

# Modeling of Extremal Earthquakes

Margarida Brito, Laura Cavalcante, and Ana Cristina Moreira Freitas

**Abstract** Natural hazards, such as big earthquakes, affect the lives of thousands of people at all levels. Extreme-value analysis is an area of statistical analysis particularly concerned with the systematic study of extremes, providing an useful insight to fields where extreme values are probable to occur. The characterization of the extreme seismic activity is a fundamental basis for risk investigation and safety evaluation. Here we study large earthquakes in the scope of the Extreme Value Theory. We focus on the tails of the seismic moment distributions and we propose to estimate relevant parameters, like the tail index and high order quantiles using the geometric-type estimators.

In this work we combine two approaches, namely an exploratory oriented analysis and an inferential study. The validity of the assumptions required are verified, and both geometric-type and Hill estimators are applied for the tail index and quantile estimation. A comparison between the estimators is performed, and their application to the considered problem is illustrated and discussed in the corresponding context.

## 1 Introduction

Earthquakes are a worldwide and ever present menace, threatening to occur at any second. A severe earthquake is one of the most frightening and destructive phenomena of nature. Experiencing an earthquake is a terrible experience, the lived moments are reported as full of panic, terror, and death. For survivors, the terrible

---

M. Brito (✉) • L. Cavalcante

Faculdade de Ciências, Universidade do Porto, Porto, Portugal  
e-mail: [mabrito@fc.up.pt](mailto:mabrito@fc.up.pt); [laucavalcante@fc.up.pt](mailto:laucavalcante@fc.up.pt)

A.C.M. Freitas

Faculdade de Economia, Universidade do Porto, Porto, Portugal  
e-mail: [amoreira@fep.up.pt](mailto:amoreira@fep.up.pt)

images remain in their memory and become part of their daily lives, as well as the constant fear of the possibility of the next big earthquake which may take lives and separate families forever. It is estimated that there are about one million earthquakes per year. However, the vast majority occurs in the midst of oceans or in sparsely populated regions, and they pass relatively unnoticed by the population. There are annually about 20 earthquakes that cause significant damage. On average, only one catastrophic earthquake occurs every year, and a highly catastrophic one every 5 years.

Since the underlying phenomena responsible for the occurrence of an earthquake are still very far from being completely understood, it is rather important to collect as much data as possible and categorize it in order to be able to provide some insight on how to diminish their negative impacts, in particular, in what concerns the reduction of number of deaths and economic losses. This is an important challenge requiring a large multidisciplinary effort. In this work, we perform a statistical analysis taking into account specific features of big earthquakes. When we are dealing with extreme events, the classical statistical models are inappropriate for the statistical modeling of earthquake size. Hence, we are particularly interested in the study of the tail distribution of the data.

The Extreme Value Theory (EVT) is one field of statistics that has been devised to study these extreme events using only a limited amount of data (see e.g. [1], and references therein). In the study of earthquakes, the EVT is a relevant tool, providing important information, such as the estimation of the probability of occurring a large earthquake over a long period of time or high quantiles (see e.g. [22]).

In the present work we consider the seismic activity in Philippines and Vanuatu Islands. The data sets are taken from the Harvard Seismic Catalog and the tail behavior of the distributions of large earthquakes seismic moments is characterized using EVT techniques. In order to apply these methods, a preliminary data analysis is performed to investigate the validity of the usual underlying assumptions. The geometric-type and the Hill estimator, as well as its bias corrected versions, are considered for the estimation of the tail index and are employed for the quantile estimation. A comparison between the estimators is carried out and their performance is discussed carefully.

All the analysis is supported by graphical tools that show, in a clear way, the features of the data that are regarded as most relevant to the study being addressed.

The paper is organized as follows. Some important concepts and results about EVT and earthquakes are briefly presented in Sect. 2. The investigation, in order to verify the validity of the usual assumptions and the analysis of the seismic moments, are performed in Sect. 3. Some final comments about the study, including an interpretation of the results in terms of the frequencies of seismic moment exceedances, are provided in Sect. 4.

## 2 Essential Notions of EVT and Earthquakes

### 2.1 Extreme Value Theory

The Extreme Value Theory is a powerful and fairly robust framework to study the tail behavior of a distribution, since it encompasses a set of probabilistic results that allow characterizing and modeling the extreme values behavior. In this way, the EVT is very useful in making statistical inferences about rare events in several areas of knowledge (e.g. meteorology, hydrology, insurance, environment, etc.), and its use may enable the implementation of appropriate prevention procedures.

More concretely, through this theory, extreme values may be modeled using the limiting distribution of the maxima of the random variables or of its excesses over a threshold. Thus, the statistical basis for applications of EVT is constituted by the following two main limit theorems.

**Theorem 1 (Fisher-Tippett-Gnedenko Theorem)** *Let  $X_1, X_2, \dots, X_n$  be independent and identically distributed (i.i.d.) random variables (r.v.) with distribution function (d.f.)  $F$  and  $M_n = \max(X_1, X_2, \dots, X_n)$  denote the maximum of the  $n$  observations. If a sequence of real numbers  $a_n > 0$  and  $b_n$  exists such that*

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

then if  $G$  is a non degenerate d.f., it belongs to one of the following types

$$\text{Type I (Gumbel)} : \Lambda(x) = \exp\{-\exp(-x)\}, x \in \mathbb{R};$$

$$\text{Type II (Fréchet)} : \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0, \\ \exp(-x^{-1/\alpha}), & x > 0; \end{cases}$$

$$\text{Type III (Weibull)} : \Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^{1/\alpha}\}, & x > 0, \\ 1, & x \geq 0; \end{cases}$$

for all continuity points of  $G$ .

If a d.f.  $F$  satisfies the conditions of the theorem, it is said that  $F$  belongs to the domain of attraction of  $G$  ( $F \in DA(G)$ ).

These three types of distributions may be combined into the single d.f.

$$G_\gamma(x) = \begin{cases} \exp\left(- (1 + \gamma x)^{-1/\gamma}\right), & \text{for } 1 + \gamma x > 0, \gamma \neq 0, \\ \exp(-\exp(-x)), & \text{for } x \in \mathbb{R}, \gamma = 0, \end{cases}$$

where  $\gamma$  is the shape parameter, known as tail index, determining the weight of the right tail of the underlying d.f.  $F$ . This distribution is known as the Generalized Extreme Value (GEV) distribution.

**Theorem 2 (Pickands-Balkema-de Haan Theorem)** *Let  $X_1, X_2, \dots, X_n$  be a sample of  $n$  i.i.d. r.v. with d.f.  $F$ ,  $x^F$  the right endpoint of  $F$  and  $F_{X-u|X>u}(x) = P\{X-u \leq x \mid X > u\}$  the excess d.f. over a (high) threshold  $u$ . Then,*

$$F \in DA(G_\gamma) \text{ iff } \lim_{u \rightarrow x^F} \sup_{0 \leq x < x^F - u} |F_{X-u|X>u}(x) - H_{\gamma, \sigma_u}(x)| = 0,$$

where  $H_{\gamma, \sigma_u}(x)$  represents the Generalised Pareto Distribution, given by:

$$H_{\gamma, \sigma_u}(x) = \begin{cases} 1 - (1 + \gamma \frac{x-u}{\sigma_u})^{-1/\gamma}, & \text{for } 1 + \gamma \frac{x-u}{\sigma_u} > 0, \gamma \neq 0, \\ 1 - \exp(-\frac{x-u}{\sigma_u}), & \text{for } x \geq u, \gamma = 0, \end{cases}$$

where  $\gamma, u, \sigma_u > 0$  are the shape, location, and scale parameter depending on threshold  $u$ , respectively.

