Modelling 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 carried out and their application to the considered problem is illustrated and discussed in the corresponding context.

1 Introduction

Earthquakes are present in everyday life of humanity worldwide. A severe earthquake is one of the most frightening and destructive phenomena of nature. Experiencing an earthquake is certainly the worst experiences anyone can have. The lived moments are reported as full of panic, terror and death.

Margarida Brito
Faculdade de Ciências, Universidade do Porto, e-mail: mabrito@fc.up.pt

Laura Cavalcante
Faculdade de Ciências, Universidade do Porto, e-mail: laucavalcante@fc.up.pt

Ana Cristina Moreira Freitas
Faculdade de Economia, Universidade do Porto, e-mail: amoreira@fep.up.pt

For survivors, the terrible images remain in memory and become part of their daily lives, as well as the constant fear within each based on the possibility of a next big earthquake that can take lives and separate families forever. It is estimated that there are about one million earthquakes per year, however, the vast majority occur in the mid 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 and some deaths. On average, only one catastrophic earthquake occurs per year and a highly catastrophic every 5 years.

Since the phenomena that trigger it is still under study and that there are uncontrollable forces of nature that dominate them, they are actually considered unpredictable and mankind will have to learn to live with them. Thus, the characterisation of the seismic activity is fundamental in order to reduce the number of deaths and economic losses. This constitutes an important challenge requiring a large multidisciplinary effort. In this work we will follow a stochastic approach, taking into account specific features of big earthquakes. When we are dealing with extreme events, the classical statistical models are inappropriate for the statistical modelling of earthquake size. In standard data analysis, unusual observed values are often considered outliers and ignored in the fitting of a statistical model. In this context, the main interest of the study relies on the analysis of the tail of the distribution that fits 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. Beirlant et al. (2004), 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 occurrence of a large earthquake over a long period of time or high quantiles (see e.g. Pisarenko et al. (2010)).

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 behaviour of the distributions of large earthquakes seismic moments is characterised using techniques from EVT. To apply these methods a preliminary data analysis is performed to investigate the validity of the underlying usual 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 carefully discussed.

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 that is addressed here.

The paper is organised as follows. Some important concepts and results about EVT and earthquakes are briefly presented in Section 2. The investigation in order to verify the validity of the usual assumptions and the analysis of the seismic moments are performed in Section 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 Section 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 behaviour of a distribution, since it encompasses a set of probabilistic results that allow characterizing and modelling the extreme values behaviour. In this way, the EVT is very useful to make 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 the 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 $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables (r.v.) with distribution function (d.f.) F and $M_n = \max(X_1, X_2, \ldots, X_n)$ denote the maximum of the n observations. If a sequence of real numbers $a_n > 0$ and b_n exists such that*

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

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

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

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

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

for all continuity points of G.

In the conditions of the theorem is said that F belongs to the domain of attraction of G $\big(F \in DA(G)\big)$.

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 γ 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 $X_1, X_2, \ldots, X_n$ 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,*

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

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

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

where γ, 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 normalised maximum is the GEV distribution and that the limit of any excess function is the GPD. Hence, they are fundamental to make possible the 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 size blocks with a previous determined amplitude and the maximum observation of each block is collected; the interest lies in the asymptotic study of maxima. In the Peaks Over Threshold (POT) method one selects the observations that exceed a certain high threshold; the interest lies in the asymptotic behaviour of the excesses over a high threshold.

Accordingly with the data set under study, one must deal with these approaches being aware that both methods have disadvantages. One major drawback of the BM method is that only one observation in a block is used to make an inference about the limiting distribution of the maximum, resulting in a small final sample size. On other hand, this method is more robust in respect to the eventual dependence among 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 modelling the extreme value distribution, the main issue to be solved is the parameter estimation. The shape parameter γ is of great interest in the analysis of the tails, since it dominates the behaviour of extremes. This parameter indicates the heaviness of the tail, since the tail function becomes more heavy as γ increases. 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 $X_1, X_2, \ldots, X_n$ is a sample of i.i.d. r.v. with d.f. F and denote by $X_{(1,n)} \leq X_{(2,n)} \leq \cdots \leq X_{(n,n)}$ the corresponding order statistics (o.s.). The estimation of γ 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,

$$k \to \infty, \qquad \frac{k}{n} \to 0 \qquad \text{as} \qquad n \to \infty. \tag{1}$$

Several estimators have been proposed for the estimation of γ (see e.g. Hill (1975), Pickands (1975), Csörgő et al. (1985), Dekkers et al. (1989)). Here we consider the following estimator for $\gamma > 0$, the geometric-type (GT) estimator

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

where

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

We also consider the commonly used Hill estimator (see Hill (1975)) 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. Brito and Freitas (2003), Deheuvels et al. (1988) and Haeusler and Teugels (1985)).

