT	Sistema de Monitoreo Atmosférico
SEDEMA	Secretaría del Medio Ambiente

References

1. United States Environmental Protection Agency. The Particle Pollution Report: Current Understanding of Air Quality and Emissions through 2003. Report No. EPA 454-R-04-002. Office of Air Quality Planning and Standards Emissions, Monitoring, and Analysis Division Research Triangle Park, North Carolina. 2004. Available online: https://www.epa.gov/sites/default/files/2017-11/documents/pp_report_2003.pdf (accessed on 10 June 2022).
2. Nemery, B.; Hoet, P.H.; Nemmar, A. The Meuse Valley fog of 1930: An air pollution disaster. *Lancet* **2001**, *357*, 704–708. [[CrossRef](#)]
3. Orru, H.; Maasikmets, M.; Lai, T.; Tamm, T.; Kaasik, M.; Kimmel, V.; Orru, K.; Merisalu, E.; Forsberg, B. Health impacts of particulate matter in five major Estonian towns: Main sources of exposure and local differences. *Air Qual. Atmos. Health* **2010**, *4*, 247–258. [[CrossRef](#)]
4. Xing, Y.F.; Xu, Y.H.; Shi, M.H.; Lian, Y.X. The impact of PM_{2.5} on the human respiratory system. *J. Thorac. Dis.* **2016**, *8*, E69–E74. [[CrossRef](#)] [[PubMed](#)]
5. Huynh, M.; Woodruff, T.J.; Parker, J.D.; Schoendorf, K.C. Relationships between air pollution and preterm birth in California. *Paediatr. Perinat. Epidemiol.* **2006**, *20*, 454–461. [[CrossRef](#)] [[PubMed](#)]
6. de Oliveira, B.F.A.; Ignotti, E.; Artaxo, P.; do Nascimento Saldiva, P.H.; Junger, W.L.; Hacon, S. Risk assessment of PM_{2.5} to child residents in Brazilian Amazon region with biofuel production. *Environ. Health* **2012**, *11*, 64. [[CrossRef](#)] [[PubMed](#)]
7. Martinelli, N.; Girelli, D.; Cigolini, D.; Sandri, M.; Ricci, G.; Rocca, G.; Olivieri, O. Access Rate to the Emergency Department for Venous Thromboembolism in Relationship with Coarse and Fine Particulate Matter Air Pollution. *PLoS ONE* **2012**, *7*, e34831. [[CrossRef](#)]
8. Pope, C.A., III; Burnett, R.T.; Thun, M.J.; Calle, E.E.; Krewski, D.; Ito, K.; Thurston, G.D. Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *JAMA* **2002**, *287*, 1132. [[CrossRef](#)] [[PubMed](#)]
9. Turner, M.C.; Krewski, D.; Pope, C.A.; Chen, Y.; Gapstur, S.M.; Thun, M.J. Long-term Ambient Fine Particulate Matter Air Pollution and Lung Cancer in a Large Cohort of Never-Smokers. *Am. J. Respir. Crit. Care Med.* **2011**, *184*, 1374–1381. [[CrossRef](#)] [[PubMed](#)]
10. Zanobetti, A.; Franklin, M.; Koutrakis, P.; Schwartz, J. Fine particulate air pollution and its components in association with cause-specific emergency admissions. *Environ. Health* **2009**, *8*, 58. [[CrossRef](#)] [[PubMed](#)]
11. Hinojosa-Baliño, I.; Infante-Vázquez, O.; Vallejo, M. Distribution of PM_{2.5} Air Pollution in Mexico City: Spatial Analysis with Land-Use Regression Model. *Appl. Sci.* **2019**, *9*, 2936. [[CrossRef](#)]
12. Aguirre-Salado, A.I.; Vaquera-Huerta, H.; Aguirre-Salado, C.A.; Reyes-Mora, S.; Olvera-Cervantes, A.D.; Lancho-Romero, G.A.; Soubervielle-Montalvo, C. Developing a Hierarchical Model for the Spatial Analysis of PM₁₀ Pollution Extremes in the Mexico City Metropolitan Area. *Int. J. Environ. Res. Public Health* **2017**, *14*, 734. [[CrossRef](#)] [[PubMed](#)]
13. Chiang, P.W.; Horng, S.J. Hybrid Time-Series Framework for Daily-Based PM_{2.5} Forecasting. *IEEE Access* **2021**, *9*, 104162–104176. [[CrossRef](#)]
14. Geng, G.; Meng, X.; He, K.; Liu, Y. Random forest models for PM_{2.5} speciation concentrations using MISR fractional AODs. *Environ. Res. Lett.* **2020**, *15*, 034056. [[CrossRef](#)]

15. Zhang, C.J.; Dai, L.J.; Ma, L.M. Rolling forecasting model of PM2.5 concentration based on support vector machine and particle swarm optimization. In Proceedings of the International Symposium on Optoelectronic Technology and Application, Beijing, China, 9–11 May 2016; Volume 10156, pp. 387–394. [\[CrossRef\]](#)
16. Masinde, C.J.; Gitahi, J.; Hahn, M. Training recurrent neural networks for particulate matter concentration prediction. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *43*, 1575–1582. [\[CrossRef\]](#)
17. Weissman, I. Estimation of parameters and large quantiles based on the k largest observations. *J. Am. Stat. Assoc.* **1978**, *73*, 812–815.
18. Tawn, J. Bivariate extreme value theory: Models and estimation. *Biometrika* **1988**, *75*, 397–415. [\[CrossRef\]](#)
19. Rosen, O.; Cohen, A. Extreme percentile regression. In *Statistical Theory and Computational Aspects of Smoothing: Proceedings of the COMPSTAT '94 Satellite Meeting, Semmering, Austria, 27–28 August 1994*; Härdle, W., Schimek, M.G., Eds.; Physica: Heidelberg, Germany, 1996; pp. 27–28.
20. Yee, T.W.; Stephenson, A.G. Vector generalized linear and additive extreme value models. *Extremes* **2007**, *10*, 1–19. [\[CrossRef\]](#)
21. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: London, UK, 2017. [\[CrossRef\]](#)
22. Chen, T.; Guestrin, C. XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016. [\[CrossRef\]](#)
23. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **1999**, *29*, 1189–1232. [\[CrossRef\]](#)
24. Coles, S. *An Introduction to Statistical Modeling of Extreme Values*; Springer: Berlin/Heidelberg, Germany, 2001; Volume 208.
25. Jenkinson, A.F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Q. J. R. Meteorol. Soc.* **1955**, *81*, 158–171. [\[CrossRef\]](#)
26. Fisher, R.A.; Tippett, L.H.C. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Camb. Philos. Soc.* **1928**, *24*, 180–190. [\[CrossRef\]](#)
27. Gumbel, E. *Statistics of Extremes*; Columbia University Press: New York, NY, USA, 1958.