Similarly with GEV, using another parameterization, the GPD is separated into three families depending on the value of the shape parameter:

- Type I (Exponential):  $H(x) = 1 - \exp(-x)$ , if  $\gamma = 0$ ,
- Type II (Pareto):  $H(x) = 1 - x^{-1/\gamma}$ , if  $\gamma > 0$ ,
- Type III (Beta):  $H(x) = 1 - (-x)^{-1/\gamma}$ , if  $\gamma < 0$ .

These two theorems state that, under their conditions, the limit distribution of the normalized maximum is the GEV distribution, and that the limit of the excess d.f. is the GPD. Hence, they are fundamental to make possible real-world applications.

In order to perform a correct inference about extreme events from the accessible data, it is necessary to properly select the extreme observations following some criterion. There are two primary methods to define such extreme observations which arise from the two main results of the classical EVT: the Block Maxima method, also known as Gumbel's approach, and the Peaks Over Threshold method.

The Block Maxima (BM) method consists in dividing the data in equal sized blocks, taking the maximum observation in each block and studying its asymptotic distribution. In the Peaks Over Threshold (POT) method one considers a certain high threshold and then studies the asymptotic distribution of the excesses over this high threshold.

Accordingly, as with the data set under study, one must be aware to consider both methods' disadvantages when applying them. One major drawback of the BM method is that only one observation in a block is used, resulting in a final sample of small size. On other hand, this method is more robust with respect to eventual dependence between the observations.

Since our interest is centered in the frequencies of exceedances of certain critical values, here we adopt the POT approach that picks up all relevant high observations and seems to make better use of the available information.

In modeling the extreme value distribution, the main issue to be solved is the parameter estimation. The shape parameter  $\gamma$  is of great interest in the analysis of the tails, since it characterizes the behavior of extremes. This parameter indicates the heaviness of the tail distribution, the tail being heavier for larger values  $\gamma$ . It also plays a crucial role in the estimation of other extreme events' parameters, namely in high quantiles estimation. In practice, the tail index is associated to the frequency with which extreme events occur and the high order quantiles are levels that are exceeded with a small probability. The adequate estimation of these quantities is the most important problem.

We assume that  $X_1, X_2, \dots, X_n$  is a sample of i.i.d. r.v. with d.f.  $F$  and denote by  $X_{(1,n)} \leq X_{(2,n)} \leq \dots \leq X_{(n,n)}$  the corresponding order statistics (o.s.). The estimation of  $\gamma$  is based on the  $k$  top o.s., where  $k = k_n$  is an intermediate sequence of positive integers ( $1 \leq k < n$ ), that is,

$$k \rightarrow \infty, \quad \frac{k}{n} \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty. \tag{1}$$

Several estimators have been proposed for the estimation of  $\gamma$  (see e.g. [6, 10, 18, 20]). Here we consider the following estimator for  $\gamma > 0$ , the geometric-type (GT) estimator

$$\widehat{GT}(k) = \sqrt{\frac{M_n^{(2)} - [M_n^{(1)}]^2}{\frac{1}{k} \sum_{i=1}^k \log^2(n/i) - \left(\frac{1}{k} \sum_{i=1}^k \log(n/i)\right)^2}} \tag{2}$$

where

$$M_n^{(j)}(k) = \frac{1}{k} \sum_{i=1}^k (\log X_{(n-i+1,n)} - \log X_{(n-k,n)})^j. \tag{3}$$

We also consider the commonly used Hill estimator (see [18]) defined by

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

The asymptotic properties of these aforementioned estimators were investigated and, under certain conditions, they share some common desirable properties, such as consistency and asymptotic normality (cf. [2, 9, 17]).

The problem of estimating high order quantiles has received increased attention as a useful tool in data modeling, which has been performed in a wide variety of problems in many different scientific areas. This field addresses interesting

questions such as the size of some extreme event that will only occur with a given small probability, or the expected time until the realization of an extreme event.

The classical quantile estimator was proposed by [23],

$$\hat{\chi}_{1-p}^W = X_{(n-k,n)} \left( \frac{k}{np} \right)^{\hat{\gamma}},$$

where  $\hat{\gamma}$  is a consistent estimator of  $\gamma$ .

Using general quantile techniques and the POT methodology, the well known POT estimator for high quantiles above the threshold  $X_{(n-k,n)}$  arises naturally and is given by

$$\hat{\chi}_{1-p}^P = \frac{\left( \frac{k}{np} \right)^{\hat{\gamma}} - 1}{\hat{\gamma}} \cdot X_{(n-k,n)} M_n^{(1)} + X_{(n-k,n)}, \quad p < \frac{k}{n}, \quad (5)$$

where  $\hat{\gamma}$ ,  $X_{(n-k,n)} M_n^{(1)}$  and  $u = X_{(n-k,n)}$  are, respectively, suitable estimators of the shape, scale and location parameters of the Generalized Pareto Distribution.

In the present work both the  $\widehat{GT}(k)$  and  $\hat{H}(k)$  are used to estimate  $\gamma$ . The high quantiles are estimated considering (5) and using  $\widehat{GT}(k)$  and  $\hat{H}(k)$  as estimators of  $\gamma$ . The asymptotic behavior of these quantile estimators was studied and their asymptotic normality was proved (cf. [3, 8, 10]).

The problem of reducing the bias of these tail index estimators was addressed in [3], where were proposed the following two asymptotic equivalent geometric-type bias corrected estimators

$$\overline{\widehat{GT}}(k) = \widehat{GT}(k) \left( 1 - \frac{\beta \left( \frac{n}{k} \right)^\rho}{(1-\rho)^2} \right),$$

and

$$\overline{\overline{\widehat{GT}}}(k) = \widehat{GT}(k) \exp \left\{ -\frac{\beta}{(1-\rho)^2} \left( \frac{n}{k} \right)^\rho \right\}.$$

Hill bias corrected estimators may be found in [4], namely

$$\overline{\hat{H}}(k) = \hat{H}(k) \left( 1 - \frac{\beta \left( \frac{n}{k} \right)^\rho}{1-\rho} \right)$$

and

$$\overline{\overline{\hat{H}}}(k) = \hat{H}(k) \exp \left\{ -\frac{\beta}{1-\rho} \left( \frac{n}{k} \right)^\rho \right\},$$

where  $\rho$  and  $\beta$  are the shape and scale parameters.

Here, in order to get bias corrected high quantiles estimators, we also consider the form (5), based on the above bias corrected estimators.

The accurate estimation of the tail index is very important, also because of its great influence on the estimation of other relevant parameters of rare events, such as the right endpoint of the underlying d.f.  $F$ . Since the impact of its influence can be considerable, the appropriate estimation of  $\gamma$  is fundamental in obtaining a suitable quantile estimator with a good performance.

## 2.2 Earthquakes

In general, everything in nature tends to an equilibrium. Due to the thermodynamic equilibrium, the constituents of the Earth's interior are in constant motion. Boosted by this movement, which causes friction with its bottom, the tectonic plates move and interchange slowly, thereby contributing to the constant evolution of the terrestrial relief.

The earthquakes mainly arise due to forces, within the earth's crust, tending to displace one mass of rock relative to another. Each time the plates interact with each other, a large amount of energy is accumulated in its rocks. When its elasticity limit is reached, they will fracture and instantly release all the energy that had been accumulated during the elastic deformation. That causes vibrations, called seismic waves, which travel outwards in all directions from the fault and give rise to violent motions at the earth's surface, unleashing an earthquake.

Therefore, earthquakes are natural shocks that occur as a result of this sudden release of huge amounts of the energy that has been slowly-accumulated over many years. If the earthquake is large enough, the seismic waves are recorded on seismographs around the world, and can cause the ground to quake strongly.

Earthquakes do not occur at random, but are distributed according to a well-defined pattern. About 90 % of earthquake activity is associated with plate-boundary processes, so the global seismicity patterns reveals a strong correlation between plate boundaries and the presence of intercontinental fault zones, indicating that earthquakes often occur at tectonic plate boundaries. We can say, without committing a gross error, that the alignments of earthquakes indicate the boundaries of tectonic plates.

