Far East Journal of Mathematical Sciences (FJMS)
Special Volume 2013, Part II, Pages 123-138
(Devoted to articles on Comput. Sci., Info. Sci., Financial Manag. & Biol. Sci.)
Published Online: May 2013
Available online at http://pphmj.com/journals/fjms.htm
Published by Pushpa Publishing House, Allahabad, INDIA

OPERATIONAL RISK MANAGEMENT: A STATISTICAL PERSPECTIVE

Francesca Ieva, Anna Maria Paganoni and Stefano Ziller

MOX - Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
e-mail: francesca.ieva@mail.polimi.it
anna.paganoni@polimi.it
zillerstefano@gmail.com

Abstract

This work presents a statistical model for operational risk management. We distinguish different types of operational event, we model the probability of event occurrence (the frequency distribution) and the economic impact of the single event (the severity distribution), and then the aggregated distribution is obtained through convolution of frequency and severity, for each event type. The main problem is the parameters estimation of the severity distribution above a suitable threshold, that we consider as an unknown parameter to be estimated as well. An application to a case study is also presented.

2010 Mathematics Subject Classification: 91G70.
Keywords and phrases: operational risk, severity distribution, generalized Pareto distribution, statistics of extremes.
Submitted by K. K. Azad
Received November 19, 2012

1. Introduction

This work presents a statistical model for operational risk management. Such risk includes losses deriving from natural disasters, system failures, human errors or frauds. All financial institutions have to set a provision up, in order to face such losses. This statistical model is based on the analysis of operational losses time series. First of all, seven operational ETs (event types) can be distinguished, according to the different causes. The idea of the model is to fit each risk class separately and then aggregate them to obtain a single distribution. Hence, the provision can be computed through the VaR (value at risk) indicator, defined as the 99.9% quantile of the aggregated distribution. The approach proposed is an actuarial one: the probability of event occurrence (the frequency distribution) and the economic impact of the single event (the severity distribution) are treated separately, and then an aggregated distribution is obtained through convolution of frequency and severity, for each ET. A first problem arises, since losses with a small economical impact are often neglected, hence they can rarely be trusted. Thus, the severity distribution is fitted with truncated distribution, above a threshold, which is fixed by the bank. Moreover, due to the sensibility of the capital at risk with respect to high level quantiles, the right tail of the severity distribution, which includes losses above a certain threshold, which has to be estimated, is fitted with the GPD (generalized Pareto distribution), which is the most appropriate in extreme values theory. On the other hand, the frequency distribution is modeled with Poisson distribution, considering only losses above the lower threshold for the estimation. Thus, according to the actuarial approach, each ET aggregated annual loss distribution is obtained through convolution, via Monte Carlo simulation, under the appropriate independence hypothesis. Finally, the ETs multivariate distribution, on which we compute the VaR, is obtained exploiting copulas, which allows to aggregate marginal distributions maintaining the desired dependence structure. In Section 2, operational risk management is introduced and the considered ETs are detailed. In Section 3, the model for severity is presented while Section 4 is devoted to the threshold selection and the estimation of GPD parameters. Section 5 contains an application to a case study.

2. Operational Risk Management

Operational risk is defined as “the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events” [1]. The main characteristic of such risks is that, apart from the introduction of mitigation and prevention procedures, they are derived from the regular operations of the bank, and therefore cannot be avoided.

The operational events are divided into seven event types (ETs), according to the specific cause:

1. Internal fraud: misappropriation of assets, tax evasion, intentional miss marking of positions, bribery.

2. External fraud: theft of information, hacking damage, third-party theft and forgery.

3. Employment practices and workplace safety: discrimination, workers compensation, employee health and safety.

4. Clients, products, and business practice: market manipulation, antitrust, improper trade, product defects, fiduciary breaches, account churning.

5. Damage to physical assets: natural disasters, terrorism, vandalism.

6. Business disruption and systems failures: utility disruptions, software failures, hardware failures.

7. Execution, delivery, and process management: data entry errors, accounting errors, failed mandatory reporting, negligent loss of client assets.

There are different methods that can be used to define the amount of capital to be set aside in order to face the losses deriving from operational risk the following year [1]. These methods are characterized by an increasing complexity and a decreasing amount of capital at risk. The most interesting from a statistical point of view are the so-called AMA (advanced measurement approach) models, where the capital requirement is computed using internal models based on operational loss data.