The problem of estimating high order quantiles has received increased attention as a useful tool in data modelling and it has been utilized 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 Weissman (1978),

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

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

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

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

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

In the present work both the $\widehat{GT}(k)$ and $\widehat{H}(k)$ are used to estimate γ. The high quantiles are estimated considering (5) and using $\widehat{GT}(k)$ and $\widehat{H}(k)$ as estimators of γ. The asymptotic behaviour of these quantile estimators was studied and their asymptotic normality was proved (cf. Brito et al. (2014), Dekkers et al. (1989) and de Haan and Rootzén (1993)).

The problem of reducing the bias of these tail index estimators was addressed in Brito et al. (2014), where were proposed the following two asymptotic equivalent geometric-type bias corrected estimators

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

and

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

Hill bias corrected estimators may be found in Caeiro et al. (2005), namely

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

and

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

where ρ and β 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 this influence can be considerable, the appropriate estimation of γ is fundamental in obtaining a suitable quantile estimator with a good performance.

2.2 Earthquakes

In general, everything in nature tends to the equilibrium. Due to the thermodynamic equilibrium, the constituents of the Earth's interior are in constant motion. Boosted by this movement that 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 causing 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.

So, the earthquakes are natural shocks, in which the ground quake strongly in the matter of seconds to minutes, that occur as a result of this sudden release of a huge amount of that energy 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, such as volcanoes, represent the more energetic and rapid manifestations of the planet's internal dynamics.

The scientific analysis of earthquakes requires measurement. The size of an earthquake can be 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 normalised for their variation with distance from the earthquake epicenter. However, the known magnitudes present a saturation point which does not allow a correct estimation of the true earthquake size of larger earthquakes, underestimating it. Moreover, it turns out that larger earthquakes, which have larger rupture surfaces, systematically radiate more long-period energy. Nowadays, the measure that is preferably adopted for scientific studies is the *seismic moment* of the displaced ground (see e.g. Howell (1990) and Day (2002)). This measure avoids the saturation problem, since it does not have an intrinsic upper bound, and describes the size of an earthquake as a essential combination of physical quantities that really matters at the earthquake source and that determines how strong the seismic motions will be.

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, μ, 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

$$M = \mu A D.$$

Because many people do not really know what means a number with the "size" of seismic moment and since the magnitude scale has been used for a very long time, the need to convert it into some kind of magnitude scale emerged. These factors have led to the definition of a new magnitude scale, the moment magnitude, m_w, based on seismic moment

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

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 modelling of earthquake data

In this section we concentrate on the entire route one need to travel from the raising of the data to our ultimate goal of modeling the tail of the distribution of earthquakes seismic moments. In order to do this we begin by explaining the entire procedure which was necessary to make it possible to apply the POT approach to the data considered in the study. More specifically, we first describe the data and perform a preliminary exploratory data analysis in which we discuss the type of distribution to which the data belongs, as well as we investigate the stationarity and independence of the data, and 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 Global Centroid-Moment-Tensor (CMT) web page (cf. e.g. Dziewonski et al. (1981) and Ekström et al. (2012)). Here, we restrict the territory of our study to earthquakes occurring within the Philippines and Vanuatu Islands, and the analysis was performed in a similar way for the both regions. In particular, we extract and analyse the information about 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 Pisarenko and Sornette (2003). The observations under this threshold were removed. Since we detect a failure in data acquisition of the Vanuatu Islands until 01-01-1980, we just consider the Vanuatu Islands data subsequent to this date. So the final data sets, on which the analysis that follows has been based, consider 821 cases for Philippines Islands and 647 cases for Vanuatu Islands. We did not exclude the aftershocks because apart from owning a greatly reduced fraction on the range of seismic moments considered, their removal may introduce a bias in the parameters estimation (cf. e.g. Pisarenko and Sornette (2003)). As the considered region has a lot of deep earthquakes, they also were not excluded. Thus, after the space, time and seismic moment has been selected, no further elimination of events is performed. In Fig. 1 the seismic moments of Philippines and Vanuatu Islands over the above mentioned period are plotted.

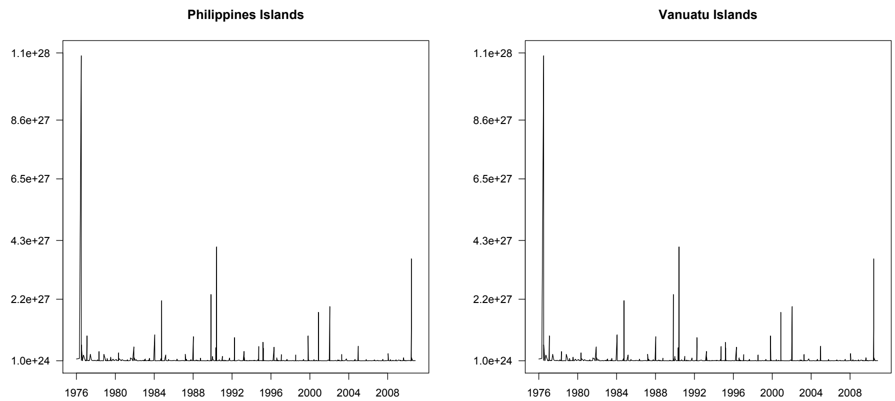

Fig. 1 Seismic moments of Philippines (left) and Vanuatu (right) Islands.