After the initial fracture, a number of secondary ruptures, corresponding to the progressive adjustment of fractured rocks, may occur, causing successive lower intensity earthquakes called aftershocks. If these vibrations occur at the sea floor, they can produce a long and smooth waving that in shallow water becomes authentic water columns known as tidal waves or tsunamis.

Therefore, earthquakes represent one of the most energetic and rapid manifestations of the planet's internal dynamics.

The scientific analysis of earthquakes requires means of measurement, and the size of an earthquake has been measured in several ways. The early methods used a kind of numerical scale based on a synthesis of observed effects, called the

*intensity* scales. Some attempts to relate intensity to the amplitude of ground motion led to a quantity called *magnitude*, based on the records of ground amplitudes, normalized for their variation with regard to the distance from the earthquake epicenter. However, the known magnitudes present a saturation point which does not allow for a correct estimation of the true earthquake size for larger earthquakes, underestimating it. Moreover, it turns out that larger earthquakes, which have larger rupture surfaces, systematically radiate more long-period energy. Nowadays, the measurement that is adopted preferably for scientific studies is the *seismic moment* of the displaced ground (see e.g. [7, 19]). This measurement avoids the saturation problem, since it does not have an intrinsic upper bound, and describes the size of an earthquake as an essential combination of physical quantities.

The seismic moment,  $M$ , provides more accurate measures of the energy released from an earthquake, taking into account the rock properties, such as its rigidity,  $\mu$ , the area of the fault plane that actually moves,  $A$ , and the amount of movement on the fault,  $D$ , combining these three factors in the following form

$$M = \mu AD.$$

Because many people do not really know the meaning of this measure, and given that the magnitude scale has been used for a very long time, the need to convert it into some kind of magnitude scale came about. These factors have resulted in the definition of a new magnitude scale, the moment magnitude,  $m_w$ , based on the seismic moment

$$m_w = \frac{2}{3} (\log M - 16.1), \quad (6)$$

where  $M$  is in units of dyne-cm.

The seismic moment, based on classical mechanics, provides, in this way, a uniform scale of earthquake size, and is considered the most consistent measure for accurate quantification of the energy released from an earthquake.

### 3 Extreme Value Modeling of Earthquake Data

In this section, we analyze the tail behavior of the distribution of the seismic moments, following the POT approach. We begin by describing the data considered for this study. We perform an exploratory data analysis, where we discuss which type of distribution may model the large seismic moments as well as the properties of stationarity and independence of the data. Then we proceed to the estimation of the tail parameters of the seismic moment distribution.



### 3.1 Description of the Earthquake Data

We consider the earthquake data obtained from the Harvard Seismic Catalog, available at the Global Centroid-Moment-Tensor (CMT) web page (cf. e.g. [11, 12, 14]). Here, we restrict the area of study to earthquakes occurring within the Philippines and Vanuatu Islands, and the analysis was performed in a similar way for both regions. In particular, we extract and analyze the information on their seismic moments covering the period 01.01.1976–31.12.2010. The original data sets contain 1255 events for Philippines Islands, and 1012 events for Vanuatu Islands. However, in order to apply the POT method we selected an adequate and large enough level  $u = 10^{24}$  dyne-cm, that corresponds to a moment magnitude  $m_w \approx 5.27$ , the same value considered in related works such as in [21]. The observations under this threshold were removed. Since we detect a failure in the data acquisition of the Vanuatu Islands until 01-01-1980, we shall consider only the Vanuatu Islands data subsequent to this date. So, the final data sets, on which the following analysis is based, consider 821 cases for Philippines Islands and 647 cases for Vanuatu Islands. We did not exclude aftershocks because, besides excluding a great fraction of the range of seismic moments considered, the removal would introduce a bias in the parameters estimation (cf. e.g. [21]). Since the considered region has a lot of deep earthquakes, they were not excluded as well. In Fig. 1 the seismic moments of Philippines and Vanuatu Islands over the above mentioned period are plotted.

### 3.2 Preliminary Data Analysis

Before considering the problem of estimating the tail parameter  $\gamma$ , it is important to discuss if the Pareto-type model provides a plausible fit to the seismic moment

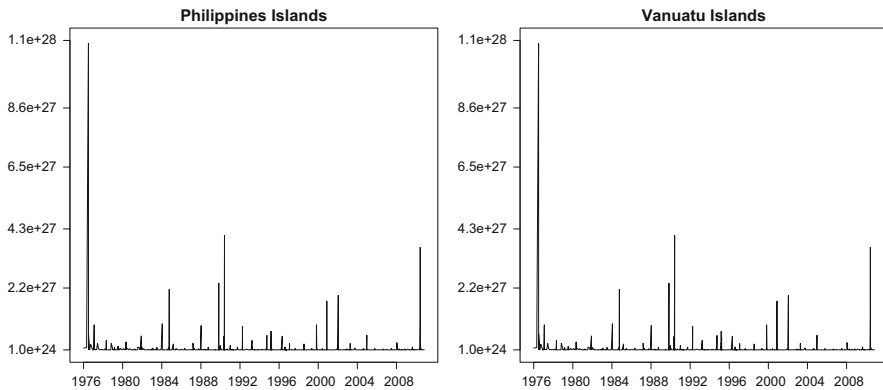
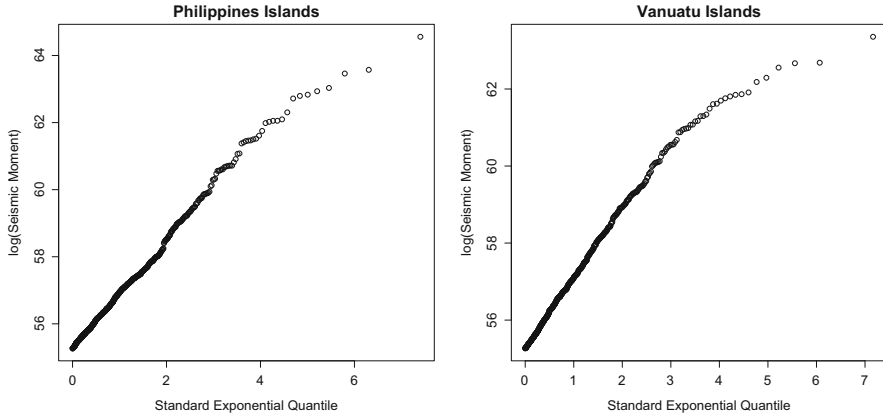


Fig. 1 Seismic moments of Philippines (*left*) and Vanuatu (*right*) Islands



**Fig. 2** Pareto QQ plot for Philippines (*left*) and Vanuatu (*right*) Islands seismic moment data

distributions of the data under study. This can be achieved graphically through quantile-quantile (QQ) plots, which constitute a very informative and powerful tool to graphically evaluate how close two distributions are from each other.

Usually, as in this case, the most convenient comparison is between the empirical quantiles and the quantiles of the assumed parametric distribution. If the sample data and the reference distribution are derived from populations with a common distribution, the QQ plot should have a linear form.

Since we believe our data is heavy tailed, we present the Pareto QQ plots of our data sets in Fig. 2.

