Hindawi
Journal of Applied Mathematics
Volume 2021, Article ID 3572555, 13 pages
https://doi.org/10.1155/2021/3572555

Research Article

An Enhanced Method for Tail Index Estimation under Missingness

F. Ayiah-Mensah,[1,2] R. Minkah,[1] L. Asiedu,[1] and F. O. Mettle[1]

[1]*Department of Statistics and Actuarial Science, School of Physical and Mathematical Sciences, College of Basic and Applied Sciences, University of Ghana, Ghana*
[2]*Department of Mathematics, Statistics and Actuarial Science, Takoradi Technical University, Takoradi, Ghana*

Correspondence should be addressed to R. Minkah; rminkah@ug.edu.gh

Received 13 May 2021; Revised 17 June 2021; Accepted 27 June 2021; Published 29 July 2021

Academic Editor: A. Bassam

Extreme events in earthquakes, wind speed, among others are rare but may lead to catastrophic effects on humans and the environment. The primary parameter in the estimation of such rare events is the tail index which measures the tail heaviness of an underlying distribution. Since extreme events are rare, the presence of missing observations may further lead to flawed. In view of this, there is a growing effort by researchers to address this problem. However, the existing methods of estimating the tail index use only the available nonmissing data. Thus, if the missing observations are influential values, ignoring them could introduce more bias and higher mean square error (MSE) in the tail index estimation and subsequently other extreme event–estimators such as high quantiles and small exceedance probabilities. In this study, we propose imputation of the missing observations before applying some standard estimators (Hill and geometric-type) to estimate the tail index. Through a simulation study, we assess the performance of the standard estimators under the proposed data enhancement method and the existing modified estimators of the tail index. The results show that the enhanced estimators have relatively lower bias and MSE. The estimation method was illustrated with a practical dataset on wind speed with missing values. Therefore, we recommend imputation mechanism as viable for enhancing the performance of tail index estimators in the case where there is missingness.

1. Introduction

Statistics of extremes is a branch of Statistics that deals with the estimation of parameters of rare events. It enables the assessment of the probability of events that are more extreme than any previous observation from a sample of random variables Coles [1]. According to Gomes and Guillou [2, 3], the occurrence of rare events in phenomena such as earthquakes, hurricanes, wind speed, sea waves, and floods can have a catastrophic impact on human beings and the environment. Modelling the occurrence of such events aids in planning to reduce or prevent the impact of such events.

Recent developments in this area focus on modelling and predicting rare events to mitigate their negative impact on humans and properties. The primary parameter of interest is the tail index or extreme value index (EVI), which measures the tail heaviness of an underlying distribution. One key challenge researchers encounter in their quest to model rare events or estimate the tail index is that the number of observations available is usually very small to none due to their unusual occurrence. Therefore, having missing observations can affect the tail index estimates computed from the sample, thereby leading to unreliable estimates for exceedance probabilities, high quantile, and return period which are the goals of extreme value analysis. In this study, we propose an enhanced method of estimating the tail index of underlying distributions, where the missing values are estimated and replaced in the data via an imputation method.

Extreme Value Theorem (EVT) was proposed to model the tails of probability distribution. According to Fisher and Tippett [4], there are three broad families of the limiting distributions in EVT, namely, Gumbel, Frechet (Pareto or Fre´chet-Pareto), and Weibull. These are also referred to as the extreme value distributions (EVD). These three families of

extreme value distributions were simplified by Jenkinson [5] as the generalised extreme value (GEV) distribution. The GEV distribution function for a random variable $X \in R$ is given by

$$\Theta_{\xi}(x) = \begin{cases} \exp\left\{-\left(1+\xi\dfrac{x-\mu}{\sigma}\right)^{-1/\xi}\right\}, & 1+\xi\dfrac{x-\mu}{\sigma} > 0, \xi \neq 0, \\ \exp\left\{-\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right\}, & -\infty < x < \infty, \xi = 0, \end{cases} \tag{1}$$

where $-\infty < \mu < \infty$, $\sigma > 0$, and ξ are the location, scale, and the tail index (or extreme value index (EVI)), respectively. According to Coles [1], ξ in (1) determines the most suitable type of tail behaviour for a dataset. The tail distribution is classified as Gumbel when $\xi = 0$, Frechet-Pareto when $\xi > 0$, and Weibull when $\xi < 0$.

Many researches conducted in statistics of extremes have been limited to estimating the tail index and other parameters of extremes with complete dataset (Beirlant et al., Ameraoui et al., Minkah et al. [6–9]). Gomes and Pestana [10] and Beirlant et al. [11] proposed reduced-bias estimators for the tail index parameter on complete data. In the case of presence of censoring, Gomes and Neves [12], Ameraoui et al. [6], and Minkah et al. [13] provide techniques for incorporating censoring in the estimation of the tail index. Li and Qi [14] employed an existing adjusted empirical likelihood method, to construct confidence intervals of the tail index so as to achieve a better accuracy. Through a simulation study, they found their method to be superior in terms of the coverage probability and length of confidence intervals. In the case of the presence of covariate information, Ma et al. [15] propose empirical likelihood-based statistics to construct confidence regions for the regression coefficient of the parametric tail index regression model. Also, Minkah et al. [16] studies several estimators of the tail index in the presence of both censoring and covariate information. However, all these estimators do not take into account missing data and hence are somewhat challenged in the presence of missing observations.

Since missing data is a common problem in statistics, some authors such Mladenovic and Piterbarg [17] worked on the estimation of the exponent of the regular variation with the use of incomplete data samples. They proved the asymptotic consistency of the Hill estimator. Ilić and Mladenovic [18] extended the works of Hsing [19] and Mladenovic and Piterbarg [17] where the authors relied on available observations (incomplete samples). Under the assumption of weak dependency, they proved the consistency of their proposed Hill-type estimator of the tail index based on an incomplete sample.

In addition, Zou et al. [20] focused on extreme value analysis without the largest values. The study revealed that the presence of missing extremes makes the choice of threshold for the top order statistics problematic. They simultaneously considered the number of missing extremes with the tail index and other parameters and proposed a functional version of the Hill estimator and named it Hill Estimator Without Extremes (HEWE). The estimator was found to be robust to missing extremes on light-tailed dataset.

Furthermore, Ilić and Veličković [21] considered the simple tail index estimation in the case of heterogeneous and dependent data samples with missing values. Their study was on the asymptotic behaviour of the median estimator and its robustness against deviations of the slowly varying function. Although under small deviations from the assumed parametric model, the proposed method provided a reliable tail index, the top values of the sample were not considered.

However, the existing methods for estimating the tail index use reduced sample size since portions of the dataset within the order statistics that are missing are ignored. Using only portions of the data may result in estimators with large bias and/or variance especially if the missing observations are influential in the top order statistics which are of interest in statistics of extremes.

Therefore, in the quest to reduce bias and variance in tail index estimation in the presence of missing observations; we propose imputation of the missing observations before applying standard tail index estimators (such as Hill and geometric-type), instead of using the modified estimators in the literature.

The rest of the paper is organised as follows. In Section 2, we present the materials and methods including the existing and the proposed method for estimating tail index. Section 3 presents the results of the simulation study and a discussion of the results. Lastly, in Section 4, we provide concluding remarks, areas for future research, and recommendations.