3.2 Preliminary data analysis

Before proceeding it would be useful to discuss if the Pareto-type model provide a plausible fit to the seismic moment 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, using for it their quantiles.

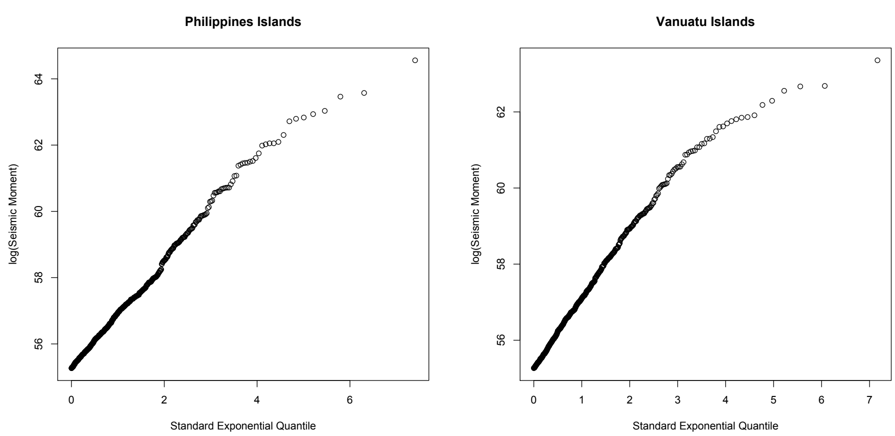

Fig. 2 Pareto QQ plot for Philippines (left) and Vanuatu (right) Islands seismic moment data.

Usually, as in this case, the most convenient comparison is between the empirical quantiles of the sample and the quantiles of the theoretical distribution intended. If the sample data and the reference distribution are derived from populations with a common distribution, the QQ plot should show a strong linear trend. So, the type of the distribution can be derived by looking for the QQ plot.

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

Given that $Y \stackrel{D}{=} \log X$, 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 resultant scatterplot a linear pattern is evident, which is indicative of the good agreement between observed values and the values predicted by the model. We carefully analyse the behaviour of the QQ plot on its upper right part, which represents the most extreme values and, although slightly less than in the remaining part of the plot, a linear tail behaviour is made apparent. The visual impressions based on the Pareto QQ plots suggests that the Vanuatu and Philippines Islands earthquake data sets do seem to follow a Pareto distribution, ie, we are dealing with a heavy-tailed underlying distribution ($\gamma > 0$).

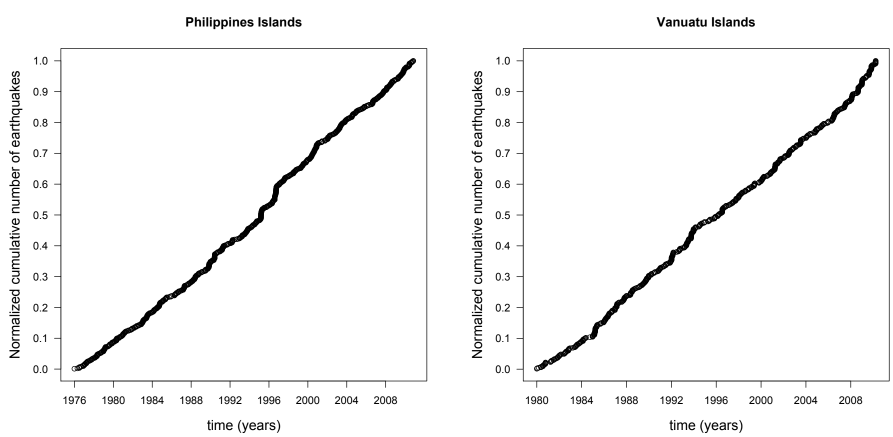

Fig. 3 Cumulative number of earthquakes normalised by the total number in the period considered as a function of time, for seismicity of Philippines (left) and Vanuatu (right) Islands with $M \geq 10^{24}$.

We investigate the stationarity of the data under study. Here we refer to strict stationarity, that is, the underlying joint distribution does not depend on time. This is a very convenient property, in particular the statistical parameters do not change over time. To analyse the stationarity we plot the normalised cumulative number of earthquakes versus time.

The linear behaviour that we can observe in Fig. 3 is an indication of the stationary behaviour of the two data sets over the selected time window, thus the data is approximately homogeneous in time and assumed as stationary.

Another relevant property that we are interested to verify before proceeding with the extreme value analysis of the data is the independence, since most of the results in EVT require it as assumption.

In our case, the goal is to analyse the existence of dependence between consecutive seismic moments, ie, verify how the seismic moment of one event, M_{i-1}, influences the seismic moment of the next, M_i.

