Instituto Complutense de Análisis Económico

Assessing the importance of the choice threshold in quantifying market risk under the POT method (EVT)

Sonia Benito Muela
Department of Economic Analysis
Faculty of Economics and Business Administration
National Distance Education University (UNED)

Carmen López-Martín
Department of Business and Accounting
Faculty of Economics and Business Administration
National Distance Education University (UNED).

Mª Ángeles Navarro
PhD. Student of the Faculty of Economics and Business Administration
National Distance Education University (UNED)

Abstract

The conditional extreme value theory has been proven to be one of the most successful in estimating market risk. The implementation of this method in the framework of the Peaks Over Threshold (POT) model requires one to choose a threshold for fitting the generalized Pareto distribution (GPD). In this paper, we investigate whether the selection of the threshold is important for the quantification of market risk. For measuring risk, we use the value at risk (VaR) measure and the expected shortfall (ES) measure. The study has been done for a large set of assets. The results obtained indicate that the quantification of the market risk through the VaR and ES measures does not depend on the threshold selected. This result is also found in a smaller sample.

Keywords Extreme Value Theory, Peaks over Threshold, Value at Risk, Expected Shortfall, Generalized Pareto Distribution.

JEL Classification G19, G29.

Working Paper nº 1820
September, 2018

UNIVERSIDAD COMPLUTENSE MADRID

ISSN: **2341-2356**
WEB DE LA COLECCIÓN: http://www.ucm.es/fundamentos-analisis-economico2/documentos-de-trabajo-del-icae

Assessing the importance of the choice threshold in quantifying market risk under the POT method (EVT)

Sonia Benito Muela[a]. Carmen López-Martín[b]. Mª Ángeles Navarro[c]

The conditional extreme value theory has been proven to be one of the most successful in estimating market risk. The implementation of this method in the framework of the Peaks Over Threshold (POT) model requires one to choose a threshold for fitting the generalized Pareto distribution (GPD). In this paper, we investigate whether the selection of the threshold is important for the quantification of market risk. For measuring risk, we use the value at risk (VaR) measure and the expected shortfall (ES) measure. The study has been done for a large set of assets. The results obtained indicate that the quantification of the market risk through the VaR and ES measures does not depend on the threshold selected. This result is also found in a smaller sample.

Keywords: Extreme Value Theory, Peaks over Threshold, Value at Risk, Expected Shortfall, Generalized Pareto Distribution.

* This work has been funded by the Spanish Ministerio de Ciencia y Tecnología (2016-2019. REF. ECO2015-67305-P) and by the Bank of Spain (PR71/15-20229).

[a] Department of Economic Analysis. Faculty of Economics and Business Administration. National Distance Education University (UNED). Senda del Rey. 11. 28040 Madrid. Spain. E-mail address: soniabm@cee.uned.es

[b] Department of Business and Accounting. Faculty of Economics and Business Administration. National Distance Education University (UNED). Senda del Rey. 11. 28040 Madrid. Spain. E-mail address. E-mail address: carmen.lopez@cee.uned.es

[c] PhD. Student of the Faculty of Economics and Business Administration. National Distance Education University (UNED). Senda del Rey. 11. 28040 Madrid. Spain. E-mail address: mnavarro900@alumno.uned.es

1. Introduction

One of the most important tasks financial institutions face is evaluating their market risk exposure. Traditionally, the market risk of a portfolio was measured through the variance. In fact, traditional financial theory defines risk as the dispersion of the results with respect to the mean return. Another way of measuring risk, which is currently the most commonly used, is to evaluate the losses that may occur when the price of the assets that makes up the portfolio decreases[1]. To evaluate those losses, two measures have been developed: (i) the value at risk (VaR) measure (J.Morgan, 1996) and (ii) the expected shortfall (ES) measure (Acerbi and Tasche, 2002). The VaR of a portfolio is defined as the worst expected loss over a given horizon under normal market conditions at a given level of confidence. Formally speaking, the $VaR(\alpha)$ of a portfolio at $(1-\alpha)\%$ confidence level is the percentile α % of the return portfolio distribution. To date, the VaR measure has been by far the most used by financial institutions and regulators[2].

However, this measure is not exempt from criticism. Certain researchers have remarked that VaR is not a coherent market risk measure as it violates the subadditivity condition, which may discourage diversification[3] (see Artzner et al., 1999). Another weakness of the VaR measure is that it fails to control tail-risk. The ES is defined as the average of all losses that are greater than or equal to VaR, i.e., the average loss in the worst α % cases. In other words, this measure provides the expected value of an investment in the worst α % of the cases. In contrast to the VaR measure, ES is a coherent risk measure, and it does not present tail-risk.

Although, to date, the VaR measure has been the most used for quantifying market risk, in the future, ES will garner more prominence, in part due to the change in the regulation set by the Basel Committee on Banking Supervision (BCBS). Under the new regulation, financial institutions must calculate the market risk capital requirements' risk based on the ES measure, replacing the VaR measure (BCBS, 2012, 2013, 2017).

To estimate those measures, several methodologies have been developed: (i) the parametric approach, (ii) the non-parametric approach (e.g., historical simulation) and (iii) the semi-parametric method (e.g., extreme value theory, filtered historical simulation and CaViar method). Among all these measures, extreme value theory (EVT) has been proven to be one of the most successful in VaR estimation (see Abad et al., 2014)[4].

The extreme value theory approach focuses on limiting the distribution of extreme returns observed over a long time period, which is essentially independent of the distribution of the returns themselves. The two main models for extreme value theory are the block maxima model (McNeil, 1998) and the peaks-over-threshold (POT) model. In the context of the POT model, extreme values

[1] In this case, the concept of risk is associated with the danger of losses.
[2] In 1996, the Basel Committee on Banking Supervision (BCBS) introduced an amendment where financial institutions were required to meet capital requirements based on VaR estimates.
[3] A risk measure ρ is called coherent if it satisfies the following conditions: (i) homogeneous, (ii) subadditive, (iii) monotonic and (iv) translation invariant.
[4] To date, there are few studies dedicated to comparing ES models.

above a high threshold are analysed using a generalized Pareto distribution (GPD). The difficulty of this method lies in finding the optimal threshold for GPD fitting. Threshold choice involves balancing bias and variance. The threshold must be sufficiently high to ensure that asymptotic underlying the GPD approximations is reliable, thus reducing bias. However, the reduced sample size for high thresholds increases the variance of the parameter estimates (see Scarrot and McDonald, 2012).

To determine the optimal threshold, several techniques have been proposed such as graphic methods, ad hoc methods or methods based on goodness-of-fit contrasts. However, none of these techniques have been proven to provide better results than the others.

Although many proposals have been made to determine the optimal threshold in the framework of the POT method, in this paper, we ask whether in the financial field; specifically, in measuring market risk, the choice of the threshold is important. The study by Iriondo (2017) offers preliminary evidence against this hypothesis. In accordance with this author, we analyse the extent to which the selection of the threshold is decisive in quantifying the market risk. To answer this question, we will analyse the impact of the threshold on the two aforementioned risk measures: VaR and ES.

The results of the study indicate that according to the literature, the choice of the threshold affects the parameter estimates of the GPD; however, the risk measures (VaR and ES) obtained from these parameters do not depend on the choice threshold. To answer this question, we analyse in detail the case of the S&P 500 and later extend that study to a set of 14 assets: 7 stock indexes (CAC40, DAX30, FTSE100, HangSeng, IBEX35, Merval and Nikkey), four commodities (Copper, Gold, Crude Oil Brent and Silver) and three rates exchange (£/€, $/€ and ¥/€). This result is also found in a smaller sample.

The remainder of the paper is organized as follows. In section 2, we present the methodology we use for the study. In section 3, we present the data and the results obtained for the particular case of the S&P 500 index. Section 4 displays a robustness analysis. The main conclusions are presented in section 5.

2. Methodology

2.1 Extreme Value Theory

The extreme value theory (EVT) approach focuses on the limiting distribution of extreme returns observed over a long time period, which is essentially independent of the distribution of the returns themselves. The two main models for EVT are (1) the block maxima models (BM) (McNeil, 1998) and (2) the peaks-over-threshold model (POT). The second model is generally considered to be the most useful for practical applications due to the more efficient use of the data at the extreme values. In the framework of the POT model, there are two types of analysis: the Semi-parametric models built around the Hill estimator and its relatives (Beirlant et al., 1996; Danielson et al., 1998) and the full parametric models based on the generalised Pareto distribution (Embrechts et al., 1999). In this paper we focus on the full parametric model.

Given a set of random variables $(r_1, r_2, \ldots, r_n)$, iid $\sim F$, we choose a low threshold u and examine all values (y) exceeding u: $\left(y_1, y_2, \ldots, y_{N_u}\right)$ where $y_i = r_i - u$ and N_u are the number of sample data greater than u. The distribution of excess losses over the threshold u is defined as:

$$F_u(y) = P\big((r-u) < y \,|\, r > u\big) = \frac{F(r+u) - F(u)}{1 - F(u)} \tag{1}$$

According to the theorem of Pickans (1975) and Balkema and de Haan (1974), for a large class of underlying distributions functions F the conditional excess distribution function $F_u(y)$, for a u large , is well approximated by $F_u(y) \approx G_{k.\xi}(\mathrm{y})$ with $u \ \to \infty$, where

$$G_{k.\xi}(\mathrm{y}) = \begin{cases} 1 - \left(1 + \frac{k}{\xi} y\right)^{-1/k} & if \ \ \xi \neq 0 \\ 1 - \exp\left(-\frac{y}{\xi}\right) & if \ \ \xi = 0 \end{cases} \tag{2}$$

$G_{k.\xi}(\mathrm{y})$ is the so-called generalized Pareto distribution (GPD), and k and ξ represent the shape parameter and the scale parameter, respectively. The shape parameter can take any value, positive or negative. The scale parameter is always positive.