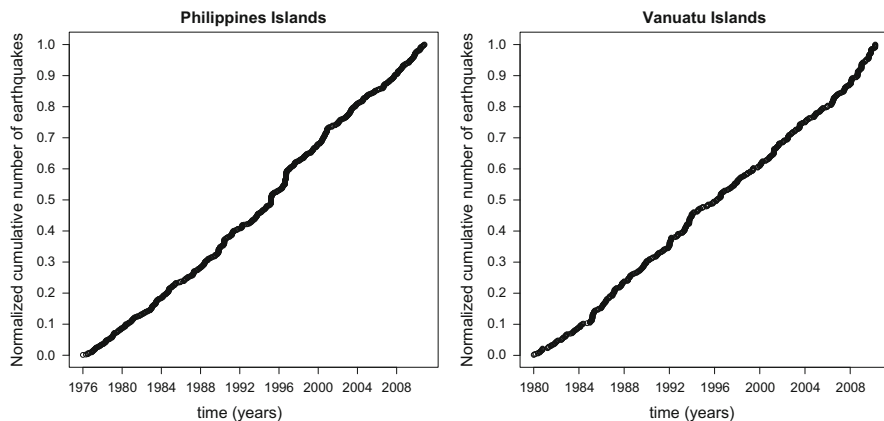
In the case  $Y \stackrel{D}{=} \log X$ , where  $X$  and  $Y$  are Pareto and Exponential distributed r.v., respectively, then the usual Pareto QQ plots are Exponential QQ plots of the log-transformed data.

In the resulting scatterplot, a linear pattern is evident, which is indicative of the good agreement between observed values and the values predicted by the model. If we analyze the behavior of the QQ plots, we may remark that, with the exception of the extreme upper points, which are based on a small number of extreme values, the plots are approximately linear. Hence, the visual impressions based on the Pareto QQ plots suggest that the Vanuatu and Philippines Islands earthquake data sets do seem to exhibit heavy tails ( $\gamma > 0$ ).

We analyse the stationarity of the data under study. More precisely, in the line of the study of Corral [5], we investigate if the mean value defined for any property of the earthquake occurrence process is approximately the same for different time windows. We plot the normalized cumulative number of earthquakes versus time.

The linear behavior that we can observe in Fig. 3 indicates that the mean seismic rate is approximately constant, and so, the data may be considered homogeneous in time.

For the application of the EVT we must analyse the independence of the data.



**Fig. 3** Cumulative number of earthquakes normalized by the total number in the period considered as a function of time, for seismicity of Philippines (*left*) and Vanuatu (*right*) Islands with  $M \geq 10^{24}$

In our case, the goal is to investigate the existence of dependence between consecutive seismic moments, i.e., verify how the seismic moment of one event,  $M_{i-1}$ , influences the seismic moment of the next,  $M_i$ . For that, let us consider the conditional probability density determined by

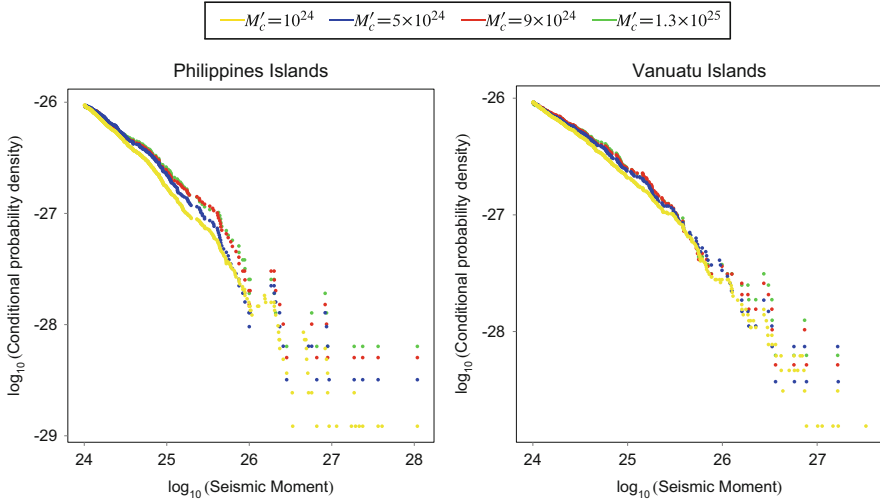
$$\frac{P(\eta \leq M_i < \eta + \Delta_\eta \mid M_{i-1} \geq M'_c)}{\Delta_\eta},$$

where  $M'_c$  is the threshold considered on the previous magnitude when this condition is imposed. Here we denote the initial threshold,  $u$ , as  $M_c$ , and the condition  $M \geq M_c$  is always satisfied (see e.g. [5]).

The conditional probability density of a seismic moment is then defined as the probability of the seismic moments are within a small interval of values, divided by the length of the small interval,  $\Delta_\eta$ , tending to zero, considering only the cases in which the seismic moment of the immediately previous event is bigger than a threshold  $M'_c$ .

If the seismic moment  $M_i$  is independent of  $M_{i-1}$ , then, as it is well known, the conditional distribution of  $M_i$  given that  $M_{i-1} \geq M'_c$ ,  $M'_c \geq M_c$ , is identical to the unconditional distribution of  $M_i$ . Note that the case  $M_c = M'_c$  gives the unconditional distribution of the considered data.

We observe in Fig. 4 that, in general, the different empirical densities, using different thresholds  $M'_c$ , share the same properties, which suggest the independence of seismic moments  $M_i$  with regards to their history. The small oscillations between the densities may be caused by the errors associated to the finite sample and the eventual dependence is apparently too weak to lead to major differences in the distributions.



**Fig. 4** Conditional probability densities of earthquake seismic moments, for seismicity of Philippines (*left*) and Vanuatu (*right*) Islands, evaluated using different thresholds  $M'_c$  and with a constant  $M_c = 10^{24}$  ( $\Delta_\eta = 10^{25}$ )

### 3.3 Estimation of Tail Parameters

In this section we formalize our main objective of investigating the extremal behavior of large earthquakes and how the proposed estimators behave with this type of data.

Then, we discuss the estimation of the tail parameters through the POT approach. The GT and the Hill estimators are considered for the estimation of the tail index and are employed on POT estimator for the quantile estimation.

Some graphical plots illustrate the tail parameters of large earthquake data, as a function of  $k$ .

From the presented bias corrected estimators, we can easily note that the bias dominant components are dependent on second order parameters, shape  $\rho$  and scale  $\beta$ . To illustrate the behavior of the corrected estimators we consider the suitable estimators of the parameter  $\rho$  proposed by [13]

$$\hat{\rho}_n^{(\tau)}(k) = - \left| \frac{3(T_n^{(\tau)}(k) - 1)}{T_n^{(\tau)}(k) - 3} \right|, \quad (7)$$

where

$$T_n^{(\tau)}(k) = \begin{cases} \frac{(M_n^{(1)}(k))^\tau - (M_n^{(2)}(k)/2)^{\tau/2}}{(M_n^{(2)}(k)/2)^{\tau/2} - (M_n^{(3)}(k)/6)^{\tau/3}}, & \text{if } \tau > 0 \\ \frac{\log(M_n^{(1)}(k)) - \frac{1}{2} \log(M_n^{(2)}(k)/2)}{\frac{1}{2} \log(M_n^{(2)}(k)/2) - \frac{1}{3} \log(M_n^{(3)}(k)/6)}, & \text{if } \tau = 0, \end{cases}$$

with  $M_n^j$  as in (3), and the  $\beta$  estimator obtained in [15]

$$\hat{\beta}_{\hat{\rho}}(k) = \left(\frac{k}{n}\right)^{\hat{\rho}} \frac{\left(\frac{1}{k} \sum_{i=1}^k \left(\frac{i}{k}\right)^{-\hat{\rho}}\right) \frac{1}{k} \sum_{i=1}^k U_i - \frac{1}{k} \sum_{i=1}^k \left(\frac{i}{k}\right)^{-\hat{\rho}} U_i}{\left(\frac{1}{k} \sum_{i=1}^k \left(\frac{i}{k}\right)^{-\hat{\rho}}\right) \frac{1}{k} \sum_{i=1}^k \left(\frac{i}{k}\right)^{-\hat{\rho}} U_i - \frac{1}{k} \sum_{i=1}^k \left(\frac{i}{k}\right)^{-2\hat{\rho}} U_i}, \quad (8)$$

where

$$U_i = i \left( \log \frac{X_{(n-i+1,n)}}{X_{(n-i,n)}} \right),$$

with  $1 \leq i \leq k < n$ .