Here we investigate this statistical dependence through the conditional probability density determined by

$$\frac{P\left(\eta \le M_i < \eta + \Delta_\eta \mid M_{i-1} \ge M_c'\right)}{\Delta_\eta},$$

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 $M \ge M_c$ is always satisfied (see e.g. Corral (2006)).

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, Δ_η, tending to zero, considering only the cases in which the seismic moment of the immediately previous event is bigger than a threshold M_c'.

The dependences will be given by the distribution described above. If the conditional distribution of M_i given that $M_{i-1} \ge M_c'$ is identical to the unconditional distribution, then the seismic moment M_i is statistically independent of an event $M_{i-1} \ge M_c'$. Note that the case $M_c = M_c'$ gives the unconditioned distribution.

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

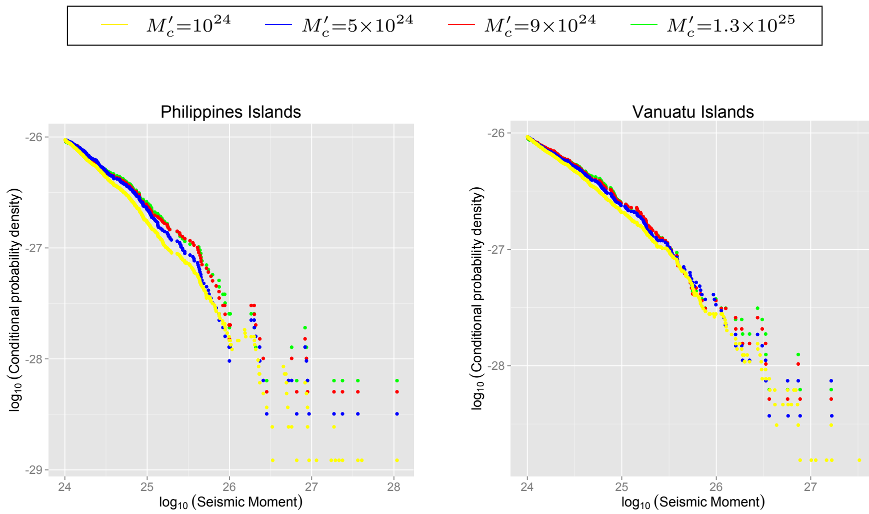

Fig. 4 Conditional probability densities of earthquake seismic moments, for seismicity of Philippines (left) and Vanuatu (right) Islands, evaluated using different thresholds M'_c and with a constant $M_c = 10^{24}$ ($\Delta_\eta = 10^{25}$).

3.3 Estimation of tail parameters

In this section we formalise our main objective of investigating the extremal behaviour of the 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.

We can easily note that the bias dominant components of the bias corrected estimators presented are dependent of the shape ρ and scale β second order parameters. To illustrate the behaviour of the corrected estimators we consider the suitable estimators of the parameter ρ proposed by Fraga Alves et al. (2003)

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

where

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

with M_n^j as in (3), and the β estimator obtained in Gomes and Martins (2002)

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

where

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

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

It is known that the external estimation of ρ and β at a larger k value than the one used for γ-estimation has clear advantages, allowing the bias reduction without increasing the asymptotic variance (see e.g. Caeiro et al. (2005)). In the lines of other studies, and among some suggestions (see e.g. Gomes et al. (2007)), the level that seemed to be the most appropriate to consider in illustrations is

$$k_h=\left\lfloor n^{1-\epsilon}\right\rfloor,\ \text{for some } \epsilon>0 \text{ small}, \tag{9}$$

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

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

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

$$k_{h1}=\left\lfloor n^{0.995}\right\rfloor \qquad \text{and} \qquad k_{h2}=\left\lfloor n^{0.999}\right\rfloor. \tag{10}$$

As usual, the means whereby we do this choice passes by portraying the sample paths of $\widehat{\rho}_\tau(k)$ in (7) for the values $\tau\in\{0,0.5,1\}$, as functions of k, in order to analyse the variations that it causes in their behaviour, and use the following algorithm as a stability criterion for large values of k:

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

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

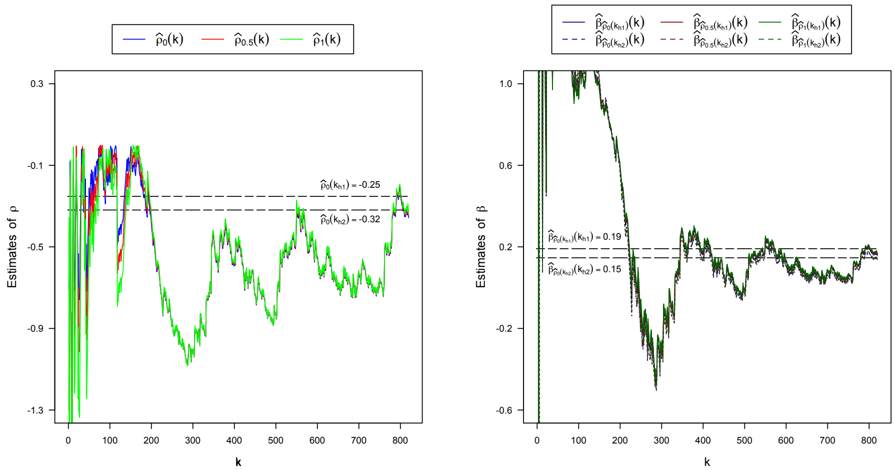

Fig. 5 Estimates of the second order parameters ρ (left) and β (right) for seismicity of Philippines Islands.