2. Material and Methods

Let $X_1, X_2, \cdots, X_n$ be a sample of independent and identically distributed observations from some process with underlying distribution F. Assume $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ to be the sample order statistics associated with the sample. The so-called semiparametric estimators of the tail index (see, e.g., Beirlant et al. [7, 22], de Haan and Resnick [23]) in the literature rely on the exceedances over a particular threshold, $X_{n-k,n}$. Thus, the dependence of tail index on the $k+1$ largest order statistics makes the selection of k critical. A careful choice of k is needed as a small value leads to large $X_{n-k,n}$ and hence few observations for estimation. This may lead to a tail index estimator with smaller bias but larger variance. On the contrary, a large k leads to small $X_{n-k,n}$ which may result in the inclusion of observations with smaller magnitude leading to a larger sample. Although, having more data will reduce the variance, it is at the expense of bias. The problem is addressed by choosing k such that as n increases, k increases but at a slower rate. Formally,

$$k = k_n \longrightarrow \infty \text{ as } n \longrightarrow \infty, \tag{2}$$

such that

$$k_n = o(n). \tag{3}$$

Equations (2) and (3) are used to obtain a number of nonzero sequence of integers k_n which are referred to as

intermediate. Next, we present some standard estimators for tail index estimation under complete dataset.

2.1. Hill Estimator. The Hill estimator (Hill [24]) is the most popular estimator in the Frechet-Pareto family under semiparametric method (Gomes and Guillou [2, 3]). The Hill estimator is valid for $\xi > 0$ and is given by

$$\xi_{k,n}^{(H)} = \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n} \tag{4}$$

where k is an intermediate sequence of integers defined in (2) and (3). Some desirable properties of the Hill estimator are its consistency and asymptotic normality (de Haan and Resnick [23]). However, it is known for its dependence on k and exhibits large bias for large k values.

2.2. Geometric-Type Estimator. The Geometric-type estimator, proposed by Brito and Freitas [25], is an adopted geometrical estimator motivated by the fact that, for large random variable X, $-\log(1 - F(X))$ is approximately linear with slope R, where R is a positive constant. The estimator is given by

$$\widehat{R}(k_n) = \sqrt{\frac{\sum_{t=1}^{k_n} \log^2(n/t) - (1/k_n)\left(\sum_{t=1}^{k_n} \log(n/t)\right)^2}{\sum_{t=1}^{k_n} X_{n-t+1,n}^2 - (1/k_n)\left(\sum_{t=1}^{k_n} X_{n-t+1,n}\right)^2}}, \tag{5}$$

where n is the sample size (number of random variables), k_n is a sequence of positive integers satisfying $1 \le k_n < n$, as well as Equations (2) and (3). Brito and Freitas [25] investigated the weak asymptotic properties of the geometric-type estimator and showed that its distribution is asymptotically normal under general conditions.

2.3. Estimators of Tail Index under Missing Observations. We now discuss existing modifications of the Hill and geometric-type estimators of the tail index when there are missing observations. For a given sample $X_1, X_2, \cdots, X_n$ with some missing observations, we consider an observed portion in the sample to be $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_{S_n}$, where S_n is the number of observed random variables among the first n terms of the sequence S_n and $S_n \le n$. The order statistics of the observed sample is $\tilde{X}_{1,S_n} \le \tilde{X}_{2,S_n} \le, \cdots, \le \tilde{X}_{S_n,S_n}$, such that the associated maximum is $\tilde{X}_{S_n,S_n}$.

According to Ilić [26], a sample with missing observations may be obtained on condition that every observed term has probability $p > 0$ which is independent of the other terms. Hence, S_n is a binomial random variable with parameters n and p. To obtain the tail index estimator, the following assumption must be satisfied:

The random variable S_n is independent of $X_1, X_2, \cdots$, and there exists a sequence of real numbers $\{\gamma_n\}$ such that

$$\lim_{n\longrightarrow\infty} \gamma_n = +\infty, \tag{6}$$

Table 1: Frechet-Pareto type distributions.

Distribution	Notation	$1 - F(x)$	Conditions	ξ
Burr	Burr (β, τ, λ)	$\left(\frac{\beta}{\beta + x^{\tau}}\right)^{\lambda}$	$x > 0\,; \beta, \tau, \lambda > 0$	$\frac{1}{\lambda\tau}$
Fréchet	Fréchet (α)	$1 - \exp(-x^{-\alpha})$	$x > 0, \alpha > 0$	$\frac{1}{\alpha}$
Pareto	Pa (α)	$x^{-\alpha}$	$x \ge 1, \alpha > 0$	$\frac{1}{\alpha}$

and

$$\frac{S_n}{\gamma_n} \xrightarrow{p} c_0 > 0 \text{ as } n \longrightarrow \infty. \tag{7}$$

Let β_n be a sequence of real numbers such that $\lim_{n\longrightarrow\infty}\beta_n = \infty$ and $\lim_{n\longrightarrow\infty}\beta_n/n = 0$. Also, let $M_n = S_n/\beta_n$ and

$$\beta_n = \begin{cases} 0, & \text{if } S_n = 0, \\ \dfrac{M_n}{S_n}, & \text{if } S_n \ge 1. \end{cases} \tag{8}$$

Then, the inequalities $\lfloor S_n/\beta_n \rfloor \le S_n/\beta_n < \lfloor S_n/\beta_n \rfloor + 1$ and $S_n/\beta_n - 1 < \lfloor S_n/\beta_n \rfloor \le S_n/\beta_n$ will hold for *bxc*, the largest integer not greater than x.

The modified Hill and geometric-type estimators are respectively given as

$$\xi_{k,n}^{(MHill)} = \mathbb{I}(S_n \ge \beta_n)\frac{1}{M_n}\sum_{t-1}^{M_n} \log \tilde{X}_{S_n-t+1,S_n} - \log \tilde{X}_{S_n-M_n,S_n}, \tag{9}$$

and

$$\widehat{R}(S_n) = \mathbb{I}(S_n \ge \beta_n)\sqrt{\frac{\sum_{t=1}^{M_n} \log^2(S_n/t) - (1/M_n)\left(\sum_{t=1}^{M_n} \log(S_n/t)\right)^2}{\sum_{t=1}^{M_n} \log^2 \tilde{X}_{S_n-t+1,S_n} - (1/M_n)\left(\sum_{t=1}^{M_n} \log \tilde{X}_{S_n-t,S_n}\right)^2}}, \tag{10}$$

where $\mathbb{I}$ is the indicator function such that for $x \in A$, $\mathbb{I}(x) = 1$, and $\mathbb{I}(x) = 0$ if $x \notin A$.

2.4. Multivariate Imputation with Chain Equations (MICE). MICE is one of the widely used imputation methods for filling missing observations in data. MICE, also known as the sequential regression or fully conditional specification multiple imputation, is a very flexible method because it can handle different variable types such as discrete and continuous. It uses fully conditional specification to preserve unique features such as bounds, skip patterns, interactions, and bracketed responses in the data (Van Buuren [27]).

The MICE operation is based on the assumption of Missing at Random (MAR) with the implication that missing value probability is independent of unobserved values but is dependent of the observed values Schafer and Graham [28].