It is known that the external estimation of  $\rho$  and  $\beta$  at a larger  $k$  value than the one used for  $\gamma$ -estimation has clear advantages, allowing bias reduction without increasing the asymptotic variance (see e.g. [4]). In line with other studies, and among some suggestions (see e.g. [16]), the level that seemed most appropriate to consider in illustrations is

$$k_h = \lfloor n^{1-\epsilon} \rfloor, \text{ for some } \epsilon > 0 \text{ small}, \quad (9)$$

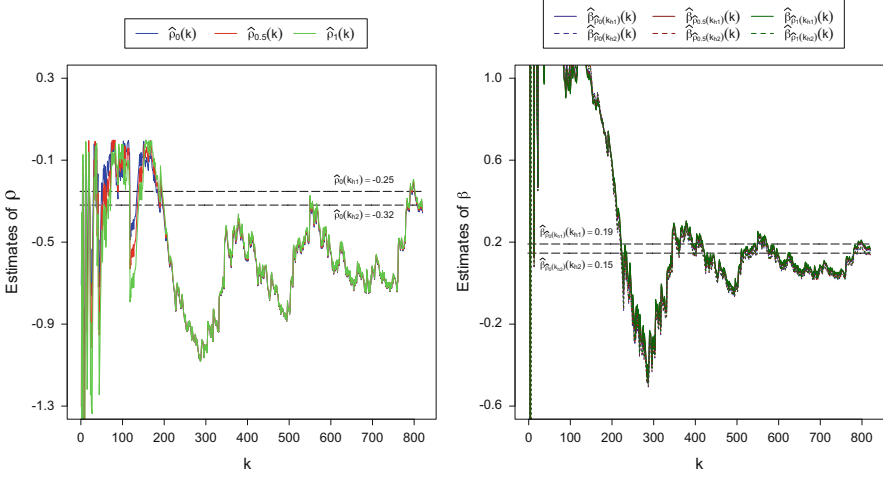
where  $\lfloor x \rfloor$  denotes the integer part of  $x$ .

We remark that the class of estimators of  $\rho$  presented above, and consequently also the  $\beta$  estimators, is dependent on a tuning parameter  $\tau \geq 0$ . Then, firstly we need to choose the tuning parameter  $\tau$ , in which we will support the estimation of the second order parameters  $\rho$  and  $\beta$ .

For this use, we consider in (9),  $\epsilon = 0.005$  and  $\epsilon = 0.001$ , i.e, we use the following  $k_h$  levels:

$$k_{h1} = \lfloor n^{0.995} \rfloor \quad \text{and} \quad k_{h2} = \lfloor n^{0.999} \rfloor. \quad (10)$$

As usual, the means whereby we do this choice, passes by portraying the sample paths of  $\hat{\rho}_\tau(k)$  in (7) for the values  $\tau \in \{0, 0.5, 1\}$ , as functions of  $k$ , in order to analyze the variations that it causes in their behavior, and use the following



**Fig. 5** Estimates of the second order parameters  $\rho$  (left) and  $\beta$  (right) for seismicity of Philippines Islands

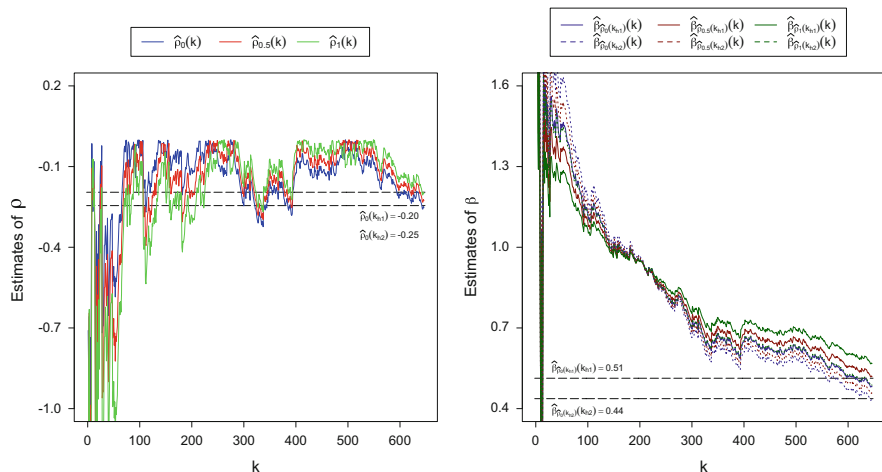
algorithm as a stability criterion for large values of  $k$ :

1. Consider  $\hat{\rho}_\tau(k)$ ,  $\tau \in \{0, 0.5, 1\}$ , for the integer values  $k \in ([n^{0.995}], [n^{0.999}])$  and compute their median, denoted by  $\chi_\tau$ ;
2. Choose the *tuning parameter*  $\tau^* = \arg \min_\tau \sum_k (\hat{\rho}_\tau(k) - \chi_\tau)^2$ ;
3. Compute the  $\rho$  estimates  $\hat{\rho}_{\tau^*}(k_{h1})$  and  $\hat{\rho}_{\tau^*}(k_{h2})$ , and the  $\beta$  estimates  $\hat{\beta}_{\rho_{\tau^*}(k_{h1})}(k_{h1})$  and  $\hat{\beta}_{\rho_{\tau^*}(k_{h2})}(k_{h2})$ , with  $k_{h1}$  and  $k_{h2}$  given by (10).

The Figs. 5 and 6 show the sample paths of the second order parameter estimators,  $\hat{\rho}$  and  $\hat{\beta}$ , based on the Philippines and Vanuatu seismic moment observations, respectively.

We can see that the sample paths of  $\hat{\rho}$ , for the three different values of  $\tau$ , have very similar behavior. It is however apparent that the behavior of  $\hat{\rho}$  is slightly better when considering  $\tau = 0$ , especially for data concerning the Vanuatu Islands. Since in both cases the algorithm described above also points to the choice of  $\tau = 0$ , we choose this value of  $\tau$  to estimate  $\rho$ .

Thus, for Philippines Islands, we have  $k_{h1} = [821^{0.995}] = 793$  and  $k_{h2} = [821^{0.999}] = 815$ , that is, the corresponding estimates of  $\rho$  are  $\hat{\rho}_0(793) \approx -0.25$  and  $\hat{\rho}_0(815) \approx -0.32$  and the corresponding estimates of  $\beta$  are  $\hat{\beta}_{\hat{\rho}_0(793)}(793) \approx 0.19$  and  $\hat{\beta}_{\hat{\rho}_0(815)}(815) \approx 0.15$ , represented both graphically through straight lines. Doing the same procedure to Vanuatu Islands, we have  $k_{h1} = [647^{0.995}] = 626$  and  $k_{h2} = [647^{0.999}] = 642$ , that is, the corresponding estimates of  $\rho$  are  $\hat{\rho}_0(626) \approx -0.20$  and  $\hat{\rho}_0(642) \approx -0.25$  and the corresponding estimates of  $\beta$  are  $\hat{\beta}_{\hat{\rho}_0(626)}(626) \approx 0.51$  and  $\hat{\beta}_{\hat{\rho}_0(642)}(642) \approx 0.44$ .