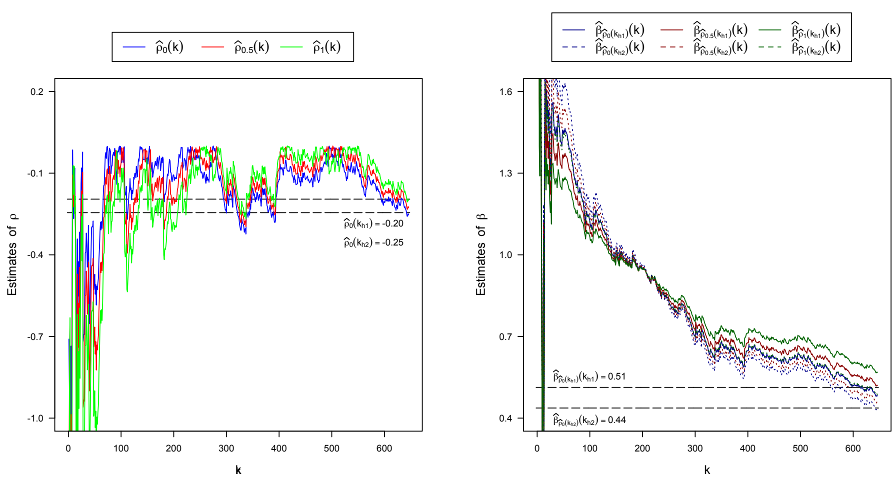

Fig. 6 Estimates of the second order parameters ρ (left) and β (right) for seismicity of Vanuatu Islands.

We might see that the sample paths of $\widehat{\rho}$ for the three different values of τ have a very similar behaviour. It is however apparent that the behaviour of $\widehat{\rho}$ is slightly better when considering $\tau = 0$, specially for data concerning the Vanuatu Islands. As the above described algorithm also points to the choice of $\tau = 0$ in both cases, we chose this value of τ to estimate ρ.

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

Since from the $\widehat{\beta}$ sample paths it is not 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 ρ fluctuations than the β ones, we use the both levels in the remaining 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 γ, through the minimisation of the asymptotic mean square error ($AMSE$) of the geometric-type estimator. Considering the following distributional representation of the geometric-type estimator (see Brito et al. (2014), 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 minimisation

$$\frac{\partial}{\partial k}\left[AMSE\left(\widehat{GT}\right)\right] = 0 \Longleftrightarrow \frac{\partial}{\partial k}\left[V\left(\widehat{GT}\right) + \left(Bias\left(\widehat{GT}\right)\right)^2\right] = 0$$

$$\Longleftrightarrow \frac{\partial}{\partial k}\left[\frac{2\gamma^2}{k} + \left(\frac{\gamma\beta}{(1-\rho)^2}\right)^2 \left(\frac{n}{k}\right)^{2\rho}\right] = 0.$$

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.

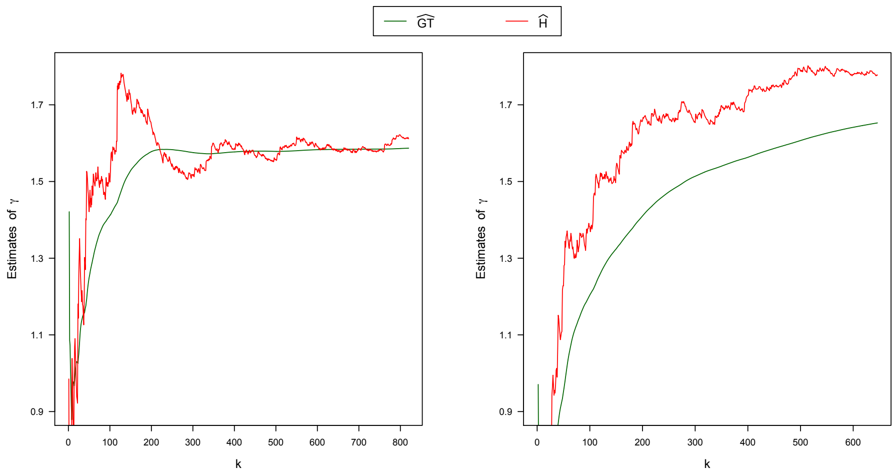

Fig. 7 Plot for the GT estimator, $\widehat{GT}$, and for the Hill estimator, $\widehat{H}$, of γ, for seismicity of Philippines (left) and Vanuatu (right) Islands.