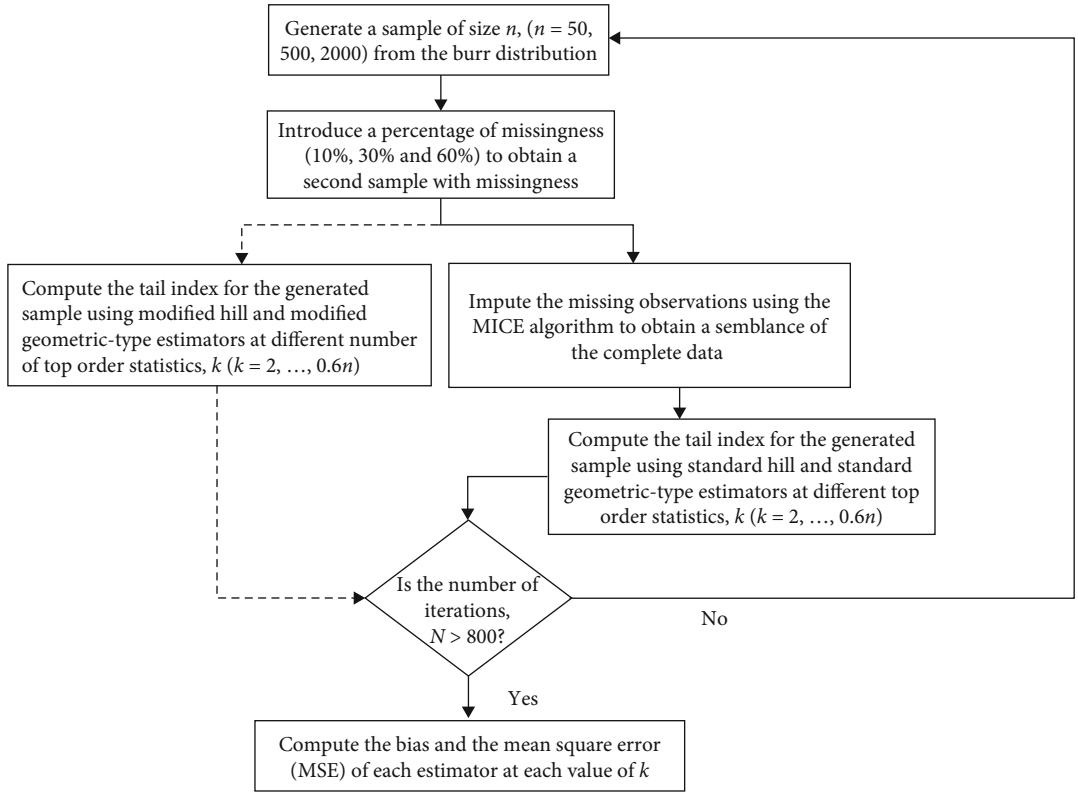


Figure 1: Simulation algorithm.

Table 2: Notation of estimators.

Name of estimator	Notation
Modified Hill	M_Hill
Modified geometric-type	Geom
Hill on MICE imputed dataset	Im.Hill
Geometric-type on MICE imputed dataset	Im.Geom

MICE has three different phases which are similar to any other multiple imputation method: imputation, analysis, and pooling. It creates multiple imputations to overcome the limitation of single imputation. MICE can handle large data sets through the use of chain equations as compared to other imputation methods that uses joint models (He et al. [29]). This makes it a flexible multiple imputation method that uses a number of regression algorithms. In this study, we use the MICE algorithm to impute missing observations.

2.5. Proposed Data Enhancement Method for Tail Index When Observations Have Missingness. This method uses MICE algorithm to impute the missing observations before applying the standard estimators of the tail index such as the Hill and geometric-type. Again, consider the sample $X_1, X_2, \cdots, X_n$, where n is the sample size of a dataset which is not fully realized due to missing observation(s). That is, the dataset available is $X_1, X_2, \cdots, X_{S_n}$, and $S_n < n$. We propose the following data enhancement and estimation method of the tail index parameter:

(1) Apply MICE on the incomplete data, $X_1, X_2, \cdots, X_{S_n}$, to generate the missing observations $X_1^*, X_2^*, \cdots, X_{n-S_n}^*$

(2) Combine the imputed observations $X_1^*, X_2^*, \cdots, X_{n-S_n}^*$ and the available observations $(X_1, X_2, \cdots, X_{S_n})$ to obtain a sample of size n with observations $X_1, X_2, \cdots, X_{S_n}, X_1^*, X_2^*, \cdots, X_{n-S_n}^*$, hereinafter referred to as an imputed dataset

(3) Obtain the order statistics $\tilde{X}_{1,n}^* \le \tilde{X}_{2,n}^* \le \cdots \le \tilde{X}_{n,n}^*$ associated with $X_1, X_2 \cdots X_{S_n}, X_1^*, X_2^* \cdots X_{n-S_n}^*$ such that the associated maximum is $\tilde{X}_{n,n}^* = \max(X_1, X_2, \cdots, X_{S_n}, X_1^*, X_2^*, \cdots, X_{n-S_n}^*)$

(4) Assume F is in the Maximum Domain of Attraction (MDA) for a suitable tail index ξ as is the case in a semiparametric framework and select the k upper order statistics in $\tilde{X}_{1,n}^* \le \tilde{X}_{2,n}^* \le \cdots \le \tilde{X}_{n,n}^*$

(5) Estimate ξ using the standard estimators (Hill and the geometric-type without any modification) based on k upper order statistics, $\tilde{X}_{n-k+1,n}^* \le \tilde{X}_{n-k+2,n}^* \le \cdots \le \tilde{X}_{n,n}^*$ where k is an intermediate sequence

2.6. Simulation Design. In this section, we present a simulation study to compare the results of the data enhancement method to some existing methods, such as Mladenovic and Piterbarg [17], for estimating the tail index as discussed in