Figure 1 illustrates the shape of the generalized Pareto distribution and the corresponding density function when the shape parameter or tail index takes negative and positive values.

[Insert Figure 1]

Assuming that, for a certain u, the distribution of excess losses above the threshold is a generalized Pareto distribution, then the distribution function of returns is given by:

$$F(r) = (1 - F(u))\ F_u(\mathrm{y}) + F(u) \tag{3}$$

and replacing $F_u(\mathrm{y})$ by GPD and $F(u)$ by its empirical estimator $(n - N_u)/n$, where n is the total number of observations and N_u the number of observations above the threshold u, we have

$$F(r) = \frac{N_u}{n}\left(1 - \left(1 + \frac{k}{\xi}(r-u)\right)^{-1/k}\right) + \left(1 - \frac{N_u}{n}\right) \tag{4}$$

which simplifies to

$$F(r) = 1 - \frac{N_u}{n}\left(1 + \frac{k}{\xi}(r-u)\right)^{-\frac{1}{k}} \tag{5}$$

For a given probability $\alpha > F(u)$, the quantile α, which is denoted by $q(\alpha)$, is calculated by inverting the tail estimation formula to obtain

$$q(\alpha) = u - \frac{\xi}{k}\left(\left(\frac{n}{N_u}\alpha\right)^{-k} - 1\right) \tag{6}$$

The expected shortfall associated with the quantile α, which is denoted by $ES(\alpha)$, is given by:

$$ES(\alpha) = q(\alpha) + E[r - q(\alpha) \,|\, r > q(\alpha)] \tag{7}$$

where the second term on the right is the mean of the excess distribution $F_{VaR_\alpha(y)}$ over the threshold $VaR(\alpha)$. It can be demonstrated that the mean of the excess distribution $F_{VaR_\alpha(y)}$ over the threshold $VaR(\alpha)$ is given by:

$$E[r - q(\alpha)|\, r > q(\alpha)] = \frac{\xi + k(q(\alpha) - u)}{1-k} \tag{8}$$

and therefore, we obtain

$$ES(\alpha) = q(\alpha) + \frac{\xi + k(q(\alpha) - u)}{1-k} = \frac{q(\alpha)}{1-k} + \frac{\xi + ku}{1-k} \tag{9}$$

2.2 Threshold selection method

One of the most difficult problems in the practical application of EVT is choosing the appropriate threshold. Threshold choice involves balancing bias and variance. An excessively low threshold may violate the asymptotic underlying the GPD approximation and, consequently, increase the bias. Conversely, an excessively high threshold may involve a smaller sample size and generate few excesses, leading to high variance in the parameter estimations (see Scarrot and McDonald, 2012).

To determine the optimal threshold, several selection methods have been proposed that can be grouped into the following categories: (i) graphic methods; (ii) ad hoc methods; (iii) methods based on goodness-of-fit contrasts; and (iv) the bootstrap bias-variance method. Due to its simplicity, the graphic method most commonly used in practice is the mean excess plot method introduced by Davison and Smith (1990). This instrument is a graphical tool based on the sample means of the excesses function (SMEF), which is defined as:

$$SMEF(u) = \frac{\sum_i^{N_u}(r_i - u)_{\{r_i > u\}}}{N_u} \tag{10}$$

The sample means excesses function (SMEF) is an estimate of the excess mean function (MEF):

$$e(u) = E[(X-u)|X > u] \tag{11}$$

For the GPD, the excess mean function is given by a linear function in u:

$$e(u) = \frac{\xi}{1-k} + \frac{k}{1-k}u \tag{12}$$

This finding means that for $0 < k < 1$ and $\xi + ku > 0$, the mean excess plot should resemble a straight line with positive slope. Thus, the general rule for the choice of optimal threshold will be to choose a value of u for which the resulting line has a positive slope. An application of this method can be found in Beirlant et al. (2004).

An alternative graphic method is to fit the GPD distribution at a range of thresholds and to seek the stability of the parameter estimates. This method involves plotting $\hat{k}$ and $\hat{\xi}$ together with confidence intervals for each of these quantities and selecting the value of u from which the estimates are no longer stable (see Coles, 2001).

The main drawback of graphic approaches is that they can be rather subjective and require substantial expertise to interpret these diagnostics as a method of threshold selection.

Other authors have developed their own techniques to identify the optimal threshold (ad hoc). Christoffersen (2003) suggests a practical rule consisting of considering extreme values but only those observations in the upper or lower decile of the distribution. Neftci (2000), followed by Bekiros and Georgoutsos (2005), proposes the estimation of the threshold as $1.176\ \sigma_0$, where σ_0 is the standard deviation of the sample. DuMouchel (1983) proposes a simple quantile rule using an upper threshold of 10%, frequently used in practice. Ferreira et al. (2003) use the square root of the number of data (n) to specify the number of exceedances (N_u). Ho and Wan (2002) and Omran and McKenzie (1999) use the rule $N_u = n^{\frac{2}{3}}/\log[\log(n)]$ proposed by Loretan and Philips (1994) to determine the optimal number of exceedances. Reiss and Thomas (2007) choose the lowest upper-order statistic N_u to minimize $\frac{1}{N_u}\sum_{i=1}^{N_u} i^{\beta}\left|\hat{k}_i - median(\hat{k}_1.....\hat{k}_{N_u})\right|$.

The last method based on the goodness of fit consists of the following: fixed to a threshold, a generalized Pareto distribution is fitted to the excess losses on the threshold yield. The goodness of fit of the distribution is then contrasted by the Kolmogorov-Smirnov test, and the p-value corresponding to the contrast statistic is extracted. This exercise is repeated for a wide range of thresholds. Theoretically, the optimal threshold is one that generates a higher p-value. J.M. van Zyl (2011) shows that the Kolmogorov-Smirnov statistic can be used not only to test the goodness of fit of the Pareto model assumption but also as an indication of where to choose the threshold.

Lastly, other researchers have suggested using techniques that provide an optimal trade-off between bias and variance. This method involves using bootstrap simulations to numerically calculate the optimal threshold considering the trade-off between bias and variance. Applications of this method can be found in Danielsson et al. (2001), Drees et al. (2000) and Ferreira et al. (2003).

2.3 Risk measure

According to Jorion (2001), the "VaR measure is defined as the worst expected loss over a given horizon under normal market conditions at a given level of confidence". Thus, the VaR is a conditional quantile of the asset return loss distribution.

Let $r_1, r_2. r_n$ be identically distributed independent random variables representing the financial returns. Using $F(r)$ to denote the cumulative distribution function, $F(r) = Pr(r < r|\Omega_{t-1})$ conditionally on the information set Ω_{t-1}that is available at time *t*-1.

Assume that $\{r_t\}$ follows the stochastic process:

$$r_t = \mu_t + \sigma_t z_t, \quad z_t \sim iid\ (0.1) \tag{13}$$

where $\sigma_t^2 = E(z_t^2|\Omega_{t-1})$ and z_t has the conditional distribution function $G(z)$, $G(z) = P(z_t < z|\Omega_{t-1})$. The VaR with a given probability $\alpha \in$ (0, 1), denoted by $VaR(\alpha)$, is defined as the α quantile of the probability distribution of financial returns: $F(VaR(\alpha)) = Pr(r_t < VaR(\alpha))$ =α.

This quantile can be estimated as follows:

$$VaR_t(\alpha) = F^{-1}(\alpha) = \mu_t + \sigma_t G^{-1}(\alpha) \tag{14}$$

where μ_t and σ_t represent the conditional mean and the conditional standard deviation (volatility) of the returns. For estimating the volatility of the return, we use an APARCH model, which is given by the next expression:

$$\sigma_t^\delta = \alpha_0 + \alpha_1(|\varepsilon_{t-1}| - \gamma\varepsilon_{t-1})^\delta + \beta\sigma_{t-1}^\delta \tag{15}$$

$$\alpha_0, \beta, \delta > 0, \quad \alpha_1 \geq 0, -1 < \gamma < 1$$

In this model, the γ parameter captures the leverage effect (Black, 1976), which means that volatility tends to be higher after negative returns.

The ES with a given probability $\alpha \in (0, 1)$, denoted by $ES(\alpha)$, is defined as the average of all losses that are greater than or equal to VaR, i.e., the average loss in the worst α % cases: $ES_t(\alpha) = E[r|\, r < VaR(\alpha)]$.

$$ES_t(\alpha) = \mu_t + \sigma_t E[z|\, z < G^{-1}(\alpha)] \tag{16}$$

Replacing expression (6) in expression (14) and equation (9) in (16), we obtain the expressions for VaR and ES, respectively, measured under the conditional extreme value theory approach.

2.4 Backtesting

a) Backtesting VaR

To evaluate the accuracy of the VaR estimates, several tests have been used. All of these tests are based on the indicator variable. We have an exception when $r_{t+1} < VaR(\alpha)$; then, the exception indicator variable (I_{t+1}) is equal to one (zero in other cases).

To check the accuracy of the VaR estimates, we have used five standard tests: unconditional (LRuc), backtesting criterion (BTC), independent and conditional coverage (LRind and LRcc) and dynamic quantile (DQ) tests.

Kupiec (1995) shows that if we assume that the probability of obtaining an exception is constant, the number of exceptions $x = \sum I_{t+1}$ follows a binomial distribution $B(N, \alpha)$, where N represents the number of observations. An accurate measure VaR(α) should produce an unconditional coverage ($\hat{\alpha} = \frac{\sum I_{t+1}}{N}$) equal to α percent. The unconditional coverage test has a null hypothesis $\hat{\alpha} = \alpha$, with a likelihood ratio statistic:

$$LR_{uc} = 2[log(\hat{\alpha}^x(1-\hat{\alpha})^{N-x}) - \log(\alpha(1-\alpha)^{N-x})] \tag{17}$$

which follows an asymptotic $\chi^2(1)$ distribution.