As a first step we estimate the tail index, γ, using GT estimator and Hill's estimator.

Concerning the shape parameter γ, the Fig. 7 displays the estimated values of GT and Hill estimators, as a function of k, for Philippines and Vanuatu Islands data. As one can observe, for Philippines Islands data both estimators give similar results stabilising around the same value of γ, which is 1.6, with basically the same scatter for moderate and high values of k, although it is worth to give emphasis to the smoothness that the geometric-type estimator shows.

For the Vanuatu Islands data, though not so explicit as to the Philippines data, the behaviour of GT tends to stabilise 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 behaviour 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 behaviour, having less variation in the initial values of k, and stabilising at slightly lower values than the uncorrected estimator.

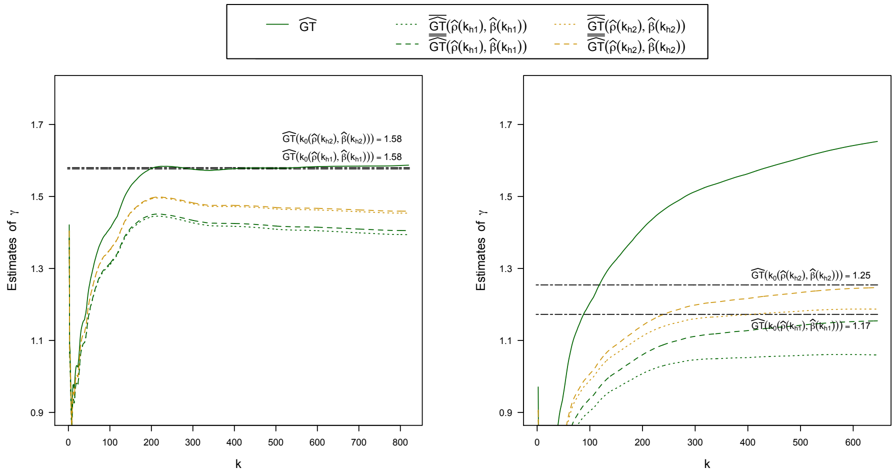

Fig. 8 Plot for the GT estimator, $\widehat{GT}$, and for the corresponding GT bias corrected estimators, $\overline{\widehat{GT}}$ and $\overline{\overline{\widehat{GT}}}$, of γ, for seismicity of Philippines (left) and Vanuatu (right) Islands.

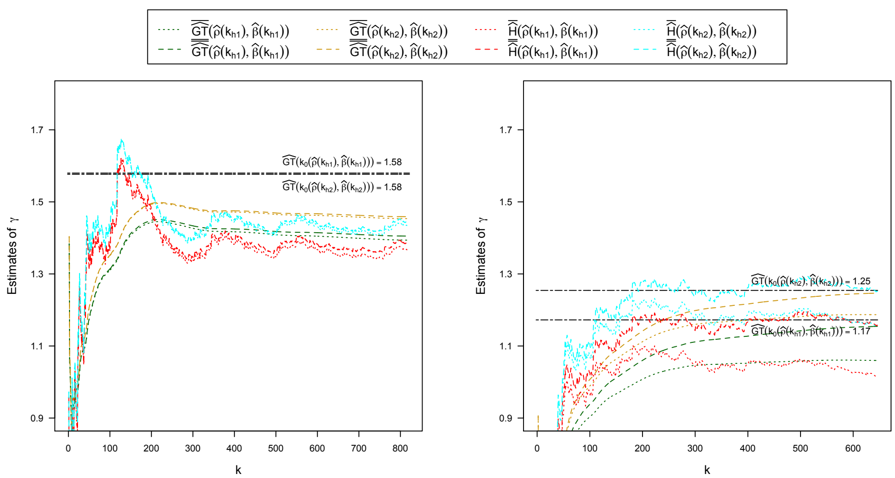

Fig. 9 Plot for the GT bias corrected estimators, $\overline{\widehat{GT}}$ and $\overline{\overline{\widehat{GT}}}$, and for the Hill ones, $\overline{\widehat{H}}$ and $\overline{\overline{\widehat{H}}}$, of γ, for seismicity of Philippines (left) and Vanuatu (right) Islands.

Depending on the unknown value of the tail index parameter, that we seek, this type of behaviour 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 stabilise sooner than the non corrected ones, showing a very satisfactory behaviour, 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 behaviour compared to the GT estimators, specially for smaller values of k.