Different data sources have to be used in the internal model development, in order to reach a more accurate determination of the capital at risk:

- Internal losses, deriving from operational events occurred in the bank.

- External losses, taken from public databases, usually available only above a certain threshold.

- Perspective losses, derived from scenario analysis, only for extreme losses.

The model used for the capital allocation firstly considers each risk class separately and then aggregates through copulas, which are particular multivariate distributions allowing to combine marginals maintaining the desired correlation structure (see [6] and [12]). Hence, the prevision can be computed through the VaR (value at risk) indicator, defined as the 99.9% quantile of the aggregated distribution [1].

The approach proposed for the single ET modeling is an actuarial one: the probability of event occurrence (i.e., the frequency distribution) and the economic impact of the single event (i.e., the severity distribution) are considered separately, and then an aggregated distribution is obtained through convolution of frequency and severity [11].

In the following, the attention will be set on the severity distribution, focusing on the main statistical techniques used to estimate it (see [14]).

3. The Severity Distribution

As mentioned before, the severity of a single operational event represents the probability distribution of the economical losses deriving from it.

The usual approach in operational risk management consists in the division of the severity distribution in two parts: the body is modeled with internal losses, while the right tail, composed by extreme losses, is modeled exploiting also external and scenario data, with a conservative approach. This technique is typical in risk management, since the distributions used to fit the body often underestimate the tail, thus obtaining an underestimation of the capital at risk ([13] and [10]).

The threshold u, above which events are considered extremes, is usually set by the bank management, equal to the lowest loss of external and scenario data. However, in the following, we present a new statistical method to choose this parameter in order to have the best fit of the right tail.

The resulting severity distribution is therefore composed as a mixture of two different distributions. Let X_{body} represent the random variable describing the severity distribution limited to the body, with relative density function f_{body}, and X_{tail} the random variable describing the severity distribution of the tail, with density f_{tail}. Thus, being X the random variable representing the economic impact of a single operational event, its density is obtained through combination of the two conditioned distributions:

- $f^*_{body}(x) = f_{body}(x)/F_{body}(u)$ is the body severity distribution conditioned to $X \leq u$;

- $f^*_{tail}(x) = f_{tail}(x)/[1 - F_{tail}(u)]$ is the tail severity distribution conditioned to $X > u$.

Thus, we obtain the severity distribution, as follows:

$$f(x) = \begin{cases} \omega \cdot f^*_{body}(x) & \text{if } x \leq u, \\ (1 - \omega) \cdot f^*_{tail}(x) & \text{if } x > u, \end{cases} \tag{1}$$

where $\omega = F_{body}(u)$ represents the weight of the body.

3.1. Estimation of body distribution

As far as the body is concerned, internal loss data are usually unreliable for small amounts. Therefore, losses above a fixed threshold H cannot be used in the fit of the severity distribution, implying the need for truncated distributions.

In details, if $X \sim F$, then its truncated distribution is the law of $X \mid X \geq H$ with the following cumulative distribution function $G(x)$:

$$G(x) = \begin{cases} 0 & \text{if } x < H, \\ \dfrac{F(x) - F(H)}{1 - F(H)} & \text{if } x \geq H. \end{cases} \tag{2}$$

In operational risk management, the most common theoretical distribution used for the modeling of the body is the lognormal and the Weibull distributions, whose densities expressions are given below:

$$g_{LN}(x; \mu, \sigma^2) = \frac{1}{\Phi\left(\dfrac{\mu - \log H}{\sqrt{\sigma^2}}\right)} \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\log x - \mu)^2} \mathbb{1}_{[H, \infty)}(x), \tag{3}$$

$$g_W(x; \theta, k) = \frac{k}{\theta} x^{k-1} e^{-\frac{1}{\theta}(x^k - H^k)} \mathbb{1}_{[H, \infty)}(x). \tag{4}$$

The relative parameters are computed via maximum likelihood estimators, using numerical optimization, since no explicit expression of the estimators can be obtained (see [2] and [3]).

3.2. Estimation of tail distribution

Considering the tail of the severity distribution (i.e., extreme losses), extreme value theory can be used. The idea is to use the peaks over threshold method, which allows to model the right tail of the distribution using only data above a certain threshold (see [4] and [5]).