A similar test for the significance of the departure of $\hat{\alpha}$ from α is the backtesting criterion statistic (BTC):

$$Z = (N\hat{\alpha} - N\alpha)/\sqrt{N\alpha(1-\alpha)} \tag{18}$$

which follows an asymptotic N(0, 1) distribution.

The conditional coverage test, developed by Christoffersen (1998), jointly examines whether the percentage of exceptions is statistically equal to the one expected ($\hat{\alpha} = \alpha$) and the serial independence of the exception indicator. The likelihood ratio statistic of this test is given by $LR_{cc} = LR_{uc} + LR_{ind}$, which is asymptotically distributed as $\chi^2(2)$, and the LR_{ind} statistic is the likelihood ratio statistic for the hypothesis of the serial independence against first-order Markov dependence[5].

Finally, the dynamic quantile test proposed by Engle and Manganelli (2004) examines if the exception indicator is uncorrelated with any variable that belong to the information set Ω_{t-1}, available when the VaR is calculated. This test is a Wald test of the hypothesis that all slopes are zero in the regression:

$$I_t = \beta_0 + \sum_{i=1}^{p} \beta_i \, I_{t-i} + \sum_{j=1}^{q} \mu_j \, X_{t-j} \tag{19}$$

where X_{t-j} are the explanatory variables contained in Ω_{t-1}. This statistic is introduced as five explanatory variable lags of VaR. Under the null hypothesis, the exception indicator cannot be explained by the level of VaR, i.e., $VaR(\alpha)$ is usually an explanatory variable to test if the probability of an exception depends on the level of the VaR.

b) Backtesting ES

In this paper, we use two backtests for the conditional expected shortfall. The first is the McNeil and Frey (2000) test, which is likely the most successful in the literature. These authors develop a test to verify that a model provides much better estimates of the conditional expected shortfall than another. The authors are interested in the size of the discrepancy between the return r_{t+1} and the conditional expected shortfall forecast $ES_t(\alpha)$ in the event of quantile violation. The authors define the residuals as follows:

$$Y_{t+1} = (r_{t+1} - ES_{t+1}(\alpha))/\sigma_{t+1} \tag{20}$$

Replacing equation (13) and equation (16) in equation (20), we have the next expression:

$$y_{t+1} = z_{t+1} - E(z|z < q_\alpha) \tag{21}$$

It is clear that, under model (5), these residuals are i.i.d. and that, conditional on $\{r_{t+1} <$ *VaRt+1(α)* or equivalent *zt+1<qα*, they have an expected value of zero. Suppose we again backtest on days in the set T. We can form empirical versions of these residuals on days when violations occur, i.e., days in which $\{r_{t+1} < VaR_{t+1}(\alpha)\}$. The authors call these residuals exceedances and denote them by

$$\{\hat{y}_{t+1}: t \,\epsilon\, T. \;\; r_{t+1} < VaR_{t+1}(\alpha)\} \;\; \text{where } \hat{y}_{t+1} = \frac{r_{t+1} - \widehat{ES}_{t+1}(\alpha)}{\hat{\sigma}_{t+1}} \tag{22}$$

---

[5] The LR_{ind} statistic is $LR_{ind} = 2\left[\log L_A - \log L_0\right]$ and has an asymptotic $\chi^2(1)$ distribution. The likelihood function under the alternative hypothesis is $L_A = (1-\pi_{01})^{N_{00}} \pi_{01}^{N_{01}} (1-\pi_{11})^{N_{10}} \pi_{11}^{N_{11}}$, where N_{ij} denotes the number of observations in state j after having been in state i in the previous period, $\pi_{01} = N_{01}/(N_{00}+N_{01})$ and $\pi_{11} = N_{11}/(N_{10}+N_{11})$. The likelihood function under the null hypothesis ($\pi_{01} = \pi_{11} = \pi = (N_{11}+N_{01})/N$) is $L_0 = (1-\pi)^{N_{00}+N_{01}} \pi^{N_{01}+N_{11}}$.

where $\widehat{ES}_{t+1}(\alpha)$ is an estimation of the conditional expected shortfall. Under the null hypothesis, in which we correctly estimate the dynamic of the process μ_{t+1} and σ_{t+1} and the first moment of the truncated innovation distribution $E(z|z < q_\alpha)$, these residuals should behave such as an i.i.d sample with a mean of zero. Thus, for testing whether the estimates of the expected shortfall are correct, we must test if the sample mean of the residual is equal to zero against the alternative that the mean of y is negative. Indeed, given a sample $\{y_{t+1}\}$ of size N (where N is the number of violations in the T period), the sample mean $\bar{y}$ converges in distribution to standard normality, as N tends to ∞ by the central limit theorem. In other words, given population mean μ_y and variance σ_y,

$$\sqrt{N}\left(\frac{\bar{y}-\mu_y}{\sigma_y}\right) \to N(0,1) \tag{23}$$

By applying the central limit theorem, the statistics for testing the null hypothesis are given by

$$t = \frac{\bar{y}}{S_y/\sqrt{N}} \sim t_{N-1} \tag{24}$$

where $\bar{y}$ and S_y are the sample mean and the sample standard deviation, respectively, of the exceedance residuals. As Wong (2010) notes, the above result will generally never be valid for sample sizes encountered in practice, due to the inherent nature of the test statistic. Therefore, we approximately interpret this backtest.

The other backtest used for the expected shortfall is the test proposed by Righi and Ceretta (2015). These authors propose an adaptation of the McNeil and Frey (2000) procedure. The authors consider the residual series (y'), which is similar to y except that they consider dispersion only for the exceptions rather than for the full sample. The dispersion is the standard deviation truncated by the VaR. The authors refer to this dispersion as shortfall deviation (SD). The SD is the square root of the truncated variance for a certain quantile conditional to the probability α, i.e., SD_{t+1}^{α}=$(var(r_{t+1}|r_{t+1} < VaR_{t+1}(\alpha)))^{1/2}$, and since $r_{t+1} = \mu_{t+1} + \sigma_{t+1}z_{t+1}$, by standardization, we obtain SD_{t+1}^{α}=$(\sigma_{t+1}^2 var[z_{t+1}|z_{t+1} < q(\alpha)])^{1/2}$, where $q(\alpha)$ is the α percentile of the innovation distribution. In this particular case, in which we assume a GPD for the innovations, the truncated variance of the innovations is given by

$$var[z|z < q(\alpha)] = \frac{1}{\alpha}\int_0^{\alpha}(q(s) - ES(s))^2\, ds \tag{25}$$

Thus, given significance level α, we can formally represent y' as follows:

$$y' = \begin{cases} SD_t^{\alpha^{-1}}(r_t - ES_t^{\alpha}). & r_i < VaR_i^{\alpha} \\ 0. & r_i \geq VaR_i^{\alpha} \end{cases} \tag{26}$$

Similar to its traditional counterpart, the Righi and Ceretta (2015) test has the null hypothesis that y' has a zero mean against the alternative that the mean of y' is negative. To avoid making any assumption about the distribution of the residuals y', the distribution of the mean ($\bar{y}'$) is found using the standard bootstrap simulation of Efron and Tibshirani (1993).

3. Case study

3.1 Dataset overview

The data consist of the S&P 500 stock index extracted from the Thomson-Reuters-Etkon database. The index is transformed into returns by taking the logarithmic differences of the closing daily price (in percentage). We use daily data for the period January 3, 2000, through December 31, 2015. The full data period is divided into a learning sample (January 3, 2000 to December 31, 2010) and a forecast sample (January 3, 2011 to December 31, 2015). Thus, we work with 4025 observations and generate 1258 out-of-sample VaR and ES forecasts. Figure 2 presents the evolution of the daily index and returns of the S&P 500. The index shows a sawtooth profile alternating periods with upward slope with a period of sudden decreases. In addition, we can observe that the range fluctuation of daily returns is not constant, which means that the variance of the returns changes over time. The volatility of S&P 500 was particularly high from 2008 to 2009, coinciding with the period known as the Global Financial Crisis. In the last years of the sample, we observe a period that is more stable. The basic descriptive statistics are provided in Table 1. The unconditional mean daily return is very close to zero (0.008%).

[Insert Table 1]

[Insert Figure 2]

The skewness statistic is negative, implying that the distribution of daily returns is skewed to the left. The kurtosis coefficient shows that the distribution has much thicker tails than the normal distribution. Similarly, the Jarque-Bera statistic is statistically significant, rejecting the assumption of normality. All this evidence shows that the empirical distribution of daily returns cannot be fit by a normal distribution, as it exhibits a significant excess of kurtosis and asymmetry (fat tails and peakness).

3.2 Parameter estimation by the maximum likelihood method

In this section, we analyse both the sensitivity of the parameters and the quantiles of the generalized Pareto distribution (GPD) to changes in threshold.

For this study, we have selected a set of 22 thresholds that correspond with the β percentiles of the S&P 500 return, for β equal to 60%, 70%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% and 99%. The value of these thresholds is presented in Table 2.