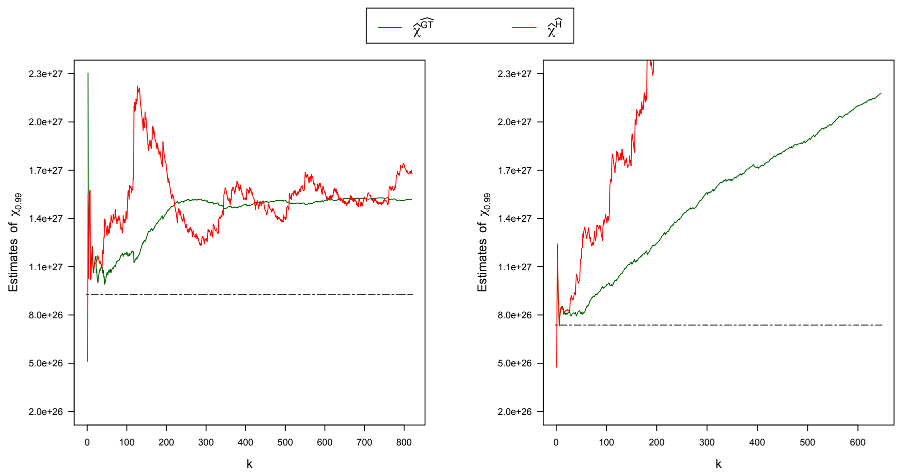

Fig. 10 Plot for the 99-quantile estimators based on the GT estimator, $\widehat{\chi}^{\widehat{GT}}$, and on the Hill estimator, $\widehat{\chi}^{\widehat{H}}$, of $\chi_{0.99}$, for seismicity of Philippines (left) and Vanuatu (right) Islands (empirical quantiles $\chi_{0.99} = 9.29 \times 10^{26}$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

It is suggestive that the value of γ that best describes the seismic moment of the Philippines Islands is a little below 1.5 and 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 analyse 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 are present, the empirical quantile is represented by the inferior one.

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 stabilise, being apparent that this stabilisation 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 a specific value for k. For the Vanuatu Islands the behaviour of both estimators is not the best, but the Hill based quantiles estimator presents a much more irregular behaviour.

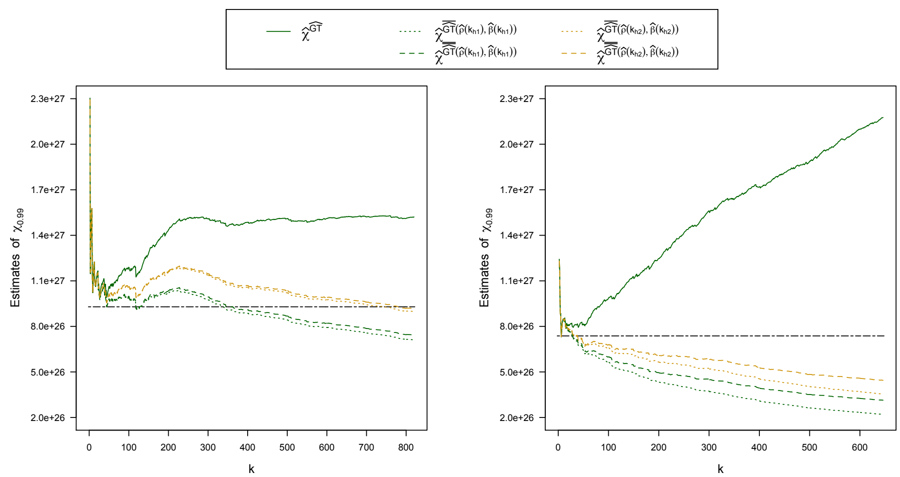

Fig. 11 Plot for the 99-quantile estimators based on the GT estimator, $\widehat{\chi}^{\widehat{GT}}$, and on the corresponding geometric-type bias corrected estimators, $\widehat{\chi}^{\overline{\widehat{GT}}}$ and $\widehat{\chi}^{\overline{\overline{\widehat{GT}}}}$, of $\chi_{0.99}$, for seismicity 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).

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 to be noted 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 Philippines Islands data, it can be seen that the quantile value calculated using the geometric-type estimator at its optimal levels $k_0^{\widehat{GT}}$, 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 behaviour 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 seem 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 behaviour may be a factor of considerable disadvantage.