In particular, exploiting the properties of the block maxima approach, it is possible to assume that the loss severity above the threshold u is distributed as the generalized Pareto distribution. The cumulative distribution function of $X \mid X \geq u$ is given by the following expression [8]:

$$F_{GPD}(u, \beta, \xi)(x) = 1 - \left(1 + \xi \frac{x - u}{\beta}\right)^{-1/\xi} \tag{5}$$

with relative density function:

$$f_{GPD}(u, \beta, \xi)(x) = \frac{1}{\beta}\left(1 + \xi \frac{x - u}{\beta}\right)^{-1/\xi - 1} \tag{6}$$

The parameters of the GPD are the location parameter u (i.e., the threshold), the scale parameter β, and the shape parameter ξ. To estimate scale and shape parameters, the most common methods used are maximum likelihood estimation (MLE) or probability weighted moments (PWM). As far as the MLE method is concerned, the GPD likelihood expression is the following:

$$L(\beta, \xi) = \prod_{j=1}^{k} f(y_i; \beta, \xi) = \prod_{j=1}^{k} \frac{1}{\beta}\left(1 + \xi \frac{y_j}{\beta}\right)^{-1/\xi - 1} \mathbb{1}_{[0, \infty)}(y_j). \tag{7}$$

It is straightforward to notice that no analytical solutions can be found deriving this expression (or the relative log-likelihood) with respect to β or to ξ. Thus, numerical optimization methods are required, in order to find the maximum of the log-likelihood. For this reason, PWM method is usually preferred. This is a generalization of the method of moments, introduced in [7]. Given a random variable X with cumulative distribution F, the probability weighted moments are defined as:

$$M_{p,r,s} = E[X^p \{F(X)\}^r \{1 - F(X)\}^s] \tag{8}$$

with $p, r, s \in \mathbb{R}$. The method consists in equalizing these expressions with the empirical weighted moments $\omega_{p,r,s}$ for particular values of p, r and s. In [8], the authors showed how the estimators can be found using $p = 1$ and $r = 0$, obtaining the following expressions:

$$\hat{\beta} = \frac{2\omega_{1,0,0}\omega_{1,0,1}}{\omega_{1,0,0} - 2\omega_{1,0,1}}, \tag{9}$$

$$\hat{\xi} = 2 - \frac{\omega_{1,0,0}}{\omega_{1,0,0} - 2\omega_{1,0,1}}. \tag{10}$$

We refer to [9] for the determination of the expression of the empirical weighted moments. If y_j are i.i.d. from a GPD(u, β, ξ), and $y_{(j)}$ are the ordered values, for $j = 1, \ldots, k$, then the expressions are the following:

$$\omega_{1,0,0} = \frac{1}{k}\sum_{j=1}^{k} y_j \omega_{1,0,1} = \frac{1}{k}\sum_{j=1}^{k} \frac{k-j}{k-1} y_{(j)}. \tag{11}$$

A final consideration on the usage of the extreme value theory arises, since the additional data used (external and scenario losses) might cause a distortion of the estimation, thus extreme losses have to be scaled considering the bank. In particular, the scale parameter would be affected by scaling, while the shape parameter is invariant to it, as will be shown in the following. Thus, the parameter ξ is estimated with external and scenario data, while β is obtained through a continuity condition between body and tail at u:

$$\omega \cdot f^{*}_{body}(u) = (1-\omega) \cdot f^{*}_{tail}(u) \tag{12}$$

which leads to:

$$\beta = \frac{1-\omega}{\omega \cdot f^{*}_{body}(u)} = \frac{1-F_{body}(u)}{f_{body}(u)}. \tag{13}$$

4. Threshold Selection

As mentioned above, the main problem in the use of the GPD for extreme value modeling is the determination of the appropriate threshold, above which losses can be considered extreme. This is a critical choice, since using an excessively low threshold would lead to a distortion in the estimation of the other parameters, while overestimating u causes the exclusion of a high number of data, thus a greater variability in estimation. The proposed approach exploits the behavior of the theoretical mean and the parameters with the variation of the threshold. In particular, we will show a possible way to choose the most appropriate threshold v. Let us define the mean of the excesses of a random variable X over a certain threshold v as $E[X - v | X > v]$. Then, if X is a random variable that can be approximated with a GPD(u, β, ξ), then it can be proved that the mean of the excesses over $v \geq u$ is:

$$E[X - v | X > v] = \frac{\beta + \xi(v - u)}{1 - \xi}. \tag{14}$$

In order to prove the expression (14), we can consider the expression of the conditional mean:

$$E[X - v | X > v] = \frac{1}{1 - F_X(v)} \int_v^{\infty} (X - v) f_X(x) dx.$$

Since we are considering only values of X above u, we have $F_X = F_{GPD}$ and $f_X = f_{GPD}$ as in (5) and (6), respectively, thus:

$$E[X - v | X > v] = \frac{1}{\left(1 + \xi \frac{v - u}{\beta}\right)^{-1/\xi}} \int_v^{\infty} (x - v) \frac{1}{\beta}\left(1 + \xi \frac{x - u}{\beta}\right)^{-1/\xi - 1} dx$$

$$= \frac{1}{\left(1 + \xi \frac{v - u}{\beta}\right)^{-1/\xi}} \left[\frac{\xi(v - u) + \beta + v - x}{\xi - 1} \left(1 + \xi \frac{x - u}{\beta}\right)^{-1/\xi} \right]_v^{\infty}$$

$$= \frac{\beta + \xi(v - u)}{1 - \xi}.$$

The most important property in this case is that the mean of excesses is linear with respect to the considered threshold. The idea is therefore to identify the correct threshold u' such that, for every v above it, the mean of excesses has a linear trend.

In particular, let us suppose to have a sample $x_1, \ldots, x_n$ from $X_1, \ldots, X_n$ i.i.d. We will denote with $x_1^*, \ldots, x_{n_v}^*$ the elements of $\{x : x \geq v\}$, being v a generical $v \geq u'$. The qualitative method here proposed consists of the identification of the value v such that the plot of

$$\left\{ v,\ \frac{1}{n_v}\sum_{j=1}^{n_v}(x_j^* - v) \right\}$$

has a linear trend. If such a behavior is observed, then we can infer that we are above the threshold such that X can be approximated with a GPD.

Of course, such considerations do not hold for high values of the threshold: the less is the amount of extreme data available, the poorer is the estimate of the mean, since the variability grows up.

In practical terms, the proposed approach considers the variation of the threshold above u_0, which is the minimum loss for external and scenario data collection by the bank. For each selected $u' \geq u_0$, a linear regression between the mean of excesses and the variable threshold $v \geq u'$ is carried on, considering, in particular, the relative values of the R^2. The candidate for the identification of the "real" EVT threshold is chosen as the u' with the highest value of R^2. Once this threshold has been chosen, a deeper analysis of goodness of fit of the linear model has to be carried out. Moving to the parameters variation with respect to the selected threshold, similar properties for β and ξ can be exploited in the determination of the correct threshold for EVT. In particular, if X can be approximated with a GPD(u, β, ξ), then for every $v \geq u$, $X \mid X \geq v$ can be approximated with a GPD with the following parameters:

$$\beta' = \beta + \xi(v - u), \tag{15}$$

$$\xi' = \xi. \tag{16}$$

Considering the cdf of $X \mid X \geq v$, we obtain:

$$F(x \mid X \geq v) = \frac{F_{GPD}(x) - F_{GPD}(v)}{1 - F_{GPD}(v)}$$

$$= \frac{1 - \left(1 + \xi \frac{x - u}{\beta}\right)^{-1/\xi} - \left[1 - \left(1 + \xi \frac{v - u}{\beta}\right)^{-1/\xi}\right]}{\left(1 + \xi \frac{v - u}{\beta}\right)^{-1/\xi}}$$

$$= 1 - \left[\frac{\beta + \xi(x - u)}{\beta + \xi(v - u)}\right]^{-1/\xi}$$

$$= 1 - \left[1 + \xi \frac{x - v}{\beta + \xi(v - u)}\right]^{-1/\xi}$$

which is the expression of the cdf of $X \sim \text{GPD}(v, \beta', \xi')$. Thus, expressions (15) and (16) show that the scale parameter β is linear with respect to the variation of $v \geq u$, while ξ is constant.

The approach we are proposing exploits these two properties as a further consolidation of the threshold identified through the analysis of the mean of excesses. In particular, different parameters estimation with different thresholds are carried on. Thus, the relative trends above the selected threshold are studied, in order to confirm the choice (if linear trend for β and constant for ξ are observed) or suggesting a review of the choice if such behaviors are not detected.