[Insert Table 2]

Let $u_1, u_2, \ldots, u_n$ be the set of thresholds selected ($n = 22$). For $j = 1, \ldots, n$, let $\hat{k}_{u_j}$ and $\hat{\xi}_{u_j}$ be the estimators of the shape and scale parameters based on the exceedances over the threshold u_j. The parameters have been estimated by maximum likelihood. In Table 3, we present the estimators of those parameters in addition to their standard deviations. Figures 3 and 4 display the estimation of k and ξ, respectively, as a function of the threshold u. We observe that as the threshold increases, the value of k increases. In the case of the scale parameter, the opposite occurs; as the threshold increases, the value of ξ is reduced. As we expected, in both cases, the accuracy of the estimations decreases as the threshold increases. The estimation of the shape parameter, which determines the weight of the tail in the distribution, is very sensitive to changes in the threshold. For instance, the value of k increases by 76% when the threshold moves from the 60th percentile to the 90th percentile. From the 60th percentile to the 99th percentile, the increase is equal to 125%. The value of the scale parameter is also sensitive to changes in the threshold; however, in this case, the changes are not that striking. When the threshold moves from the 60th percentile to the 90th percentile, the value of ξ is reduced by 27%. From the 60th percentile to the 99th percentile, the decrease is 29%. Thus, in accordance with the literature, we find that the parameter estimations are very sensitive to the threshold we selected for estimating PGD.

[Insert Table 3]

[Insert Figures 3 and 4]

Second, we analyse the sensitivity of the α quantiles' generalized Pareto distribution to changes in the threshold (for α equal to 80%, 85%, 90%, 95%, 96%, 97%, 98% and 99%). These quantiles have been calculated using the expression (6). Figure 5 displays these quantiles as a function of the threshold u.

[Insert Figure 5]

At first sight, it appears that the α quantile of the GPD does not depend on the choice threshold. We observe certain differences in the quantiles calculated from the threshold corresponding to the 60th, 70th, 80th, and 99th percentiles. In Table 4, we present the differences between the α quantile obtained below the optimal threshold and the α quantile obtained for the set of thresholds selected. To calculate the optimal threshold, we have used the "excess mean plot" method (see section 2.2). Applying this technique, we find that for the S&P 500, the optimal threshold is a 1.1% return, which corresponds to the 90th percentile.

For a large set of thresholds, from a return corresponding to the 85th percentile to a return corresponding to the 93rd percentile, the differences in quantile estimation do not exceed the 2 basis points. For the thresholds corresponding to the 60th, 70th, 80th and 99th percentiles, the differences are more pronounced, achieving 20 basis points in certain quantiles. However, focusing on the higher quantiles (95th, 96th, 97th, 98th and 99th), which are relevant for risk measuring, the differences for the cited percentiles do not exceed 11 basis points. This preliminary analysis may suggest that the choice of the threshold in the framework of the POT method is not relevant in quantifying risk.

[Insert Table 4]

3.3 Sensitivity of the risk measures to changes in the threshold

From the analysis presented in the previous section, we can conclude that it is in accordance with the literature; we observe that the estimates of the parameters that describe the generalized Pareto distribution depend significantly on the threshold selected for the estimation. Thus, the results presented in the previous section justify the numerous efforts made in the literature to develop techniques to detect the optimal threshold. However, in this section, we want to go a step further by assessing the extent to which the selection of the threshold affects the quantification of financial risk. With this objective, a set of 22 thresholds has been selected. The parametric estimates corresponding to these thresholds were presented in the previous section.

To quantify the risk, we use VaR and ES measures, which were presented in section 2.3. The expression for these measures is given by:

$$VaR_t(\alpha) = F^{-1}(\alpha) = \mu_t + \sigma_t G^{-1}(\alpha) \qquad ES_t(\alpha) = \mu_t + \sigma_t E[z|\, z < G^{-1}(\alpha)] \qquad (27)$$

where σ_t represents the conditional standard deviation of the return, G^{-1} (α) is the percentile α of the GPD, and μ_t is the conditional mean return that is assumed constant ($\mu_t = \mu$). For the estimation of the conditional standard deviation of the yields, we use an APARCH model (eq. (15)).

For calculating the VaR and ES measures, the sample period is divided into a learning sample from January 3, 2000 to December 31, 2010 and a forecast sample from January 3, 2011 to the end of December 2015. For each day of the forecast period, we will generate estimations of the value at risk measure and the expected shortfall measure. These forecasting measures are obtained one day ahead at the 95% and 99% confidence levels.

In Table 5, we present the descriptive statistics of the differences between the VaR estimates from the optimal threshold (90th percentile) and VaR estimates we obtain from the remainder of the thresholds selected.

For a large set of thresholds, from a return corresponding to the 80th percentile to a return corresponding to 96th percentile, the mean of the differences does not exceed the 3 basis points with a standard deviation between 1 and 2 basis points. For the thresholds correspond to the 60th, 70th, 97th and 99th percentiles the mean of the differences in the VaR estimate at 95% confidence level, increases moving between 6 and 11 basis points. The standard deviation of these differences also increases, moving between 4 and 14 basis points. For these percentiles, the minimum difference becomes 28 basis points (60th percentile), while the maximum difference becomes 76 basis points (99th percentile). For VaR estimates at the 99% confidence level, we find similar results.

In Table 6, we present certain descriptive statistics of the differences between the ES estimates obtained from the optimal threshold (90th percentile) and the ES estimates we obtain from the rest of the thresholds selected. The results are very similar to those obtained for the VaR measure. For a large set of thresholds (from the 82nd percentile to the 96th percentile), the mean and standard deviation of the differences are very reduced, not exceeding 2 basis points. Only in the case of the threshold corresponding to the 60th, 70th and 99th percentiles, the differences are more striking.

As a resume, we find that for a large set of thresholds (the return corresponding to the 80th percentile to the 96th percentile) the quantification of risk that we obtain from VaR measures is similar. This result is keeping on for the ES measure. Thus, we can conclude that in the range noted, the choice of threshold in the framework of the POT method may not be relevant in quantifying market risk.

3.4 Analysing the quality of the risk estimates

In this section, we are interested in analysing the accuracy of the risk measures (VaR and ES) obtained from the conditional extreme value theory. In addition, we will analyse if the quality of these measures depends on the threshold selected for applying EVT. Therefore, we will use the backtesting techniques presented in section 2.4.

To evaluate the accuracy of the VaR estimates, we have used five standard tests: unconditional (LR_{uc}), backtesting criterion (BTC), independent (LR_{ind}), conditional coverage (LR_{cc}) and dynamic quantile (DQ) tests. The results of these tests are presented in Table 7. In this table, we also present the number and the percentage of exception.

The first thing that pay our attention when viewing Table 7 is that for a large set of thresholds (from the 82th percentile to the 93th percentile), the number of exceptions is exactly equal to the expected one[6]. In the cases in which the number of the exception differs from the theoretical one, the differences are very reduced. Thus, at the 95% confidence level, the percentage of exceptions ranges from 4.45% to 6.04%, corresponding to the 60th percentile and the 99th percentile. At the 99% confidence level, the percentage of exceptions ranges from 0.95% to 1.19%, also very similar to the expected one (1%).

To test statistically whether the number of exceptions is equal to the theoretical one, we use the aforementioned test. We cannot reject the null hypothesis "that the VaR estimates are accurate" for any of the thresholds selected. Only for the threshold corresponding to the 99th percentile, the backtesting criterium test (BTC) rejects this hypothesis at the 95% confidence level.

To test whether the ES estimations are correct, we use the procedure proposed by McNeil and Frey (2000) and the Righi and Ceretta (2015) test. The results of these tests are displayed in Table 8. In no case do we find evidence against the null hypothesis that the average of the discrepancy measure is equal to zero.

The results presented in this section indicate that the choice of threshold in the framework of the POT method may not be relevant in quantifying market risk when we use the VaR and ES measures for this task.

4. Robustness Analysis

In the above section, we show that the choice of threshold in the framework of the POT method may not be relevant in quantifying market risk. To corroborate the validity of this result, we carry out two robustness exercises. It appears reasonable to think that in a small sample, the

[6] For the forecasting period considered in this study, which has 1258 observations, the expected number of exceptions is 62 at a 95% confidence level and 13 at a 99% confidence level.

quantification of risk may be more sensitive to the selected threshold. Therefore, initially, we analyse the validity of this result in a smaller sample (section 4.1). Then, we extend the S&P 500 index study to a set of 14 assets: 7 stock market indexes (CAC40, DAX30, FTSE100, HangSeng, IBEX35, Merval and Nikkey), four commodities (Copper, Gold, Crude Oil Brent and Silver) and three rates of exchange (£ /€, $/€ and ¥/€) (section 4.2).

4.1 Sample size robustness

In this section, we repeat the study of the S&P 500 in a smaller sample. The sample used in this section runs from January 2010 to the end of December 2015. The full period is split into a learning sample (2010 to 2013) and a forecast period (2014 to 2015). In this case, we work with 1058 observations and generate 504 VaR and ES forecasting measures. In section 3, we worked with 4025 observations and generated 1258 VaR and ES forecasting measures.

We choose this sample size because, in market risk, there are usually no problems in obtaining the asset price data, such that to work with a very small sample is not usual. However, in other areas of risk management, such as in operational risk where one of the problems is the small sample size, it may be interesting to extend this study to much smaller samples.