**Fig. 6** Estimates of the second order parameters  $\rho$  (left) and  $\beta$  (right) for seismicity of Vanuatu Islands

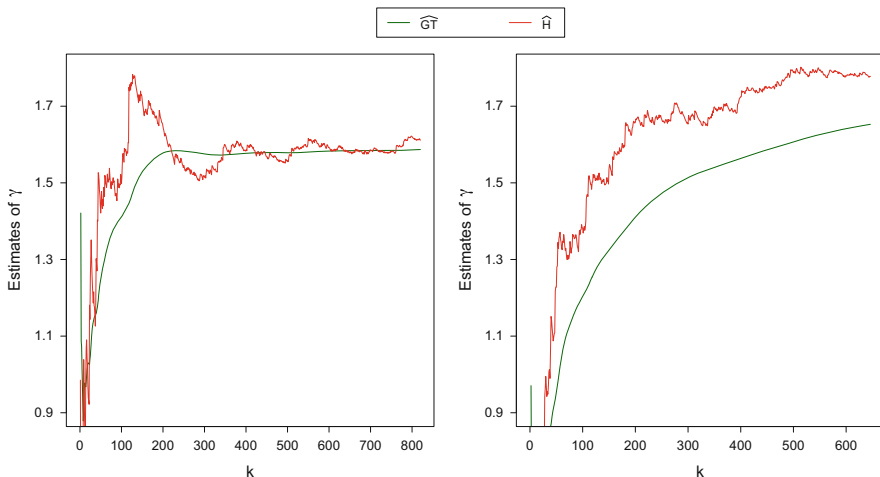
Since from the  $\hat{\beta}$  sample paths, there are no readily apparent significant differences between the use of  $k_{h1}$  or  $k_{h2}$ , and due to the fact that the tail index estimation is more affected by the  $\rho$  fluctuations than the  $\beta$  ones, we use both levels in the rest of the study.

Moreover, here we also present a possible optimal level  $k_0$  of top observations to consider when the geometric-type estimator is used to estimate  $\gamma$ , through the minimization of the asymptotic mean square error (AMSE) of the geometric-type estimator. Considering the following distributional representation of the geometric-type estimator (see [3, Theorem 2.2]).

$$\widehat{GT}(k) \stackrel{D}{=} \gamma + \frac{\gamma}{2\sqrt{k}}Q_n - \frac{\gamma}{\sqrt{k}}P_n + \frac{A\left(\frac{n}{k}\right)}{(1-\rho)^2} + o_p\left(A\left(\frac{n}{k}\right)\right) + O_p\left(\frac{\log^2 k}{k}\right),$$

we get what we need to calculate the  $AMSE(\widehat{GT})$  and provide for their minimization

$$\begin{aligned} \frac{\partial}{\partial k} [AMSE(\widehat{GT})] = 0 &\iff \frac{\partial}{\partial k} \left[ V(\widehat{GT}) + (Bias(\widehat{GT}))^2 \right] = 0 \\ &\iff \frac{\partial}{\partial k} \left[ \frac{2\gamma^2}{k} + \left( \frac{\gamma\beta}{(1-\rho)^2} \right)^2 \left( \frac{n}{k} \right)^{2\rho} \right] = 0. \end{aligned}$$



**Fig. 7** Plot for the GT estimator,  $\widehat{GT}$ , and for the Hill estimator,  $\widehat{H}$ , of  $\gamma$ , for seismicity of Philippines (*left*) and Vanuatu (*right*) Islands

Solving the equation in order to  $k$  and denoting the result as  $k_0^{\widehat{GT}}$ , we obtain

$$k_0^{\widehat{GT}} = \left[ \frac{(1 - \rho)^2}{-2\rho\beta^2} \right]^{1/(1-2\rho)} n^{-2\rho/(1-2\rho)}.$$

Although this is not the optimal value for the bias corrected estimators, the value of the tail index and quantiles calculated with the geometric-type estimator at the  $k_0^{\widehat{GT}}$  level is represented in some illustrations for comparison.

As a first step, we estimate the tail index,  $\gamma$ , using the GT and Hill’s estimators.

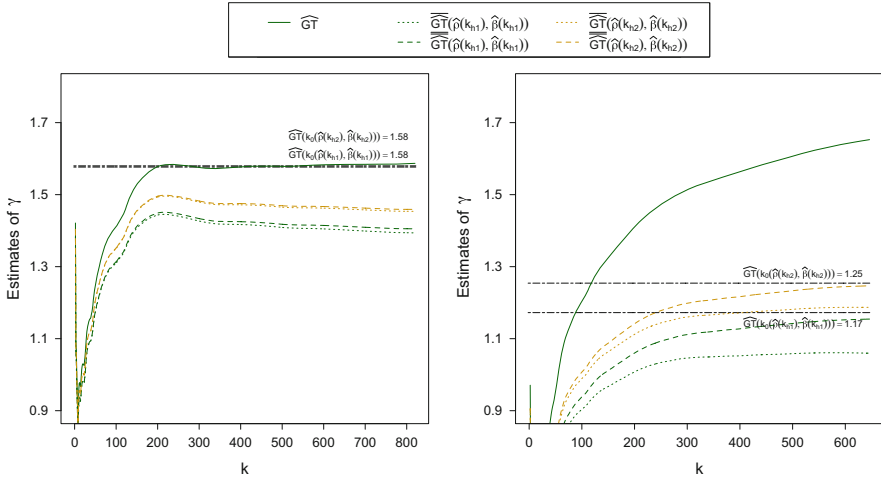
Concerning the shape parameter  $\gamma$ , Fig. 7 displays the estimated values of the GT and Hill estimators, as a function of  $k$ , for Philippines and Vanuatu Islands data. As can be observed, for Philippines Islands data both estimators stabilize around the same value of  $\gamma$ , which is 1.6, with identical scatter plots for moderate and high values of  $k$ , although it is worth to give emphasis to the smoothness that the geometric-type estimator displays.

For the Vanuatu Islands data, though not so explicit as to the Philippines data, the behavior of GT tends to stabilize around the value of 1.64 as  $k$  increases. The same is true for the Hill estimator around the value of 1.78, although in a slightly more erratic way.

The GT estimator presents the best performance specially for Philippines Islands data, displaying almost a straight line around 1.58 for  $k$ -values larger than 300.

In Fig. 8 it is possible to compare the behavior of the GT estimator with its corrected versions,  $\overline{\widehat{GT}}$  and  $\overline{\overline{\widehat{GT}}}$ . We note that the corrected estimators maintain the good behavior; that is, they have less variation in the initial values of  $k$ , and





**Fig. 8** Plot for the GT estimator,  $\widehat{GT}$ , and for the corresponding GT bias corrected estimators,  $\overline{\widehat{GT}}$  and  $\underline{\widehat{GT}}$ , of  $\gamma$ , for seismicity of Philippines (*left*) and Vanuatu (*right*) Islands

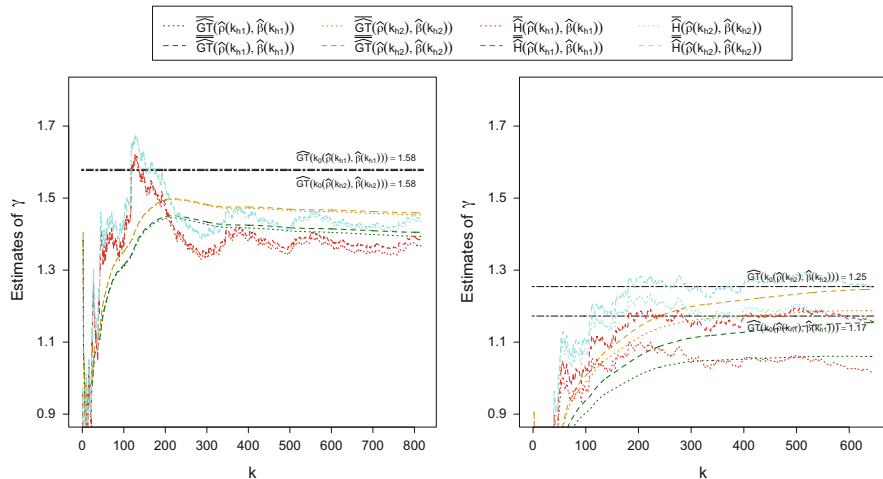
stabilize at slightly lower values than the uncorrected estimator. Depending on the unknown value of the tail index parameter that we seek, this type of behavior seems to be indicative of a better performance of the corrected estimators. Particularly for Vanuatu Islands data, this improvement seems to be evident since the corrected estimators begin to stabilize sooner than the non corrected ones, showing a very satisfactory behavior, to the right from the initial values of  $k$ .