In the following section, we show how the previous techniques can be applied to a real database, considering, in particular, the selection of the EVT threshold.

5. Application

The data we considered are provided by one of the major Italian banks, and they have been scaled in order to maintain the privacy constraints. As far as the tail is concerned, we take into account internal, external and scenario data, which are available only above the threshold u_0, fixed by the bank. The main task is, therefore, for each event type, to identify a threshold $u' \geq u_0$ which corresponds to the best fit of the GPD, as previously discussed. We show this behavior for one selected ET. The same discussions can be carried out for the others.

We firstly considered the properties introduced through the expression (14), trying to detect which threshold would correspond to the best linear regression between the different thresholds and relative empirical means of excesses. Figure 1 shows the different values of R^2 obtained through the regressions. It results that the highest value is at $u_1 = 2798$.

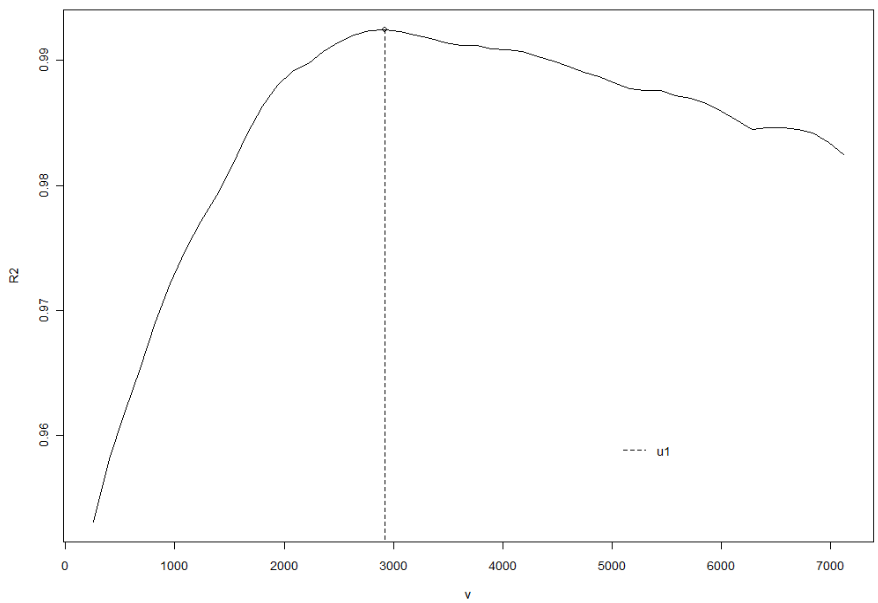

Figure 1. Identification of the most appropriate threshold u^*, obtained as the u value corresponding to the regression model between the means of the excesses and the different thresholds with the highest R^2.

We provide in Figure 2 the graph of the empirical means with respect to the different thresholds, showing the accurateness of the linear regression. We can observe how the linear trend is not obtained for all the values greater than u_0, whereas such behavior is present above the identified threshold u_1.