Overall, the results obtained in this section are very similar to those presented in section 3. First, related to the parameter estimates, we find that the shape parameter increases as the threshold increases, while the scale parameter decreases as the threshold increases. As expected, independently of the threshold selected, the accuracy of the estimate is now lower and decreases as the threshold increases. Second, we analyse the sensitivity of the α quantile generalized Pareto distribution to changes in the threshold (for α equal to 80%, 85%, 90%, 95%, 96%, 97% 98% and 99%). Focusing on the higher quantiles (95th, 96th, 97th, 98th and 99th), which are relevant for risk measuring, we find that for a large set of thresholds, from the return corresponding to the 60th percentile to the return corresponding to the 97th percentile, the differences in the quantiles do not exceed 6 basis points[7]. Again, this preliminary analysis suggests that the selection of the threshold may not be relevant in quantifying market risk. Lastly, we quantify risk through VaR and ES measures and evaluate the quality of these forecasting measures. In Tables 9 and 10, we present the results of the backtesting for the ES and VaR measures. For all the thresholds considered and for the 95% confidence level, the number of exceptions is very similar to the theoretical one, which is 25 (see Table 10). In the cases where we find certain differences, those are very reduced. In addition, for confidence levels of 95% and 99%, the accuracy test indicates that the VaR measures are all accurate, independent of the threshold selected. The results obtained for the ES measure are also robust to the threshold chosen. According to the Righi and Ceretta (2015) test, all ES measures are correct, independent of the confidence level and the threshold chosen. However, when we use McNeil and Frey's (2000) test, the results depend on the confidence level. The measures obtained at

[7] We do not include the tables with the results to save space, but they can be obtained from the authors upon request.

the 99% confidence level are all correct, as there are no cases in which the null hypothesis that the average of the discrepancy measure is equal to zero is rejected. However, at the 95% confidence level, this hypothesis is rejected in all cases[8]. Regardless of whether the estimates are good or bad, the important thing is that the results are robust to the selected threshold.

Thus, the results presented in this section corroborate those obtained in section 3, indicating that in the case of small samples, the choice of the threshold in the framework of the POT method may not be relevant in quantifying market risk.

4.2 Asset robustness

In this section, we extend the S&P 500 index study to a set of 14 assets. The sample period considered for these assets ranges from January 2000 through December 2015. The full data period is divided into a learning sample (January 3, 2000 to December 31, 2010) and a forecast sample (January 3, 2011 to December 31, 2015).

In accordance with the performed study for the S&P 500, for each of these assets, we select a set of 22 thresholds and apply the conditional extreme value theory for forecasting, 1 day ahead, the value of the risk measure and the expected shortfall measure. Both measures have been calculated at the 95% and 99% confidence levels.

For evaluating the accuracy of the VaR estimates, we use the standard tests that we presented in section 2.4: LR_{uc}, BTC, LR_{ind}, LR_{cc} and DQ. For each asset, Table 11 displays the number of times that each of these tests is rejected for the 22 thresholds selected. In a footnote, we indicate the set of thresholds for which the null hypothesis is rejected. For instance, for CAC40, the backtesting criterium (BTC) test is rejected once for the threshold corresponding to the 99th percentile. The results obtained for VaR are as follows. According to LRuc tests, in 10 of the 14 considered assets, we do not find evidence against the null hypothesis that the "VaR(5%) estimate is accurate". This result is independent of the selected threshold, although for certain indexes, this hypothesis is rejected for certain tests for the threshold corresponding to the 99th percentile. Conversely, in certain cases, the accuracy tests provide evidence against the null hypothesis; however, in these cases, the rejection does not depend on the threshold selected. For instance, for the NIKKEY index, the DQ test rejects the null hypothesis in 21 occasions. In another example, the backtesting criterium test is rejected 21 times for gold and 18 times for current exchange £/€. Again, we find that the results obtained with respect to the accuracy of the VaR estimates do not depend on the threshold selected. Only in two punctual cases, for the silver and the current exchange $/€, the BCT test provides different results as a function of the selected threshold. For the silver, the BCT test is rejected in 12 cases, which correspond to the range percentiles [85th, 95th] plus the 99th percentile. For the current exchange $/€, the BCT test is rejected in 7 cases, which correspond to the range percentiles [88th,

[8] In contrast to McNeil and Frey (2000) using the central limit theorem to test if the mean of the discrepancy measures is zero, the results obtained with this test are less reliable than those obtained by the Righi and Ceretta (2015) test.

93th] and the 98th percentile. The results found for VaR at the 99% confidence level are very similar to those for VaR at the 95% confidence level. These results suggest that the quantification of the risk through the VaR measure does not depend on the threshold selected for this objective.

To test whether the ES estimations are correct, we use the procedure proposed by McNeil and Frey (2000) and the Righi and Ceretta (2015) test. Table 12 displays for each asset the number of the times that each of these tests is rejected for the 22 thresholds selected. Overall, we do not find evidence against the null hypothesis that the average of the discrepancy measure is equal to zero from any of these tests. Only for DAX, gold and the rate exchange ¥/€, Student's t test rejects the null hypothesis for a threshold corresponding to the 99th percentile.

The results presented in this section corroborate those obtained in the previous section, indicating that the quantification of market risk through the VaR and ES measures does not depend on the threshold selected for applying the POT method.

5. Conclusions

The conditional extreme value theory has been proven to be one of the most successful in estimating market risk. The implementation of this method in the framework of the POT model requires choosing a threshold return for fitting the generalized Pareto distribution. Threshold choice involves balancing bias and variance. To determine the optimal threshold, several techniques have been proposed such as graphic methods, ad hoc methods or methods based on goodness-of-fit contrasts. However, none of these techniques have been proven to provide better results than others.

Although many proposals have been made to determine the optimal threshold in the framework of the POT method, in this paper, we ask whether, in the financial field and specifically in measuring market risk, it is important to choose the threshold. In other words, in this study, we assess to what extent the selection of the threshold is decisive in quantifying the market risk. To measure market risk, we have used the value at risk (VaR) and expected shortfall (ES) measures. The study has been done for the S&P 500 index.

To answer the aforementioned question, how the selection of the threshold affects the estimates of the parameters of the generalized Pareto distribution and the percentiles of that distribution have been previously studied. The results obtained are as follows. First, we find that in accordance with the literature, the parameter estimations are very sensible to the selected threshold for estimating GPD. However, the quantiles of the GPD do not change much when the threshold changes, particularly for high quantiles (95th, 96th, 97th, 98th and 99th), which are relevant in risk estimation. Third, for a large set of thresholds (from the 80th percentile to the 96th percentile), the VaR estimations are practically equivalent. A similar finding occurs for the expected shortfall measure. This last result shows that in the framework of the POT method, the choice of the threshold is not relevant in the estimation of risk. When we analyse the validity of the risk measures (VaR and ES), the results are highlighted more. With the exception of certain thresholds, such as the return corresponding to the 99th percentile, all the thresholds considered provide correct estimations of VaR and ES. This result is robust to the sample size, at least for a sample size that is not inferior to 1000 data points. Consequently, we can conclude that in market risk estimation, where there is usually no problem in obtaining historical data, the researchers and practitioners should not focus excessively on the threshold choice, as a wide range produces the same risk estimates.

To corroborate these results, we have extended the S&P 500 index study to a set of 14 assets (stock market indexes, commodities and rate exchange). The results obtained for these assets corroborate the results obtained for S&P 500, indicating that the quantification of market risk through the VaR and ES measures does not depend on the threshold selected to apply the POT method.

Figure 1. Shape of the generalized Pareto distribution and the corresponding density function for ξ=1.

Panel (a). GPD

Panel (b). Tail of the density function

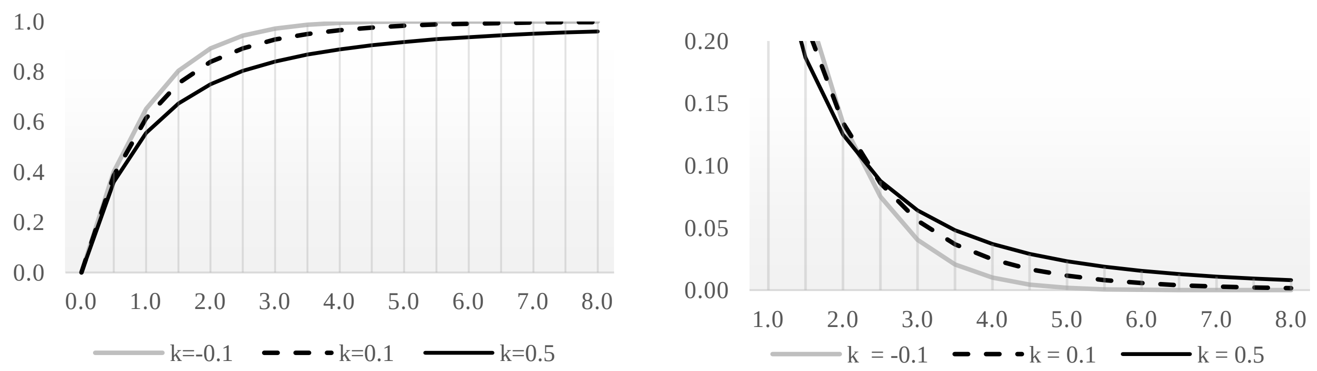


Figura 2. S&P500

Index (-) and return (-)