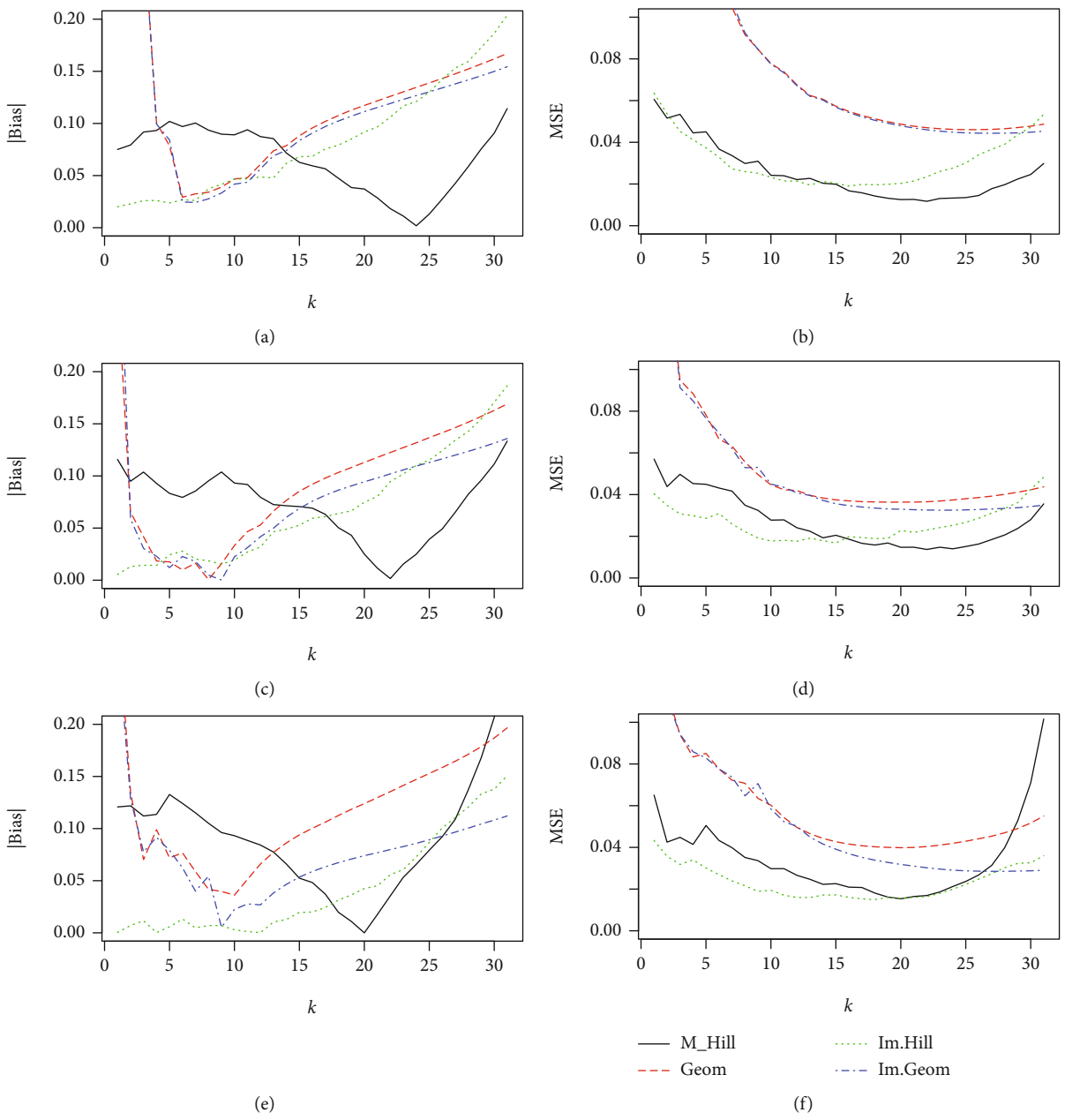


FIGURE 2: Results of $\widehat{\xi}(k)$ from Burr distribution with $n = 50$. (a, c, and e) Absolute bias. (b, d, and f) MSE. (a, b), (c, d), and (e, f) are 10%, 30%, and 60% missingness, respectively.

Section 2.3. We generate samples from distributions from the Pareto domain of attraction for the simulation study. Table 1 contains the distribution functions used for the simulation study and their characteristics.

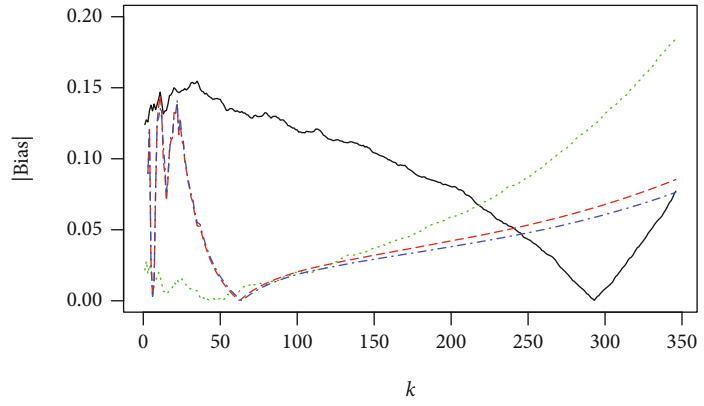
A step-by-step procedure for the simulation study is outlined in the flowchart in Figure 1.

In the next section, we present and discuss the simulation results. However, for brevity and ease of presentation, we provide the simulation results and discussion of the performance of the estimators for samples generated from the Burr distribution only. The results from the other distributions did not differ significantly from the Burr distribution and are available upon request from the authors.

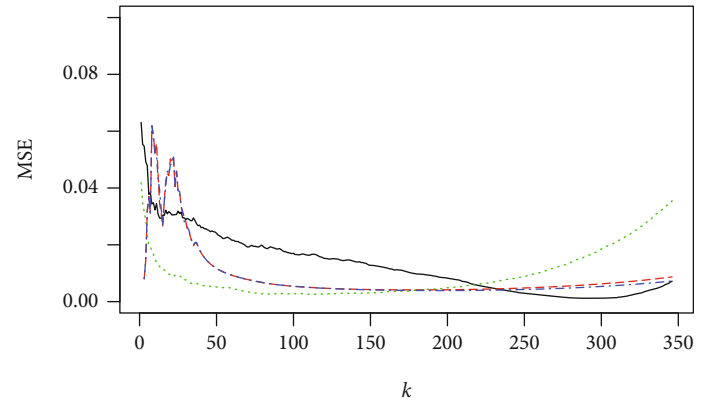
3. Results and Discussion

Generally, an estimator with relatively smaller bias and MSE is preferred. In addition, we require such an estimator to be stable as k increases. An estimator that is less sensitive to the changing values of k maintains a stable outlook throughout the evolution of k. Such an estimator is deemed as the most appropriate for the estimation of the tail index as it maintains a better balance between bias and variance.

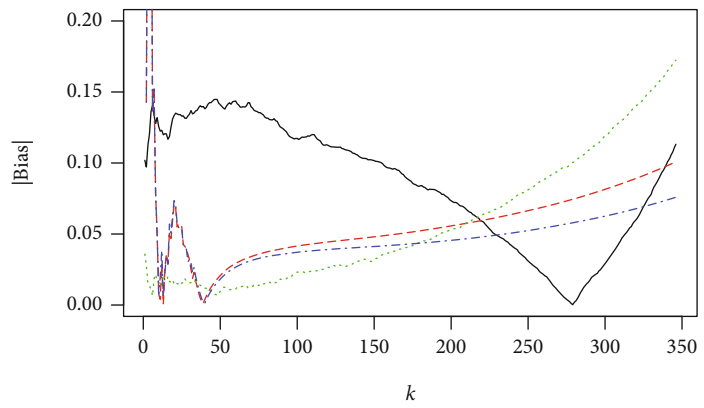
Since extreme value analysis for the right tail concerns larger observations, we assess the estimators' performances for k up to 60% of the sample size. Thus, this enables the inclusion of smaller order statistics and the assessment of



(a)



(b)



(c)

Figure 3: Continued.