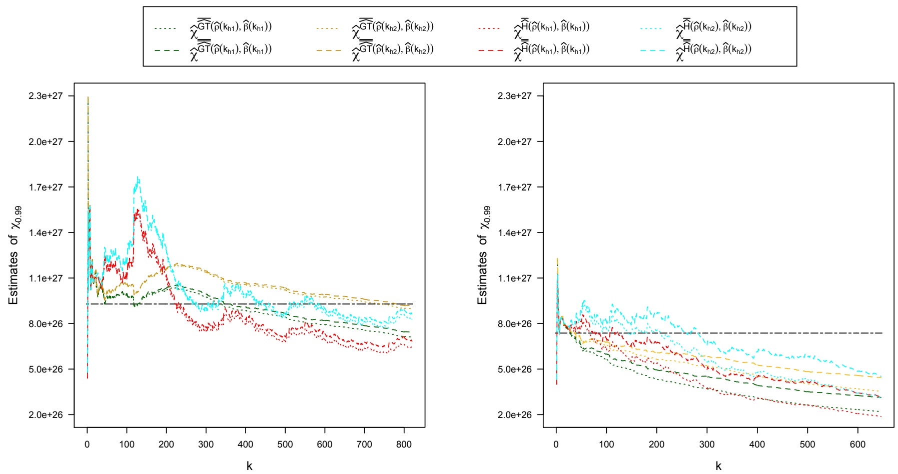

Fig. 12 Plot for the 99-quantile estimators based on the geometric-type bias corrected estimators, $\widehat{\chi}^{\overline{\widehat{GT}}}$ and $\widehat{\chi}^{\overline{\overline{\widehat{GT}}}}$, and on the Hill bias corrected estimators, $\widehat{\chi}^{\overline{\widehat{H}}}$ and $\widehat{\chi}^{\overline{\overline{\widehat{H}}}}$, 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 analysing 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 to emphasise 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 to go beyond the applications. 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 emphasise that in some situations the Hill bias corrected estimators present an erratic and less stable behaviour. This is a real disadvantage for example in choosing a specific value for k.

In general, it is possible to conclude that the smoother behaviour 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 behaviour 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 1.51×10^{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 1.51×10^{27}. Since, in average, there are 23.43 earthquakes per year, we may say that an earthquake exceeding a seismic moment of 1.51×10^{27} is expected to happen in Philippines Islands once in every 4.35 years. Moreover, we also may conclude that the probability of occurring an earthquake with seismic moment larger than 1.51×10^{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 a uniformly best estimator. Nevertheless, from the results of the 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 when used in the POT estimator 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

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*, John Wiley and Sons, Ltd., England.

Brito, M., Cavalcante, L. and Freitas, A.C.M. (2014). Bias corrected geometric-type estimators, *Preprint CMUP 2014-6*.

Brito, M. and Freitas, A.C.M. (2003). Limiting behaviour of a geometric estimator for tail indices, *Insur. Math. Econ.*, **33**, 221–226.

Caeiro, F., Gomes, M.I. and Pestana, D. (2005). Direct reduction of bias of the classical Hill estimator, *Revstat*, **3**, 113–136.

Corral, A. (2006). Dependence of earthquake recurrence times and independence of magnitudes on seismicity history, *Tectonophysics*, **424**, 177–193.

Csörgő, S., Deheuvels, P. and Mason, D.M. (1985). Kernel estimates of the tail index of a distribution, *Ann. Statist.*, **13**, 1050–1077.

Day, R.W. (2002). *Geotechnical Earthquake Engineering*, McGraw-Hill Professional, New york.

de Haan, L. and Rootzén, H. (1993). On the estimation of high quantiles, *J. Statist. Plann. Inference*, **35**, 1–13.

Deheuvels, P., Haeusler, E. and Mason, D.M. (1988). Almost sure convergence of the Hill estimator, *Math. Proc. Camb. Phil. Soc.*, **104**, 371–381.

Dekkers, A.L.M., Einmahl, J.H.J. and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution, *Ann. Stat.*, **17**, 1833–1855.

Dziewonski, A.M., Chou, T.-A. and Woodhouse, J.H. (1981). Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *J. Geophys. Res.*, **86**, 2825–2852.

Ekström, G., Nettles, M. and Dziewonski, A.M. (2012). The global CMT project 2004-2010: Centroid-moment tensors for 13,017 earthquakes, *Phys. Earth Planet. Inter.*, 200–201, 1-9.

Fraga Alves, M.I., Gomes, M.I. and de Haan, L. (2003). A new class of semi-parametric estimators of the second order parameter, *Portugaliae Math.*, **60**, 193–213.

Global CMT Catalogue. Available from <http://www.globalcmt.org/>. [last accessed August 2013].

Gomes, M.I. and Martins, M.J. (2002). “Asymptotically unbiased” estimators of the tail index based on external estimation of the second order parameter, *Extremes*, **5**, 5–31.

Gomes, M.I, Martins, M.J. and Neves, M. (2007). Improving second order reduced bias extreme value index estimation, *Revstat* **5**, 177–207.

Haeusler, E. and Teugels, J.L. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation, *Ann. Statist.* **13**, 743–756.

Hill, B.M. (1975). A simple approach to inference about the tail of a distribution, *Ann. Stat.*, **3**, 1163–1174.

Howell, B.F., Jr (1990). *An Introduction to Seismological Research*, Cambridge University Press.

Pickands, J. (1975). Statistical inference using extreme order statistics, *Ann. Stat.*, **3**, 119–13.

Pisarenko, V.F., Sornette, D. and Rodkin, M.V. (2010). Distribution of maximum earthquake magnitudes in future time intervals, Application to the seismicity of Japan (1923-2007), *Earth Planets Space* **62**, 567–578.

Pisarenko, V.F. and Sornette, D. (2003). Characterization of the frequency of extreme events by Generalised Pareto Distribution, *Pure and Applied Geophysics* **160**, 2343–2364.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations, *J. Amer. Statist.* **73**, 812–815.