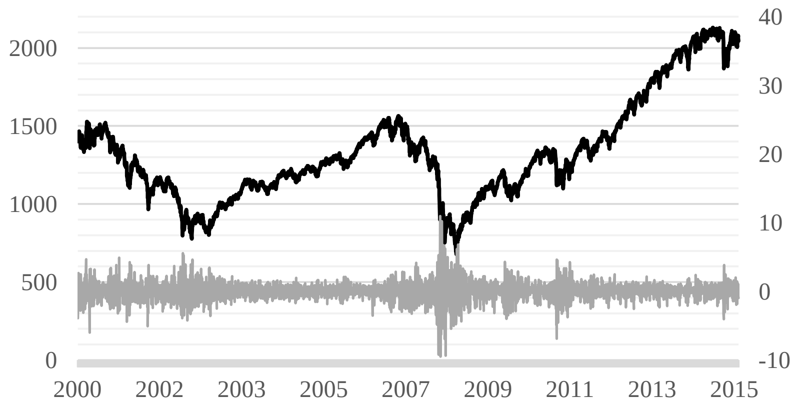


Figure 3. Shape parameter

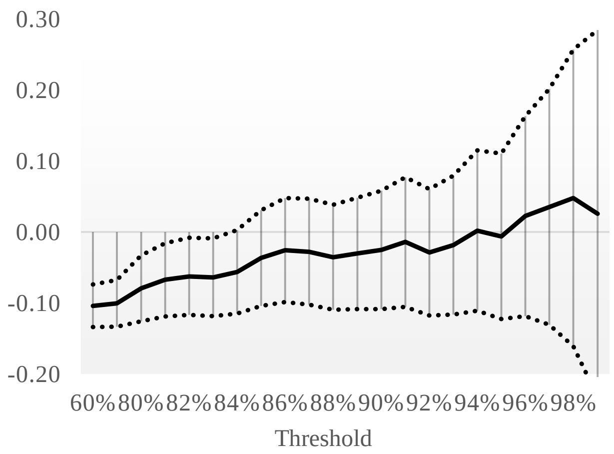
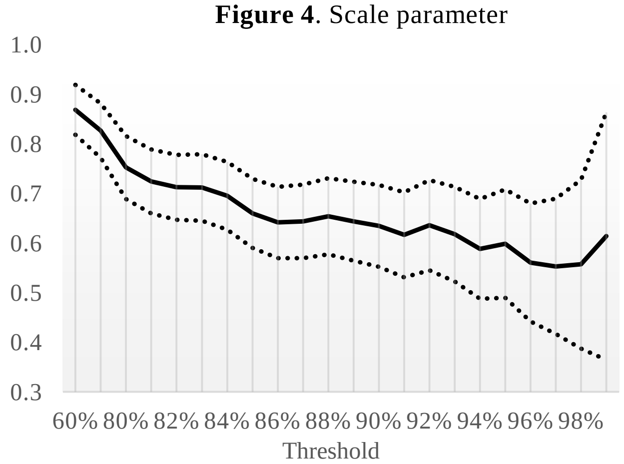


Figure 4. Scale parameter



Note: The dot lines represent the confident interval at 95% confidence level

Figure 5. Quantiles GPD

(80th, 85th, 90th, 95th, 96th, 97th, 98th and 99th)

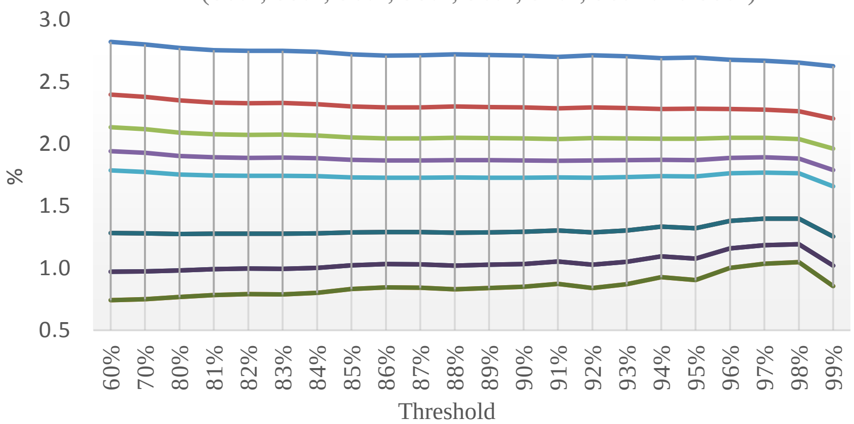


Table 1. Descriptive Statistics

	Mean	Median	Maximum	Minimum	Std. Dev.	Skewness	Kurtosis	Jarque Bera
S&P 500	0.0084	0.0535	10.957	-9.469	1.267	-0.1859* (0.039)	11.01* (0.077)	10781 (0.001)

Note: This Table presents the descriptive statistics of the daily returns of S&P 500. The sample period is from January 3rd, 2000 to December 31th, 2015. The index return is calculated as R_t=100(ln(I_t)-ln(I_{t-1})) where I_t is the index level for period t. Standard errors of the skewness and excess kurtosis are calculated as $\sqrt{6/n}$ and $\sqrt{24/n}$ respectively. The JB statistic is distributed as the Chi-square with two degrees of freedom. (*) denotes significance at the 5% level.

Table 2. Thresholds selected

Percentiles	**60%**	**70%**	**80%**	**81%**	**82%**	**83%**	**84%**	**85%**	**86%**	**87%**	**88%**
Returns	0.11	0.31	0.60	0.63	0.67	0.70	0.76	0.80	0.84	0.89	0.95
Percentiles	**89%**	**90%**	**91%**	**92%**	**93%**	**94%**	**95%**	**96%**	**97%**	**98%**	**99%**
Returns	1.02	1.10	1.16	1.24	1.32	1.44	1.57	1.72	1.95	2.26	2.78

Note: The returns are standardized (%)

Table 3. Maximum likelihood estimations

Percentiles	**60%**	**70%**	**80%**	**81%**	**82%**	**83%**	**84%**	**85%**	**86%**	**87%**	**88%**
k	-0.104 (0.015)	-0.101 (0.017)	-0.079 (0.023)	-0.067 (0.026)	-0.063 (0.027)	-0.064 (0.027)	-0.056 (0.029)	-0.037 (0.034)	-0.026 (0.037)	-0.028 (0.037)	-0.036 (0.037)
ξ	0.869 (0.025)	0.826 (0.027)	0.752 (0.032)	0.724 (0.032)	0.712 (0.033)	0.712 (0.033)	0.695 (0.034)	0.660 (0.035)	0.641 (0.036)	0.643 (0.037)	0.654 (0.038)
Percentiles	**89%**	**90%**	**91%**	**92%**	**93%**	**94%**	**95%**	**96%**	**97%**	**98%**	**99%**
k	-0.030 (0.039)	-0.025 (0.042)	-0.014 (0.046)	-0.029 (0.045)	-0.019 (0.049)	0.002 (0.057)	-0.006 (0.058)	0.022 (0.071)	0.035 (0.083)	0.048 (0.105)	0.026 (0.129)
ξ	0.064 (0.040)	0.635 (0.041)	0.616 (0.043)	0.636 (0.045)	0.618 (0.048)	0.588 (0.050)	0.599 (0.055)	0.561 (0.059)	0.553 (0.068)	0.557 (0.085)	0.614 (0.125)

Note: k: Shape parameter; ξ: scale parameter. The standard deviation is given in parenthesis.

Table 4. Differences in quantiles

	Quantiles							
Threshold	**80%**	**85%**	**90%**	**95%**	**96%**	**97%**	**98%**	**99%**
60%	-11(*)	-6	-1	6	7	9	10	11
70%	-10	-6	-1	5	6	7	9	9
80%	-8	-5	-2	3	4	5	6	6
81%	-6	-4	-2	2	3	3	4	4
82%	-6	-4	-2	1	2	3	4	4
83%	-6	-4	-2	2	2	3	4	4
84%	-5	-3	-1	1	2	2	3	3
85%	-2	-1	-1	0	1	1	1	1
86%	0	0	0	0	0	0	0	0
87%	-1	0	0	0	0	0	0	0
88%	-2	-1	-1	0	0	1	1	1
89%	-1	-1	0	0	0	0	0	0
90%	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
91%	2	2	1	0	0	0	-1	-1
92%	-1	-1	0	0	0	0	0	0
93%	2	2	1	0	0	0	0	-1
94%	8	6	4	1	1	0	-1	-2
95%	5	4	3	1	0	0	-1	-2
96%	15	12	9	4	2	1	-1	-3
97%	19	15	11	4	2	0	-2	-4
98%	20	16	11	3	2	-1	-3	-6
99%	1	-1	-4	-7	-8	-8	-9	-8

(*) For the case of the S&P500, the difference in the percentile 80th of the generalized Pareto distribution obtained for a threshold corresponding to the 60th percentile and the optimal threshold is equal to 11 basis points. The optimal threshold corresponds to the 90th percentile. We shaded in light gray the differences that oscillate between 3 and 4 basis points. Differences greater than 4 basis points are shaded in dark gray.

Table 5. Differences between VaR estimates. Descriptive statistics.

	95% confidence level				99% confidence level			
Threshold (u)	Mean	S.D.	Max	Min	Mean	S.D.	Max	Min
60%	-7	4	-3	-28	-7	4	-1	-28
70%	-6	3	-2	-26	-5	3	-1	-23
80%	-3	1	-1	-8	-1	1	0	-4
81%	-2	1	-1	-8	-1	0	0	-3
82%	-2	1	-1	-5	0	0	0	-2
83%	-1	1	-1	-4	0	0	1	-2
84%	-1	1	0	-4	0	0	1	-1
85%	-1	1	0	-3	0	0	1	0
86%	0	0	0	-2	0	0	1	0
87%	0	0	1	0	0	0	0	-1
88%	0	0	0	0	0	0	0	-1
89%	0	0	0	0	0	0	0	0
90%	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
91%	0	0	1	0	0	0	1	-2
92%	0	0	1	0	0	0	0	-1
93%	0	0	2	-1	0	1	1	-3
94%	-2	1	-1	-7	2	1	6	1
95%	-2	2	0	-15	2	1	9	0
96%	-3	2	0	-10	2	1	7	1
97%	-8	5	-3	-36	4	2	12	2
98%	-8	3	-2	-18	4	1	10	2
99%	11	14	76	-5	3	1	9	1

In this Table we present some descriptive statistics of the differences between the VaR estimations obtained under the threshold u_j $(j = 1,2,\dots,22)$ and the VaR estimates obtained under the optimal threshold. The optimal threshold is given by the 90th percentile. We shaded in light gray the differences that oscillate between 3 and 4 basis points. Differences greater than 4 basis points are shaded in dark gray.