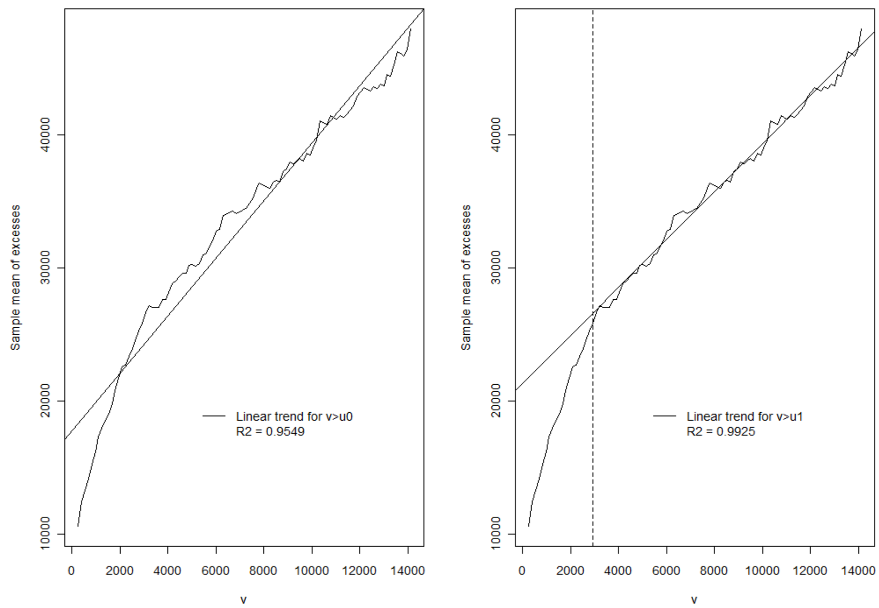

Figure 2. Graph of the empirical means of the excesses with respect to the variation of the threshold above u_0, considering the linear trends starting the regression, respectively, from u_0 and from u_1.

As before-mentioned, the same considerations can be done considering parameters estimation. In particular, we considered the two described methods (MLE and PWM) in order to further assess the threshold choice. We estimated ξ and β for different thresholds, obtaining as a result the confirmation of the necessity to choose a different value for u.

In Figure 3, we report the trend of the scale parameter β estimations with respect to different thresholds, using the two methods. The graph clearly shows how a linear trend is obtained only above the chosen threshold u_1, as expected from the expression (15). This is confirmed by the fact that the R^2 values of the two linear regressions are close to 1.

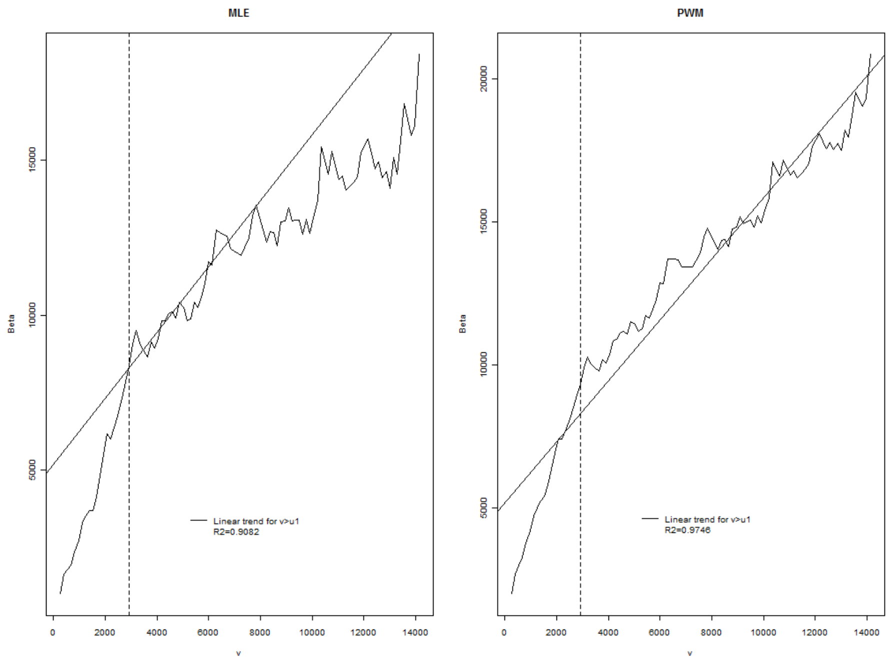

Figure 3. Graphs with the estimated values of β with methods MLE and PWM resulting from the variation of the threshold.

As far as the shape parameter ξ is concerned, the expression (15) proves that this parameter is constant with respect to the threshold variation. For instance, Figure 4 reports the estimations for different threshold, and it can be observed how MLE method shows that the most appropriate threshold is u_1, while the same conclusions cannot be obtained considering PWM method.