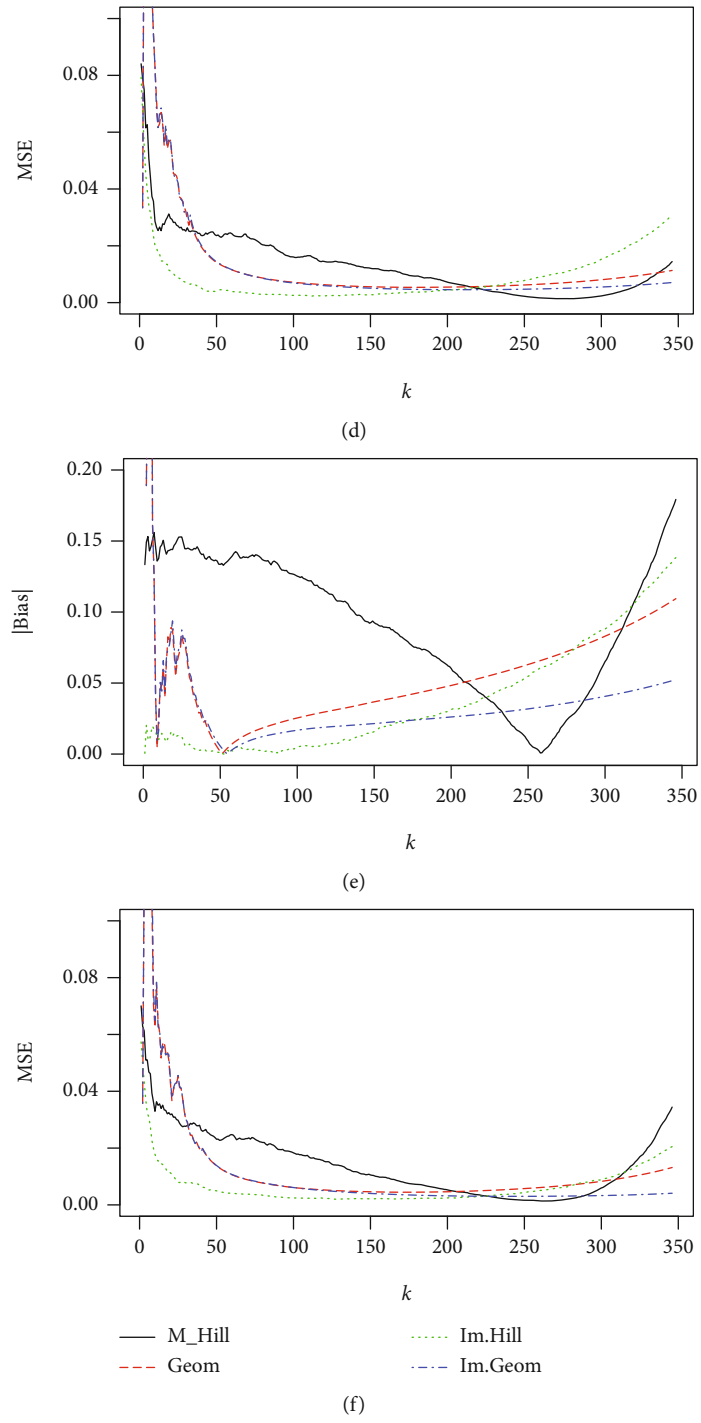


FIGURE 3: Results of $\widehat{\xi}(k)$ from Burr distribution with $n = 500$. (a, c, and e) Absolute bias. (b, d, and f) MSE. (a, b), (c, d), and (e, f) are 10%, 30%, and 60% missingness, respectively.

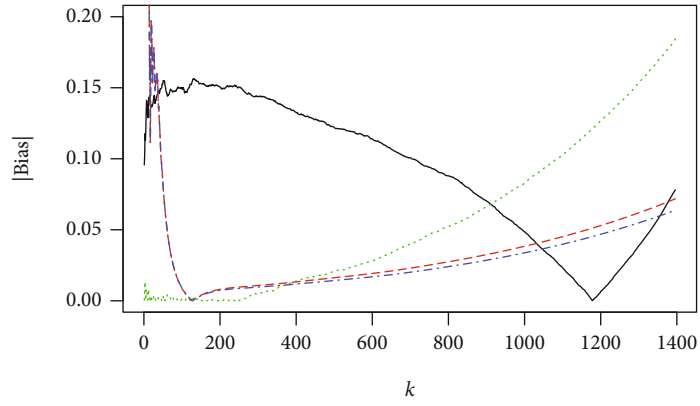
estimators' performance across a broad spectrum of k where bias is expected to be prevalent.

Table 2 contains the notations of estimators used in the study.

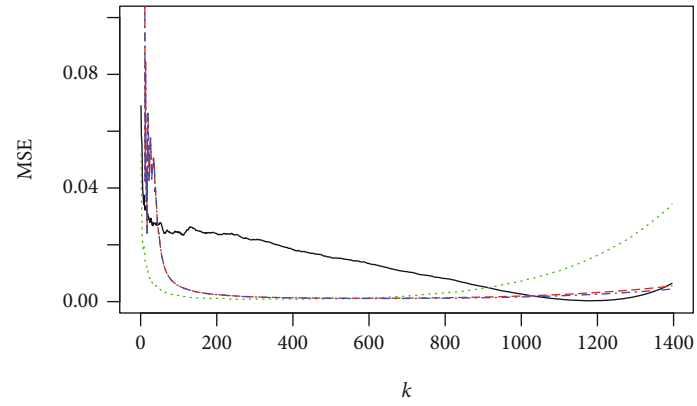
Also, all the simulation and practical application (i.e., Section 3.2) results were obtained using the R statistical package, and the codes are available upon request from the authors.

3.1. Results from the Burr Distribution. For each figure discussed in this subsection, the left and the right panels show the absolute bias and the MSE of an estimator. Also, the top, middle, and bottom panels depict 10%, 30%, and 60% missing observations, respectively relative to the sample size.

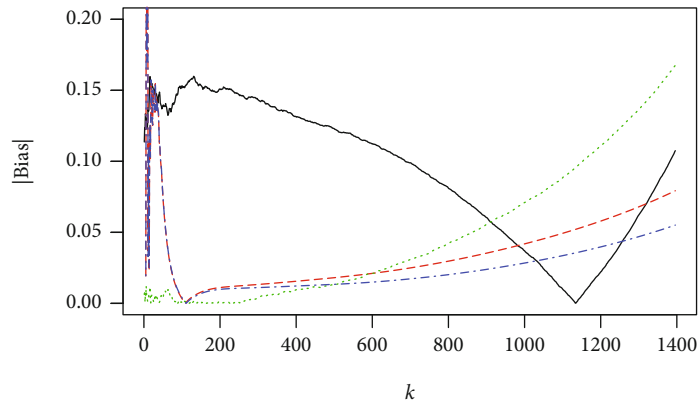
Figure 2 shows the absolute bias and MSE of the estimators of the tail index for samples of size, $n = 50$, generated from the Burr distribution. It is evident from the left



(a)



(b)



(c)

Figure 4: Continued.

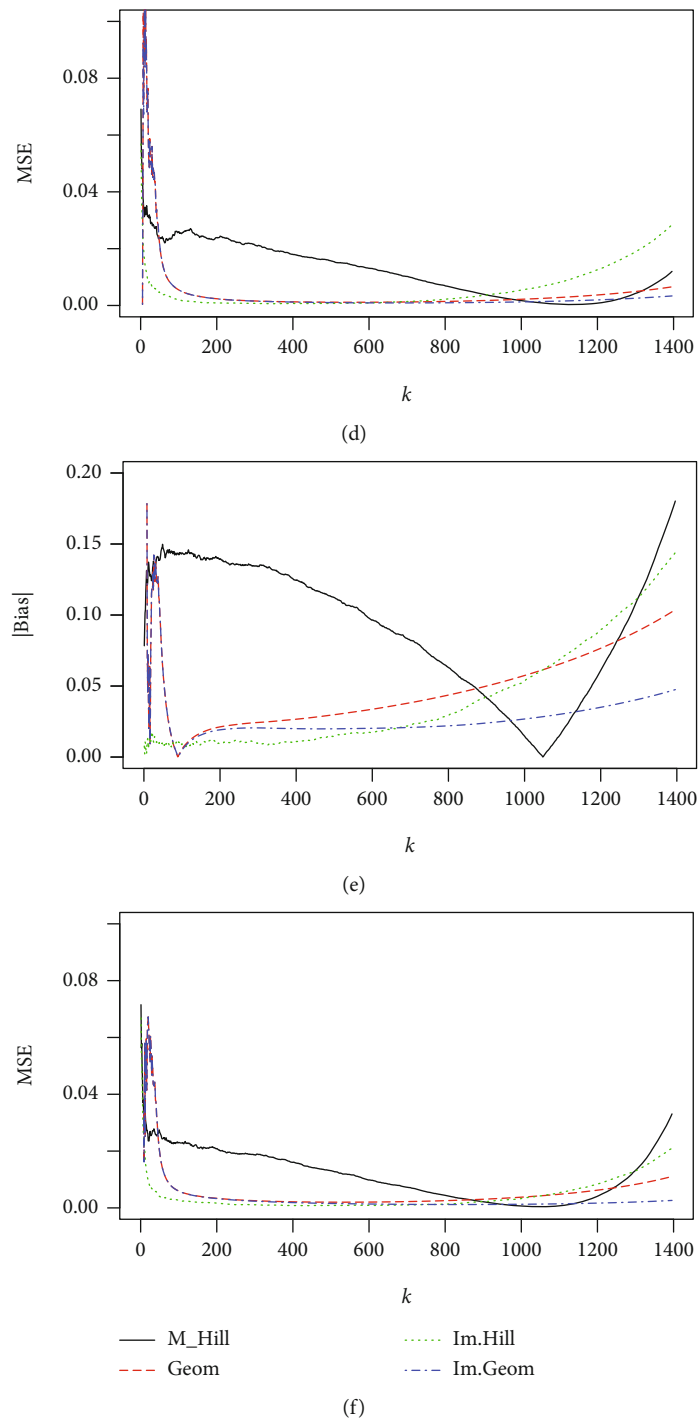


FIGURE 4: Results of $\widehat{\xi}(k)$ from Burr distribution with $n = 2000$. (a, c, and e) Absolute bias. (b, d, and f) MSE. (a, b), (c, d), and (e, f) are 10%, 30%, and 60% missingness, respectively.

panel (i.e., consisting of 10%, 30%, and 60% missingness) of Figure 2 that the Hill on MICE imputed dataset (Im.Hill) has the smallest absolute bias for small values of k. Generally, the bias of Im.Hill increases as k increases regardless of the percentage of missing observations. Also, Im.Hill is not stable as k increases above the first five top order statistics.