Table 6. Differences between ES estimates. Descriptive statistics.

Threshold (u)	95% confidence level				99% confidence level			
	Mean	s.d.	Max	Min	Mean	s.d.	Max	Min
60%	-10	5	0	-39	-7	4	-1	-28
70%	-8	4	0	-35	-5	3	-1	-23
80%	-4	2	0	-11	-1	1	0	-4
81%	-3	2	0	-11	-1	0	0	-3
82%	-2	1	0	-7	0	0	0	-2
83%	-2	1	0	-6	0	0	1	-2
84%	-1	1	0	-5	0	0	1	-1
85%	-1	1	0	-4	0	0	1	0
86%	0	0	1	-2	0	0	1	0
87%	0	0	2	0	0	0	0	-1
88%	0	0	1	-1	0	0	0	-1
89%	0	0	1	-1	0	0	0	0
90%	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
91%	0	0	1	-2	0	0	1	-2
92%	0	0	0	-1	0	0	0	-1
93%	0	0	1	-2	0	1	1	-3
94%	1	0	2	0	2	1	6	1
95%	0	0	2	0	2	1	9	0
96%	0	0	3	0	2	1	7	1
97%	-1	1	1	-5	4	2	12	2
98%	0	1	6	-3	4	1	10	2
99%	9	9	48	0	3	1	9	1

In this Table we present some descriptive statistics of the differences between the ES estimations obtained under the threshold u_j $(j = 1,2,\ldots,22)$ and the ES estimates obtained under the optimal threshold. The optimal threshold is given by the 90th percentile. We shaded in light gray the differences that oscillate between 3 and 4 basis points. Differences greater than 4 basis points are shaded in dark gray.

Table 7. Backtesting VaR S&P500 (2011-2015)

Threshold	VaR 95%							VaR 99%						
	Nº excep	% excep	LRuc	BTC	LRind	LRcc	DQ	Nº excep	% excep	LRuc	BTC	LRind	LRcc	DQ
60%	56	4.45	0.549	0.814	0.823	0.815	0.185	12	0.95	0.913	0.565	0.751	0.945	0.248
70%	56	4.45	0.549	0.814	0.823	0.815	0.183	12	0.95	0.913	0.565	0.751	0.945	0.247
80%	59	4.69	0.737	0.693	0.738	0.894	0.124	12	0.95	0.913	0.565	0.751	0.945	0.247
81%	61	4.85	0.871	0.597	0.683	0.908	0.147	13	1.03	0.938	0.453	0.731	0.940	0.330
82%	62	4.93	0.939	0.546	0.656	0.903	0.159	13	1.03	0.938	0.453	0.731	0.940	0.330
83%	62	4.93	0.939	0.546	0.656	0.903	0.158	13	1.03	0.938	0.453	0.731	0.940	0.329
84%	62	4.93	0.939	0.546	0.656	0.903	0.159	13	1.03	0.938	0.453	0.731	0.940	0.330
85%	62	4.93	0.939	0.546	0.656	0.903	0.159	13	1.03	0.938	0.453	0.731	0.940	0.331
86%	62	4.93	0.939	0.546	0.656	0.903	0.158	13	1.03	0.938	0.453	0.731	0.940	0.330
87%	62	4.93	0.939	0.546	0.656	0.903	0.159	13	1.03	0.938	0.453	0.731	0.940	0.330
88%	62	4.93	0.939	0.546	0.656	0.903	0.158	13	1.03	0.938	0.453	0.731	0.940	0.329
89%	62	4.93	0.939	0.546	0.656	0.903	0.158	13	1.03	0.938	0.453	0.731	0.940	0.329
90%	62	4.93	0.939	0.546	0.656	0.903	0.158	13	1.03	0.938	0.453	0.731	0.940	0.330
91%	62	4.93	0.939	0.546	0.656	0.903	0.158	13	1.03	0.938	0.453	0.731	0.940	0.330
92%	62	4.93	0.939	0.546	0.656	0.903	0.158	13	1.03	0.938	0.453	0.731	0.940	0.329
93%	62	4.93	0.939	0.546	0.656	0.903	0.158	13	1.03	0.938	0.453	0.731	0.940	0.331
94%	61	4.85	0.871	0.597	0.683	0.908	0.147	14	1.11	0.795	0.344	0.711	0.903	0.411
95%	61	4.85	0.871	0.597	0.683	0.908	0.147	14	1.11	0.795	0.344	0.711	0.903	0.410
96%	60	4.77	0.803	0.646	0.710	0.905	0.133	14	1.11	0.795	0.344	0.711	0.903	0.410
97%	54	4.29	0.437	0.875	0.882	0.732	0.219	15	1.19	0.661	0.246	0.692	0.840	0.466
98%	55	4.37	0.492	0.847	0.853	0.776	0.483	15	1.19	0.661	0.246	0.692	0.840	0.466
99%	76	6.04	0.279	**0.045**	0.896	0.552	0.208	15	1.19	0.661	0.246	0.692	0.840	0.467

Note: The table shows p-value for the following statistics: (i) the unconditional coverage test (LRuc); (ii) the back-testing criterion (BTC); (iii) statistics for serial independence (LRind); (iv) the Conditional Coverage test (LRcc) and (v) the Dynamic Quantile test (DQ). Shaded cell indicates that the null hypothesis is rejected at 5% level of significance.

Table 8. Backtesting ES.
S&P500 (2011-2015)

Threshold	ES(95%) McNeil and Frey (2000)	ES(95%) Righi and Ceretta (2015)	ES(99%) McNeil and Frey (2000)	ES(99%) Righi and Ceretta (2015)
60%	0.60	0.61	0.34	0.67
70%	0.69	0.57	0.42	0.65
80%	0.66	0.54	0.55	0.62
81%	0.62	0.52	0.47	0.62
82%	0.60	0.52	0.49	0.61
83%	0.62	0.51	0.50	0.60
84%	0.64	0.49	0.50	0.59
85%	0.67	0.49	0.51	0.60
86%	0.69	0.49	0.51	0.62
87%	0.72	0.48	0.51	0.63
88%	0.71	0.48	0.51	0.59
89%	0.72	0.49	0.51	0.59
90%	0.71	0.47	0.51	0.60
91%	0.70	0.47	0.50	0.60
92%	0.69	0.50	0.50	0.61
93%	0.70	0.44	0.50	0.62
94%	0.78	0.42	0.45	0.63
95%	0.78	0.37	0.46	0.63
96%	0.81	0.36	0.46	0.61
97%	0.97	0.32	0.41	0.59
98%	0.91	0.19	0.41	0.60
99%	0.39	0.32	0.38	0.52

Not: The table display the p-value of the tests.

Table 9. Backtesting ES.
S&P500 (2014-2015)

Threshold	ES(95%) McNeil and Frey (2000)	ES(95%) Righi and Ceretta (2015)	ES(99%) McNeil and Frey (2000)	ES(99%) Righi and Ceretta (2015)
60%	0.00	0.84	0.13	0.83
70%	0.01	0.80	0.13	0.83
80%	0.00	0.77	0.13	0.83
81%	0.00	0.78	0.13	0.83
82%	0.00	0.81	0.13	0.84
83%	0.00	0.82	0.13	0.84
84%	0.01	0.83	0.13	0.85
85%	0.01	0.82	0.13	0.85
86%	0.00	0.83	0.13	0.84
87%	0.00	0.81	0.12	0.85
88%	0.01	0.81	0.12	0.84
89%	0.01	0.82	0.12	0.83
90%	0.01	0.84	0.12	0.85
91%	0.01	0.83	0.12	0.84
92%	0.01	0.82	0.12	0.80
93%	0.01	0.81	0.12	0.79
94%	0.01	0.80	0.12	0.82
95%	0.01	0.82	0.12	0.84
96%	0.03	0.84	0.12	0.85
97%	0.02	0.50	0.13	0.83
98%	0.04	0.37	0.11	0.81
99%	0.00	0.88	0.12	0.74

Note: The table display the p-value of the tests. Shaded cell indicates that the null hypothesis is rejected at 5% level of significance.

Table 10: Backtesting VaR
S&P500 (2014-2015)

	VaR 95%							VaR 99%						
Threshold	Nº excep.	% excep.	LRuc	BTC	LRind	LRcc	DQ	Nº excep.	% excep.	LRuc	BTC	LRind	LRcc	DQ
60%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
70%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
80%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
81%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
82%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
83%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
84%	25	4,96	0.98	0.60	0.67	0.91	0.36	2	0.40	0.31	0.84	0.93	0.59	1.00
85%	25	4,96	0.98	0.60	0.67	0.91	0.36	2	0.40	0.31	0.84	0.93	0.59	1.00
86%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
87%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
88%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
89%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
90%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
91%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
92%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
93%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
94%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
95%	26	5,16	0.91	0.61	0.73	0.94	0.41	2	0.40	0.31	0.84	0.93	0.59	1.00
96%	23	4,56	0.76	0.64	0.56	0.81	0.24	2	0.40	0.31	0.84	0.93	0.59	1.00
97%	24	4,76	0.87	0.61	0.62	0.87	0.30	2	0.40	0.31	0.84	0.93	0.59	1.00
98%	22	4,37	0.66	0.68	N.C.	0.92	0.61	2	0.40	0.31	0.84	0.93	0.59	1.00
99%	29	5,75	0.62	0.70	0.51	0.71	0.52	2	0.40	0.31	0.84	0.93	0.59	1.00

Note: The table shows p-value for the following statistics: (i) the unconditional coverage test (LRuc); (ii) the back-testing criterion (BTC); (iii) statistics for serial independence (LRind); (iv) the Conditional Coverage test (LRcc) and (v) the Dynamic Quantile test (DQ).