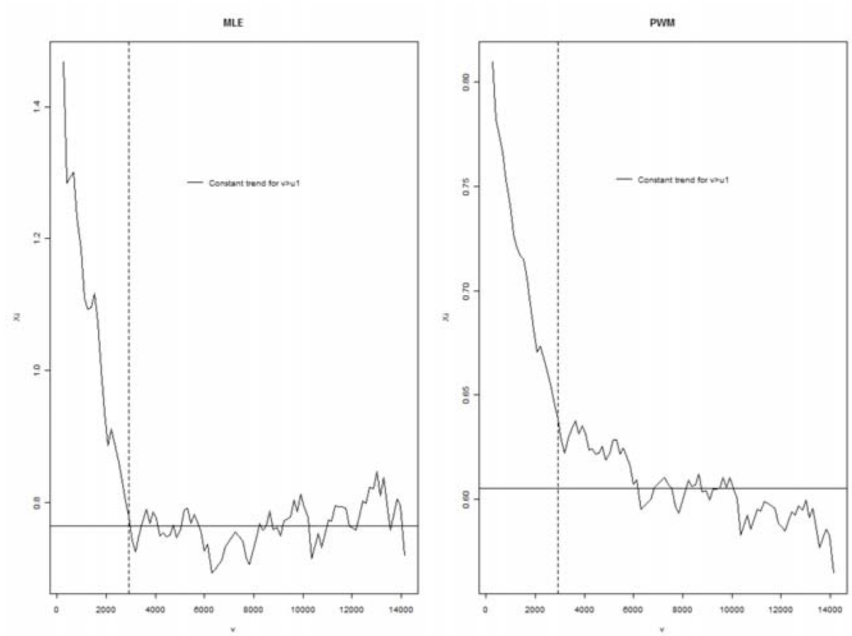

Figure 4. Graphs with the estimated values of ξ with methods MLE and PWM resulting from the variation of the threshold.

6. Conclusions

In this paper, we proposed a statistical approach to the operational risk management. In particular, we focused our attention in modeling severity distribution and choosing the right threshold for the excesses, in order to have the best fit of the right tail. According to the actuarial approach, each aggregated annual loss distribution is then obtained via Monte Carlo simulation. In general, the right threshold, above which events are considered extremes, is usually set by the bank management equal to the lowest loss of external and scenario data. The main novelty of this work is the data-driven statistical approach to the threshold estimation.

Acknowledgement

This work has been done within a collaboration between authors and Deloitte Consulting.

References

[1] Banca d'Italia, Nuove Disposizioni di Vigilanza Prudenziale per le Banche, Circolare n. 263 del 27 dicembre 2006.

[2] M. Bee, On maximum likelihood estimation of operational loss distributions, Discussion Paper No. 3, Dipartimento di Economia, Universitá degli Studi di Trento, 2005.

[3] A. C. Cohen, Maximum likelihood estimation in the Weibull distribution on complete and on censored samples, Technometrics 7(4) (1965), 579-588.

[4] S. Coles, An Introduction to Statistical Modelling of Extreme Values, Springer Series in Statistics, 2001.

[5] P. Embrechts, C. Klüppelberg and T. Mikosch, Modelling Extremal Events, Springer Series: Stochastic Modelling and Applied Probability, Vol. 33, 1997.

[6] P. Embrechts, F. Lindskog and A. McNeil, Modelling dependence with copulas and application to risk management, Handbook for Heavy Tailed Distributions in Finance, Elsevier, 2003, pp. 331-385.

[7] J. A. Greenwood, J. M. Landwehr, N. C. Matalas and J. R. Wallis, Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form, Water Resources Research 15 (1979), 1049-1054.

[8] J. R. M. Hosking and J. R. Wallis, Parameter and quantile estimation for the generalized Pareto distribution, Technometrics 29(3) (1987), 339-349.

[9] J. M. Landwehr, N. C. Matalas and J. R. Wallis, Estimation of parameters and quantiles of Wakeby distributions, Water Resources Research 15 (1979), 1361-1379.

[10] A. J. McNeil, Extreme value theory for risk managers, Internal Modelling and CAD II, RISK Books, 1999, pp. 93-113.

[11] A. J. McNeil, R. Frey and P. Embrechts, Quantitative Risk Management: Concepts, Techniques and Tools, Princeton Series in Finance, 2005.

[12] A. Nelsen, R. Frey and P. Embrechts, An introduction to copulas, Lecture Notes in Statistics, Vol. 139, Springer, 1999.

[13] R. L. Smith, Estimating tails of probability distributions, Ann. Statist. 15 (1987), 1174-1207.

[14] S. Ziller, Modelli Statistici per la gestione dei Rischi Operativi, Tesi di Laurea Magistrale in Ingegneria Matematica, Politecnico di Milano, 2011. [online] https://www.politesi.polimi.it/browse?type=author&order=ASC&rpp=20&value=ZILLER%2C+STEFANO