Although the M_Hill estimator is the most biased estimator for small values of k, it has the least bias for larger values of k. The Im.Hill estimator is competitive to the M_Hill for larger values of k.

In terms of MSE, the right panel of Figure 2 shows that all the estimators diverge as k increases. However, the Im.Hill is closer to zero for smaller values of k in all the percentages of

missingness, i.e., 10%, 30%, and 60%. Also, for large k, generally, the Im.Hill estimator has smaller MSE comparable to that of the M_Hill in the case of samples with high percentages of missingness.

Therefore, in terms of bias and MSE, the proposed Im.Hill can be considered as the most appropriate across a large spectrum of k.

In addition, the simulation results for samples of size $n = 500$ are shown in Figure 3. In the case of bias, the Im.Hill estimator exhibits the least bias and relative stable for k less than 40% of the sample size. Beyond this value of k, M_Hill relatively has the smallest bias but unstable as it diverges as k increases. For MSE, Im.Hill has the least MSE for k approximately less than 50% of the sample size. Again, M_Hill has better MSE values than all the other estimators for k over 50% of the sample size. However, for high percentage of missingness, the proposed Im.Hill outperforms the M_Hill estimator. Thus, overall, the Im.Hill provides a better estimator of the tail index in terms of bias and MSE across the values of k.

Furthermore, in the case of samples of size $n = 2000$, the estimators exhibit similar performance characteristics to the two preceding cases discussed. More importantly, a closer look at the corresponding graphs for all the sample sizes indicates that the MSE generally decreases as n increases. This is empirically consistent with the consistency property of estimators of the tail index.

Moreover, the right panel which shows the estimators' performance in terms of MSE indicates that the Im.Hill, Geom, and Im.Geom estimators are more stable within the 50 and 200 upper order statistics, whereas M_Hill is not stable within the 50 and 200 upper order statistics. Here again, the Hill on MICE imputed dataset (Im.Hill) has the smallest MSE within the 200 upper order statistics. Hence, Im.Hill is the preferred estimator for estimating the tail index under missing observations when the sample of size $n = 500$ is drawn from the Burr distribution.

The results of the tail index estimators on a sample of size $n = 2000$ drawn from the Burr distribution are presented in Figure 4. The results in the left panel (subgraphs (a), (c) and (e) representing absolute bias for the 10%, 30%, and 60% missingness) indicate that Geom, Im.Geom, and M_Hill are not stable within the first 200 upper order statistics.

Specifically, the absolute bias of M_Hill decreases as k increases within 200 and 1000 upper order statistics whereas the absolute bias of Im.Hill, Geom, and Im.Geom increases as k increases. Comparatively, Im.Hill is stable within the first 200 upper order statistics with smallest absolute bias. Therefore, Im.Hill can be said to be the best/preferred estimator in terms of low bias. Figures 4(b), 4(d), and 4(f) present the MSE of the tail index estimators on a sample of size $n = 2000$ drawn from the Burr distribution. Within 200 and 800 upper order statistics, the MSE of M_Hill decreases as k increases. The MSE of Im.Hill, Geom, and Im.Geom are more stable as k increases. Im.Hill has the smallest MSE within the first 400 upper order statistics and hence is selected as the most appropriate estimator for estimating the tail index under missing observations using a sample drawn from the Burr distribution with size $n = 2000$.

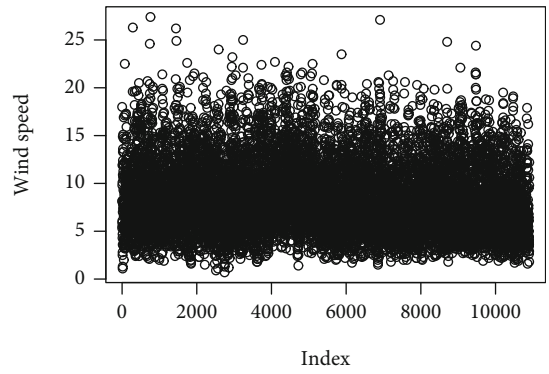


Figure 5: Scatter plot of the wind speed data.

3.2. Application to Real-Life Data. In this section, we illustrate the enhanced and existing methods of estimating tail index under missing observations, on a real-life wind speed data obtained from the extremefit package in R (Durrieu et al. [30]). The wind speed data contains the average wind speed in meters per second (m/s) per day in Brest (France) from 1976 to 2005. The data contains 10903 observations with minimum wind speed of 0.700, maximum of 27.400, and an average daily mean wind speed of 8.553.

High wind speeds are known to cause collapse of buildings, ships, difficulties in aircraft takeoff and landing, among others (see, e.g., Marchigiani et al. [31]). Therefore, modelling the tail behaviour of an underlying distribution of wind speed will help in planning and mitigating the effects of extreme wind speeds. The presence of missing values in historic wind data speed needs to be taken care of in the modelling process. In the case of the present wind speed data, there are 6 missing values. In addition, in order to further assess the suitability of the proposed data enhancement method for tail index estimation, we introduced missingness up to 45% (i.e., 10%, 25%, and 45% to represent small, medium, and large percentage of missingness, respectively) of the sample size.

The application of the proposed data enhancement method of estimating tail index begins with a search for the domain of attraction of the underlying distribution of the wind speed dataset. Figure 5 (the scatter plot of the wind speed data) shows some few observations are detached from the majority of the data values. Thus, the detached values (large values) may be outliers or extreme observations.

It is evident from the histogram that the wind speed data is positively skewed which suggests that the data has a heavy tail. Also, the general increasing trend of the mean excesses as the threshold decreases indicates that the wind speed data has an underlying distribution which is heavy-tailed than the exponential. Again, the QQ-plots at the bottom of Figures 6(c) and 6(d) compare the sample quantiles of the wind speed data to the theoretical quantiles of the exponential and Pareto distributions. Both plots support the assertion from the previous graphs that the underlying distribution has a Pareto-type tail.