Table 11. Backtesting VaR

	95% confidence level					99% confidence level				
	LRuc	BTC	LRind	LRcc	DQ	LRuc	BTC	LRind	LRcc	DQ
CAC40	0	1(1)	0	0	1(1)	0	0	0	0	0
DAX30	1(1)	1(1)	0	1(1)	0	0	0	0	0	0
FTSE100	0	0	0	0	0	0	0	0	0	13(8)
HANG SENG	0	1(1)	0	0	0	0	0	0	0	0
IBEX35	0	0	0	0	0	0	0	0	0	0
MERVAL	0	0	0	0	0	0	0	0	0	0
NIKKEY	0	0	0	0	21(2)	0	0	0	0	0
S&P500	0	1(1)	0	0	0	0	0	0	0	0
COPPER	0	0	1(1)	0	1(1)	0	0	5(9)	0	22
GOLD	2	21(3)	1(1)	2(4)	1(1)	0	10(10)	0	0	12(11)
OIL BRENT	1(1)	0	0	1(1)	0	0	0	0	0	0
SILVER	0	12(5)	0	0	0	0	2(4)	0	0	20(12)
$/€	0	7(6)	0	0	0	0	10(13)	0	0	22
£/€	0	18(7)	0	0	0	0	15(14)	0	0	0
¥/€	1(1)	1(1)	0	1(1)	1(1)	0	0	0	0	0

Note: The table counts the number of rejections for the 22 thresholds (u) considered. Reject for: (1)) threshold corresponding to 99th percentile (u=99%); (2) all thresholds except for 99th percentile; (3) all except for 60th percentile; (4) thresholds corresponding to 98th and 99th percentiles; (5) thresholds in the percentiles range [85th, 95th] and the 99th percentile; (6) thresholds in the percentiles range [88th, 93th]; (7) all thresholds expect 60th, 95th, 97th and 99th; (8) thresholds in the percentiles range [60th, 90th]; (9) thresholds corresponding to percentiles 80th, 81th, 82th and 83th; (10) thresholds in the percentiles range [88th, 94th] and 96th, 97th and 98th; (11) thresholds in the percentiles range [88th, 99th]; (12) all thresholds except for 98 and 99th; (13) thresholds corresponding to percentiles 89th and 90th and the range [92th, 99th]; (14) threshold corresponding to the range percentiles [85th, 99th].

Table 12. Backtesting ES

	95% confidence level		99% confidence level	
	McNeil and Frey (2000)	Righi and Ceretta (2015)	McNeil and Frey (2000)	Righi and Ceretta (2015)
CAC40	0	0	0	0
DAX30	1(1)	0	0	0
FTSE100	0	0	0	0
HANGSENG	0	0	0	0
IBEX35	0	0	0	0
MERVAL	0	0	0	0
NIKKEY	0	0	0	0
S&P500	0	0	0	0
COPPER	0	0	0	0
GOLD	1(1)	0	0	0
OIL BRENT	0	0	0	0
SILVER	0	0	0	0
$/€	0	0	0	0
£/€	0	0	0	0
¥/€	1(1)	0	0	0

Note: The table counts the number of rejections for all thresholds (u) considered. (1) Rejected for the threshold corresponding to the 99th percentile.

References

[1] Abad P., Benito S. & López-Martín C. (2014). A Comprehensive Review of Value at Risk Methodologies. The Spanish Review *of Financial Economic*, 12, 15-32.

[2] Acerbi C. & Taasche D. (2002). On the Coherence of Expected Shortfall. *Journal of Banking and Finance*, 26(7), 1487-1503.

[3] Artzner P., Delbaen F., Eber J.M. & Heath D. (1999). Coherent Measures of Risk. *Mathematical Finance*, 9(3), 203-28.

[4] Balkema A., & De Haan L. (1974). Residual Life Time at Great Age. *The Annals of Probability*, *2*(5), 792-804.

[5] Basel Committee on Banking Supervision. (2017). High-level summary of Basel III reforms. Basel, Switzerland: Bank for International Settlements (BIS). Available at: https://www.bis.org/bcbs/publ/d424_hlsummary.pdf

[6] Basel Committee on Banking Supervision. (2013). Fundamental review of the tradingbook: A revised market risk framework. Basel, Switzerland: Bank for InternationalSettlements (BIS). Available at http://www.bis.org/bcbs/publ/d305.htm

[7] Basel Committee on Banking Supervision. (2012). Fundamental review of the trad-ing book. Basel, Switzerland: Bank for International Settlements (BIS). Available at http://www.bis.org/publ/bcbs212.pdf

[8] Beirlant J, Goegebeur Y, Segers J. & Teugels, J. (2004). Statistics of Extremes: Theory and Applications, Wiley, London.

[9] Beirlant J., Vynckie P. & Teugels J.L. (1996). Tail Index Estimation, Pareto Quantile Plots and Regression Diagnostics. *Journal of the American Statistical Association*, 91, 1659-1667.

[10] Bekiros and Georgoutsos (2005). Estimation of Value-at-Risk by extreme value and conventional methods: A comparative evaluation of their predictive performance. *Journal of International Financial Markets Institutions and Money* 15(3):209-228.

[11] Black F. (1976). Studies in Stock Price Volatility Changes. Proceedings of the 1976 Business Meeting of the Business and Economics Statistics Section, American Association: 177-181.

[12] Coles S. (2001). An introduction to statistical modeling of extreme values. *British library Cataloguing in Publication Data. Springer Series in Statistics*, pp:78-84.

[13] Christoffersen (2003). Elements of Financial Risk Management. *Academic Press.*

[14] Christoffersen P. (1998). Evaluating Interval forecasting. *International Economic Review*, 39, 841-862.

[15] Dannielson J., de Haan L., Peng L. & de Vries C.G. (2001). Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation. *Journal of Multivariate Analysis* 76, 226-248.

[16] Danielsson J., Hartmann P. & de Vries C. (1998). The Cost of Conservatism. *Risk*, 11 (1), 101-103.

[17] Davison, A. & Smith, R. (1990). Models for Exceedances over High Thresholds. *Journal of the Royal Statistical Society*, 52, 3, 393-442.

[18] Drees, H., de Haan, L. & Resnick, S. (2000). How to make a Hill Plot. *Annals of Statistics*, 28(1), 254–274.

[19] DuMouchiel, M. (1983). Estimating the stable index α in order to measure tall thickness: A critique. *The Annals of Statistics,* 11(4), 1019-1031.

[20] Efron B. & Tibshirani R. J. (1993). An Introduction to the Bootstrap: Monographs on Statistics and Applied Probability, Vol. 57. *New York and London: Chapman and Hall/CRC.*

[21] Embrechts P., Resnick S. & Samorodnitsky G. (1999). Extreme Value Theory as a Risk Management Tool. *North American Actuarial Journal*, 26. pp. 30-41

[22] Engle R. & Manganelli S. (2004). CAViaR: Conditional autoregressive Value at Risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4), 367-381.

[23] Ferreira A. de Haan L & Peng L. (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics: A Journal of Theoretical and Applied Statistics.* 37(5), 401-434.

[24] Ho A.K. & Wan A.T. (2002). Testing for covariance stationarity of stock returns in the presence of structural breaks: an intervention analysis. *Applied Economics Letters* 9(7), 441-447.

[25] Iriondo A. (2017). Análisis de sensibilidad de las medidas de riesgo financiero VaR y CVaR ante cambios en el rendimiento umbral en el marco del método POT (Peak over Threshold). Master students' final work, Faculty of Economic and Business Administration (UNED).

[26] Jorion P. (2001). Value at Risk: The new benchmark for managing financial risk. McGraw-Hill.

[27] Kupiec P. (1995). Techniques for Verifying the Accuracy of Risk Measurement Models. *Journal of Derivatives*, 2, 73-84.

[28] Loretan M. & Phillips P.C.B. (1994). Testing the covariance stationarity of heavy-tailed time series: An overview of the theory with applications to several financial datasets. *Journal of Empirical Finance* 1(2), 211-248.

[29] McNeil A.J. (1998). Calculating Quantile Risk Measures for Financial Time Series Using Extreme Value Theory. Departmet of Mathematics, ETS. Swiss Federal Technical University E-Collection. http://e-collection.ethbib.etchz.ch/

[30] McNeil A.J. & Frey R. (2000). Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: an Extreme Value Approach. *Journal of Empirical Finance*, nº 7, pp. 271-300.

[31] Morgan J.P. (1996). Riskmetrics Technical Document, 4th Ed. New York.

[32] Neftci S. (2000) Value at Risk Calculations, Extreme Events and Tail Estimation. *The Journal of Derivatives.*

[33] Omran M.F. & McKenzie E. (1999). Heteroscedasticity in Stock Returns Data Revisited: Volume versus GARCH Effects. *Applied Financial Economics* 10(5):553-60.

[34] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3, 119–131.

[35] Reiss, R and Thomas, M (2007). Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. *Springer Science+Business Media.*

[36] Righi M.B. & Ceretta P.S. (2015). A comparison of Expected Shortfall estimation models. *Journal of Economics and Business*, 78, 14-47.

[37] Scarrot, A & McDonald, A (2012). A Review of Extreme Value Threshold Estimation and Uncertainty Quantification. *REVSTAT – Statistical Journal.* Volume 10, Number 1, March 2012, 33–60.

[38] Van Zyl J.M. (2011). Application of the Kolmogorov–Smirnov Test to Estimate the Threshold When Estimating the Extreme Value Index. *Communications in Statistics-Simulation and Computation*, 40(2), 199-207.

[39] Wong W.K. (2010). Backtesting Value-at-Risk Based on Tail Losses. *Journal of Empirical Finance.* Vol. 17(3), pp. 526-538.