In order to make the comparison between the bias corrected GT estimators and the Hill ones, we draw the sample paths of one against the other.

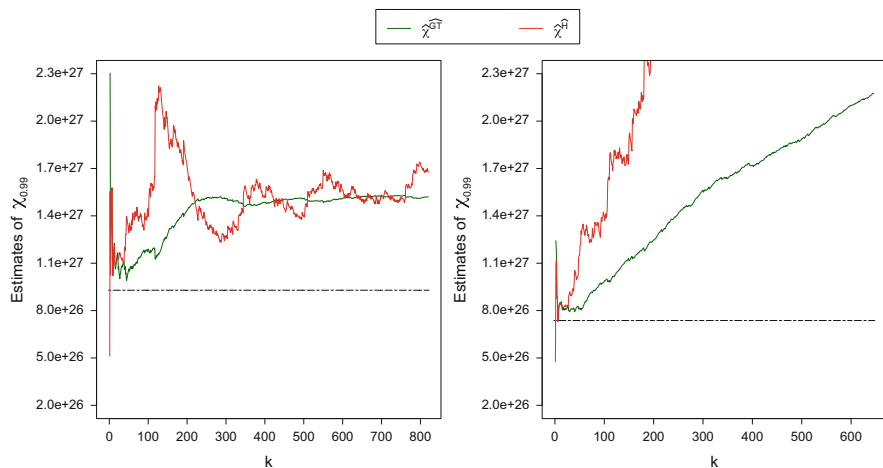
We might see from Fig. 9 that the estimates provided by the corrected Hill estimators are around the same values of the estimates given by the corrected GT estimators. However, it is quite clear that the Hill estimators hold a rather irregular behavior compared to the GT estimators, especially for smaller values of  $k$ .

It is suggestive that the value of  $\gamma$  that best describes the seismic moment of the Philippines Islands is a little below 1.5, and that of the Vanuatu Islands is slightly above 1.

As in most of the applications, the main interest lays not on the tail index but in the quantiles of the extreme distributions, which are more stable and robust. Now we analyze the sample paths of the quantiles estimators. We estimate the values of POT high quantiles estimator, in (5), based on the GT and Hill estimators, as a function of  $k$ , for Philippines and Vanuatu Islands data, considering the percentile 99%. Each tail index estimator leads to a different estimation of large quantiles, which is also dependent on  $k$ . The straight dashed line represents the estimate of the empirical 99% quantile. When more than one straight line is present, the empirical quantile is represented by the inferior one.

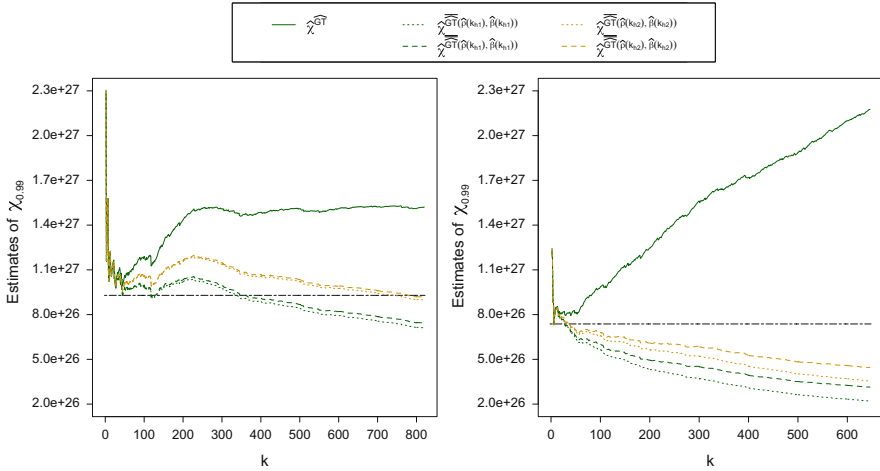


**Fig. 9** Plot for the GT bias corrected estimators,  $\overline{\overline{GT}}$  and  $\overline{\overline{GT}}$ , and for the Hill ones,  $\overline{\overline{H}}$  and  $\overline{\overline{H}}$ , of  $\gamma$ , for seismicity of Philippines (*left*) and Vanuatu (*right*) Islands



**Fig. 10** Plot for the 99-quantile estimators based on the GT estimator,  $\widehat{\chi}^{GT}$ , and on the Hill estimator,  $\widehat{\chi}^H$ , of  $\chi_{0.99}$ , for seismicity of Philippines (*left*) and Vanuatu (*right*) Islands (empirical quantiles  $\chi_{0.99} = 9.29 \times 10^{26}$  and  $\chi_{0.99} = 7.37 \times 10^{26}$ , for Philippines and Vanuatu Islands, respectively)

We might see from Fig. 10 that, for the Philippines Islands, both estimates do not present values close to the empirical quantile. For values of  $k$  larger than 300, the estimates tend to stabilize, and it is apparent that this stabilization process is significantly more regular for the GT based quantiles estimator. The uneven performance that the Hill quantile plot shows make it extremely hard to decide upon

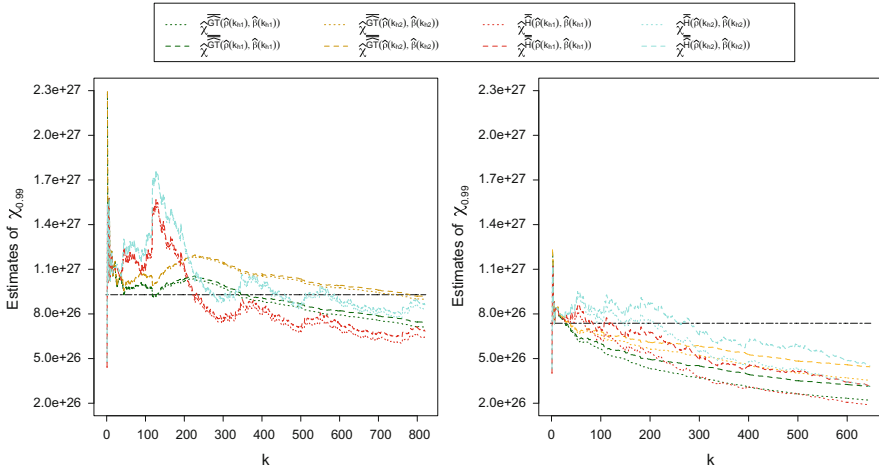


**Fig. 11** Plot for the 99-quantile estimators based on the GT estimator,  $\hat{\chi}^{GT}$ , and on the corresponding geometric-type bias corrected estimators,  $\hat{\chi}^{GT}$  and  $\hat{\chi}^{GT}$ , of  $\chi_{0.99}$ , for seismicity of Philippines (*left*) and Vanuatu (*right*) Islands (empirical quantiles  $\chi_{0.99} = 9.29 \times 10^{26}$  and  $\chi_{0.99} = 7.37 \times 10^{26}$ , for Philippines and Vanuatu Islands, respectively)

a specific value for  $k$ . For the Vanuatu Islands the behavior of both estimators is not the best, but the Hill based quantiles estimator presents a much more irregular behavior.

Now comparing the GT based quantiles estimator with its corrected versions, we can observe in Fig. 11 that the improvement caused by the correction is quite remarkable. It is also worth noting that considering the  $k_{h2}$  level to estimate the second order parameters, the performance seems to be a little better. Also in Fig. 11, and for the Philippines Islands data, it can be seen that the quantile value calculated using the geometric-type estimator at its optimal levels  $k_0^{GT}$ , represented by the superior straight lines, almost coincides with the value of the quantiles estimator based on the geometric-type estimation for  $k$ -values larger than 200, which highlights the fairly stable behavior of this quantiles estimator in this range of values.