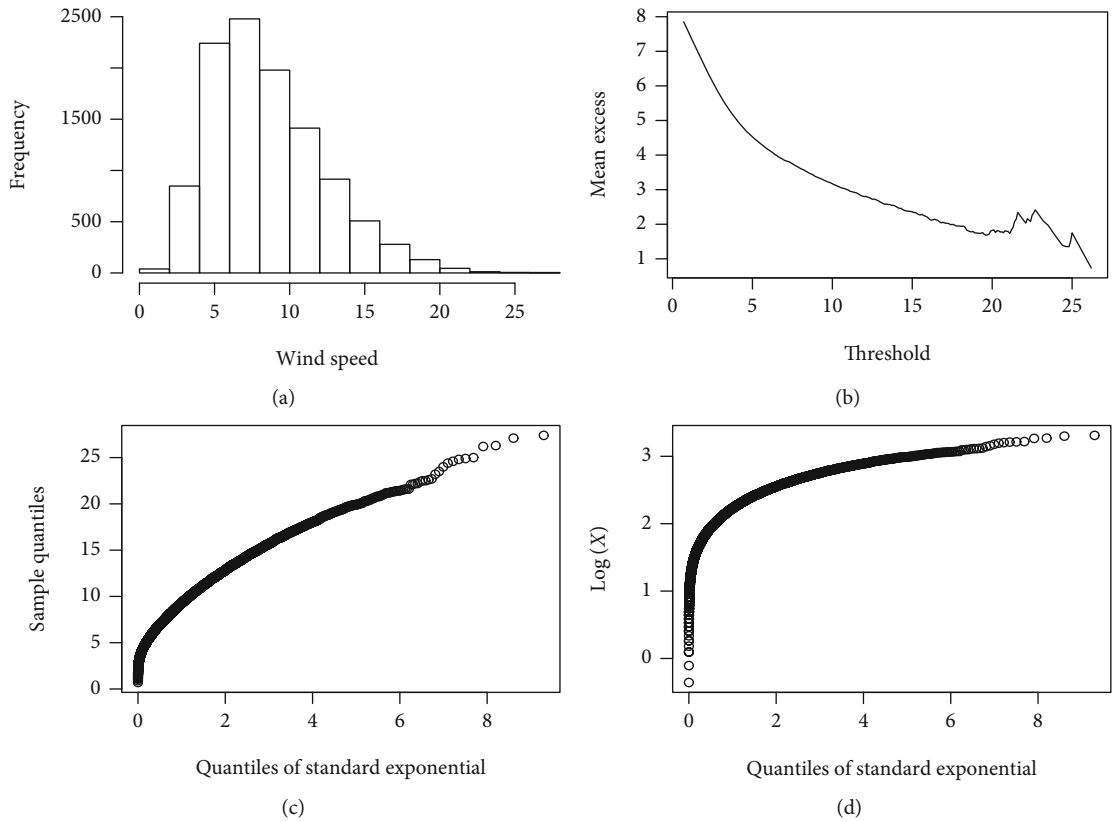


Figure 6: Preliminary analysis: (a) histogram, (b) mean excess plot, (c) exponential QQ-plot, and (d) Pareto QQ-plot.

Next, we apply the geom estimator to estimate the tail index of the available data and call this the "gold standard." Subsequently, a set of missing percentages of data relative to the sample size, i.e., 10%, 25%, and 45% were created randomly in the data using the `ampute()` function in the R package MICE. The existing modified Hill and geometric-type estimators in the literature are used to estimate the tail index of the underlying distribution of the wind speed datasets with the missing observations. However, in the application of our method, we impute the missing observations using the `mice()` function in each of the three datasets containing the various percentages of missing observations. Thereafter, we use the standard Hill and geometric-type estimators to estimate the tail index of the underlying distribution from each sample.

Figure 7 presents the results of the tail index estimators on the wind speed data as a function of the number of top order statistics, k.

For each of the estimators of the tail index, the performance is assessed on their closeness to the standard geometric-type estimator (i.e., C Geom, of which the estimation is done on a complete dataset of the wind speed data) at different values of k.

It can be seen from Figure 7 that, as k increases, almost all the estimators deviate from the standard geometric-type on complete dataset (C_Geom). From Figure 7(a) (i.e., 10% missingness), estimates of Geom and Im.Geom are almost the same as estimates of the C Geom, whereas estimates of Im.Hill and M_Hill are farther away from C_Geom. Also, in Figure 7(b) (with 25% missingness) the Im.Geom estimator is closer to C_Geom but it diverges as the number of top order statistics increases. Also, the proposed Im.Hill is closer to C_Geom than the rest of the estimators for $k \in [1000, 4000]$ and diverges beyond this range. In the case of the introduction of high percentage of missingness (i.e., 45%), Figure 7(c) shows that the estimates of Im.Hill are almost the same as those of C_Geom and quite stable compared with the other estimators of the tail index of the underlying distribution of the wind speed data.

In all the cases considered, it is evident that the M_Hill deviates more from the standard as compared to the Geom, Im.Geom, and Im.Hill. Thus, it can be ruled out as not good for the tail index of the wind speed data.

Generally, Im.Hill and Im.Geom are relatively closer to standard (C_Geom) whereas M_Hill deviates from the standard in all the scenarios. Therefore, the estimators of tail index that are based on our proposed data enhancement method can be considered as appropriate for estimating the tail index of the underlying distribution of the wind speed data. With these estimates, other parameters of extreme events such as high exceedance probability, extreme quantiles, and return periods for certain wind speeds, which are the focus of extreme value analysis, can be obtained more readily.

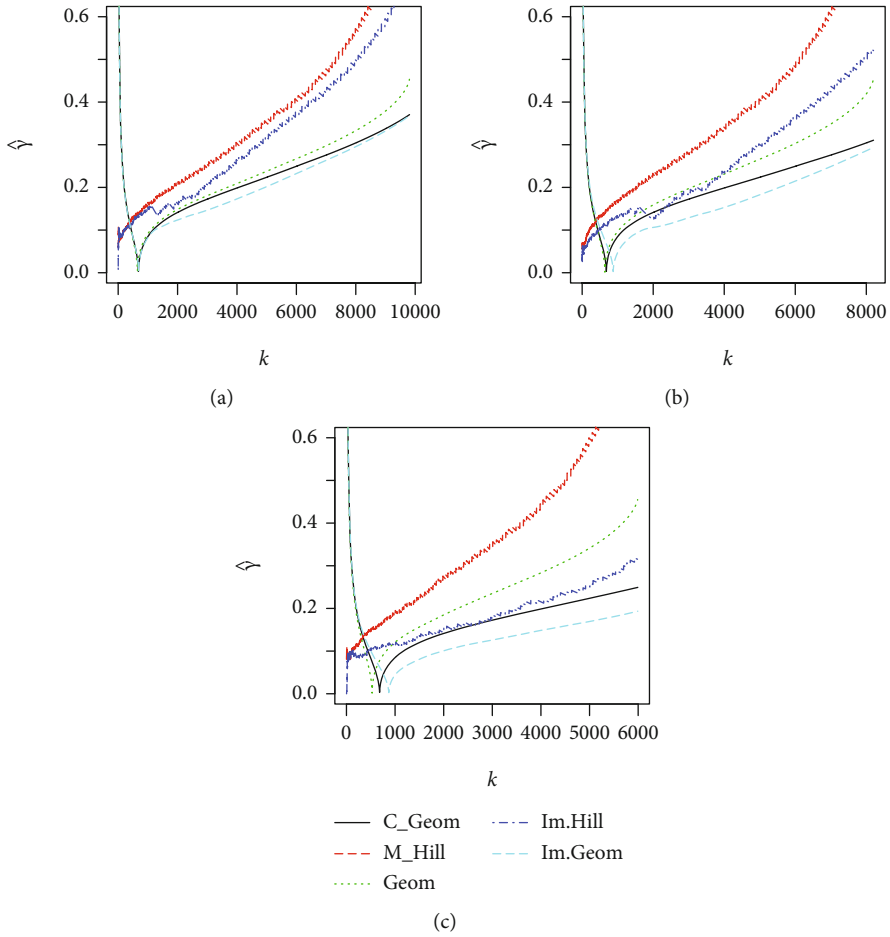


Figure 7: Tail index estimators on wind speed dataset with missing observations. (a), (b), and (c) represent 10%, 25%, and 45% missingness in the datasets, respectively.

4. Conclusion

In this paper, a data enhancement method is proposed in the estimation of tail index of an underlying distribution of a dataset when some observations are missing. This method involves imputing the missing data with an appropriate imputation method and thereafter the application of standard tail index estimators such as Hill and geometric-types. This method is contrary to the existing approach where standard estimators are augmented to use only the nonmissing part of a dataset to estimate the tail index.

The estimators based on the data enhancement method are compared with the existing estimators of tail index in the presence of missingness using a simulation study. The results of the simulation study show that no estimator is universally the best across a broad spectrum of the number of top order statistics and percentages of missingness. However, generally, the proposed estimators based on the data enhancement method exhibit smaller bias and MSE across larger spectrum of top order statistics. More importantly, in the presence of high percentage of missingness, the estimators based on the proposed data enhancement method show smaller bias and MSE and can thus be considered appropriate for estimating the tail index under missing observations.

In addition, the proposed data enhancement method of estimating tail index together with the existing estimators were illustrated with a practical dataset on wind speed. The results show that the estimators based on the data enhancement method are competitive when there are few missing observations and are more suitable when there is a high percentage of missing observations. Therefore, the imputation of missing data to obtain a semblance of the complete data offers a good approach in tail index estimation. In this regard, the MICE algorithm is recommended as a suitable imputation mechanism for enhancing the performance of tail index estimators under missingness.

Data Availability

The wind speed data used in this study is publicly available in the R package, extremefit, and it is named dataWind.

Conflicts of Interest

The authors declare that there is no conflict of interest.

References

[1] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer, London, UK, 2001.

[2] M. I. Gomes and A. Guillou, "Estimation of the extreme value index for randomly censored data," *Biometrical Letters*, vol. 48, no. 1, pp. 1–22, 2015.

[3] M. I. Gomes and A. Guillou, "Extreme value theory and statistics of univariate extremes: a review," *International Statistical Review*, vol. 83, no. 2, pp. 263–292, 2015.

[4] R. Fisher and L. Tippett, "On the estimation of the frequency distributions of the largest or smallest member of a sample," *Proceedings of the Cambridge Philosophical Society*, vol. 24, pp. 80–190, 1928.

[5] A. F. Jenkinson, "The frequency distribution of the annual maximum (or minimum) values of meteorological elements," *Quarterly Journal of the Royal Meteorological Society*, vol. 81, no. 348, pp. 158–171, 1955.

[6] A. Ameraoui, K. Boukhetala, and J. F. Dupuy, "Bayesian estimation of the tail index of a heavy tailed distribution under random censoring," *Computational Statistics & Data Analysis*, vol. 104, pp. 148–168, 2016.

[7] J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels, *Statistics of Extremes: Theory and Applications*, Wiley, England, 2004.

[8] R. Minkah, T. de Wet, and A. Ghosh, "Robust estimation of Pareto-type tail index through an exponential regression model," *Communications in Statistics - Theory and Methods*, pp. 1–19, 2021.

[9] R. Minkah, "Tail index estimation of the generalised Pareto distribution using a pivot from a transformed Pareto distribution," *Science and Development Journal*, vol. 4, no. 1, pp. 19–27, 2020.

[10] M. Ivette Gomes and D. Pestana, "A simple second-order reduced bias' tail index estimator," *Journal of Statistical Computation and Simulation*, vol. 77, no. 6, pp. 487–502, 2007.

[11] J. Beirlant, F. Figueiredo, M. Ivette Gomes, and B. Vandewalle, "Improved reduced-bias tail index and quantile estimators," *Journal of Statistical Planning and Inference*, vol. 138, no. 6, pp. 1851–1870, 2008.

[12] M. I. Gomes and M. M. Neves, "Estimation of the extreme value index for randomly censored data," *Biometrical Letters*, vol. 48, no. 1, pp. 1–22, 2011.

[13] R. Minkah, T. de Wet, and K. Doku-Amponsah, "On extreme value index estimation under random censoring," *African Journal of Applied Statistics*, vol. 5, no. 2, pp. 419–445, 2018.

[14] Y. Li and Y. Qi, "Adjusted empirical likelihood method for the tail index of a heavy-tailed distribution," *Statistics & Probability Letters*, vol. 152, pp. 50–58, 2019.

[15] Y. Ma, Y. Jiang, and W. Huang, "Empirical likelihood based inference for conditional Pareto-type tail index," *Statistics & Probability Letters*, vol. 134, pp. 114–121, 2018.

[16] R. Minkah, T. de Wet, and E. N. N. Nortey, "A simulation comparison of estimators of conditional extreme value index under right random censoring," *African Journal of Applied Statistics*, vol. 5, no. 1, pp. 337–349, 2018.

[17] P. Mladenovic and V. Piterbarg, "On estimation of the exponent of regular variation using a sample with missing observations," *Statistics & Probability Letters*, vol. 78, no. 4, pp. 327–335, 2008.

[18] I. Ilić and P. Mladenovic, "Incomplete samples and tail estimation for stationary sequences," *Novi Sad Journal of Mathematics*, vol. 38, no. 3, pp. 97–104, 2008.

[19] T. Hsing, "On tail index estimation using dependent data," *The Annals of Statistics*, vol. 19, no. 3, pp. 1547–1569, 1991.

[20] J. Zou, R. A. Davis, and G. Samorodnitsky, "Extreme value analysis without the largest values: what can be done?," *Probability in the Engineering and Informational Sciences*, vol. 34, no. 2, pp. 200–220, 2017.

[21] I. Ilić and V. M. Veličković, "Simple tail index estimation for dependent and heterogeneous data with missing values," *Brazilian Journal of Probability and Statistics*, vol. 33, no. 1, pp. 192–203, 2019.

[22] J. Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys, "Tail index estimation and an exponential regression model," *Extremes*, vol. 2, no. 2, pp. 177–200, 1999.

[23] L. de Haan and S. Resnick, "On asymptotic normality of the hill estimator," *Stochastic Models*, vol. 14, no. 4, pp. 849–866, 1998.

[24] B. Hill, "A simple general approach to inference about the tail of a distribution," *Annals of Statistics*, vol. 3, pp. 1163–1174, 1975.

[25] M. Brito and A. C. M. Freitas, "Limiting behaviour of a geometric-type estimator for tail indices," *Insurance: Mathematics and Economics*, vol. 33, no. 2, pp. 211–226, 2003.

[26] I. Ilić, "On tail index estimation using a sample with missing observations," *Statistics & Probability Letters*, vol. 82, no. 5, pp. 949–958, 2012.

[27] S. Van Buuren, *Flexible Imputation of Missing Data*, CRC press, 2018.

[28] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art," *Psychological Methods*, vol. 7, no. 2, pp. 147–177, 2002.

[29] Y. He, A. M. Zaslavsky, M. Landrum, D. Harrington, and P. Catalano, "Multiple imputation in a large-scale complex survey: a practical guide," *Statistical Methods in Medical Research*, vol. 19, no. 6, pp. 653–670, 2010.

[30] G. Durrieu, I. Grama, K. Jaunatre, Q.-K. Pham, and J.-M. Tricot, "Extremefit: a package for extreme quantiles," *Journal of Statistical Software*, vol. 87, no. 12, 2018.

[31] R. Marchigiani, S. Gordy, J. Cipolla et al., "Wind disasters: a comprehensive review of current management strategies," *International Journal of Critical Illness and Injury Science*, vol. 3, no. 2, pp. 130–142, 2013.