In Fig. 12, we can observe that the bias corrected Hill quantiles estimators present estimate values very similar to the ones presented by the bias corrected GT quantiles estimators. Although the corrected Hill quantiles estimators, using the  $k_{h2}$  level to compute the second order parameters, appear to have values more close to the empirical quantile than the corresponding corrected GT quantiles estimators, in case of Philippines Islands only for  $k$ -values greater that 300, their erratic and much less stable behavior may be a factor of considerable disadvantage.



**Fig. 12** Plot for the 99-quantile estimators based on the geometric-type bias corrected estimators,  $\hat{\chi}_{GT}^{\bar{H}}$  and  $\hat{\chi}_{GT}^{\bar{H}}$ , and on the Hill bias corrected estimators,  $\hat{\chi}^{\bar{H}}$  and  $\hat{\chi}^{\bar{H}}$ , of  $\chi_{0.99}$ , for seismicity of Philippines (*left*) and Vanuatu (*right*) Islands (empirical quantiles  $\chi_{0.99} = 9.29 \times 10^{26}$  and  $\chi_{0.99} = 7.37 \times 10^{26}$ , for Philippines and Vanuatu Islands, respectively)

### 4 Final Considerations

In this study we consider the seismic moments of the Philippines and Vanuatu Islands larger than the level  $10^{24}$  recorded during 35 years. We begin by analyzing the data in order to investigate the presence of heavy tails, the stationarity and the independence of the observations. In this way, we verify that the exceedances can be modeled by heavy tailed distributions. We use the geometric-type estimator and its bias corrected versions for estimating the tail index and high quantiles. For the sake of comparison we also consider the corresponding Hill estimators.

The geometric-type estimator shows a better performance when compared to the Hill estimator, namely it is worth emphasizing the contrast between the smoothed behavior of the geometric-type estimator and the irregular behavior exhibited by the Hill estimator.

It is well known that the considerable bias that appears in several estimators reveals a difficult problem that goes well beyond the application. In order to deal with this problem we also study and apply corrected versions of the geometric-type estimator. As expected, its performance is improved. We may emphasize that in some situations the Hill’s bias corrected estimators present an erratic and less stable behavior. This is a real disadvantage for example in choosing a specific value for  $k$ .

In general, it is possible to conclude that the smoother behavior is a common quality shared by the estimates obtained for the GT tail index estimators, as by GT-based quantiles estimates, which show a very small variability, reflecting the more regular behavior of the GT estimators.

Regarding the case of Philippines Islands, and when considering the geometric-type estimator, we obtain an estimate for the seismic moment 0.99-quantile of  $1.51 \times 10^{27}$ . In a more practical way, we may say that it is expected that one out of a hundred earthquakes has a seismic moment larger than  $1.51 \times 10^{27}$ . Since in average there are 23.43 earthquakes per year, we may say that an earthquake exceeding a seismic moment of  $1.51 \times 10^{27}$  is expected to happen in Philippines Islands once in every 4.35 years. Moreover, we may also conclude that the probability of occurring an earthquake with seismic moment larger than  $1.51 \times 10^{27}$  next year is approximately  $1 - 0.99^{23.43}$ , that is, 21%.

As one knows, the performance of the estimators depends on the distribution of the data, and there is not an uniformly agreed best estimator. Nevertheless, from results of practical example conducted here, one could say that, for this type of data, the GT estimator turns out to be the best choice for tail index estimator, and the POT estimator when used for high quantiles.

On the whole, the application of the EVT to the problem under study seems quite promising since it provides reasonable estimates of the tails of the seismic moment distribution.

**Acknowledgements** ACMF is partially supported by FCT grant SFRH/BPD/66174/2009 and LC is supported by FCT grant SFRH/BD/60642/2009. All three authors are supported by FCT project PTDC/MAT/120346/2010. Research funded by the European Regional Development Fund through the programme COMPETE and by the Portuguese Government through the FCT—Fundação para a Ciência e a Tecnologia under the project PEst—C/MAT/UI0144/2013. The authors also thank the referees for their comments.

## References

1. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.: *Statistics of Extremes: Theory and Applications*. Wiley, Chichester (2004)
2. Brito, M., Freitas, A.C.M.: Limiting behaviour of a geometric estimator for tail indices. *Insur. Math. Econ.* **33**, 221–226 (2003)
3. Brito, M., Cavalcante, L., Freitas, A.C.M.: Bias corrected geometric-type estimators. Preprint CMUP 2014-6 (2014)
4. Caeiro, F., Gomes, M.I., Pestana, D.: Direct reduction of bias of the classical Hill estimator. *Revstat* **3**, 113–136 (2005)
5. Corral, A.: Dependence of earthquake recurrence times and independence of magnitudes on seismicity history. *Tectonophysics* **424**, 177–193 (2006)
6. Csörgő, S., Deheuvels, P., Mason, D.M.: Kernel estimates of the tail index of a distribution. *Ann. Stat.* **13**, 1050–1077 (1985)
7. Day, R.W.: *Geotechnical Earthquake Engineering* McGraw-Hill, New York (2002)
8. de Haan, L., Rootzén, H.: On the estimation of high quantiles. *J. Stat. Plann. Inference* **35**, 1–13 (1993)
9. Deheuvels, P., Haeusler, E., Mason, D.M.: Almost sure convergence of the Hill estimator. *Math. Proc. Camb. Philos. Soc.* **104**, 371–381 (1988)
10. Dekkers, A.L.M., Einmahl, J.H.J., de Haan, L.: A moment estimator for the index of an extreme-value distribution. *Ann. Stat.* **17**, 1833–1855 (1989)

11. Dziewonski, A.M., Chou, T.-A., Woodhouse, J.H.: Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *J. Geophys. Res.* **86**, 2825–2852 (1981)
12. Ekström, G., Nettles, M., Dziewonski, A.M.: The global CMT project 2004–2010: centroid-moment tensors for 13,017 earthquakes. *Phys. Earth Planet. Inter.* **200–201**, 1–9 (2012)
13. Fraga Alves, M.I., Gomes, M.I., de Haan, L.: A new class of semi-parametric estimators of the second order parameter. *Port. Math.* **60**, 193–213 (2003)
14. Global CMT Catalogue: Available from <http://www.globalcmt.org/> (2013). Last accessed Aug 2013
15. Gomes, M.I., Martins, M.J.: “Asymptotically unbiased” estimators of the tail index based on external estimation of the second order parameter. *Extremes* **5**, 5–31 (2002)
16. Gomes, M.I., Martins, M.J., Neves, M.: Improving second order reduced bias extreme value index estimation. *Revstat* **5**, 177–207 (2007)
17. Haeusler, E., Teugels, J.L.: On asymptotic normality of Hill’s estimator for the exponent of regular variation. *Ann. Stat.* **13**, 743–756 (1985)
18. Hill, B.M.: A simple approach to inference about the tail of a distribution. *Ann. Stat.* **3**, 1163–1174 (1975)
19. Howell, B.F. Jr.: *An Introduction to Seismological Research*. Cambridge University Press, Cambridge (1990)
20. Pickands, J.: Statistical inference using extreme order statistics. *Ann. Stat.* **3**, 119–13 (1975)
21. Pisarenko, V.F., Sornette, D.: Characterization of the frequency of extreme events by Generalised Pareto Distribution. *Pure Appl. Geophys.* **160**, 2343–2364 (2003)
22. Pisarenko, V.F., Sornette, D., Rodkin, M.V.: Distribution of maximum earthquake magnitudes in future time intervals, application to the seismicity of Japan (1923–2007). *Earth Planets Space* **62**, 567–578 (2010)
23. Weissman, I.: Estimation of parameters and large quantiles based on the  $k$  largest observations. *J. Am. Stat.* **73**, 812–815 (1978)