Treatment of incomplete data in the field of operational risk: the effects on parameter estimates, EL and UL figures

Marco Moscadelli [1], Anna Chernobai [2] and Svetlozar T. Rachev [3]

Forthcoming in "Operational Risk", June 2005

1. Introduction

In the operational risk field, the computation of the capital charge is based, in the most of the cases, on the Loss Distribution Approach, which estimates the aggregated loss distribution and derives from it appropriate figures for the Expected Losses (*EL*) and Unexpected Losses (*UL*).

The process of computing the aggregated distribution of losses is, from the statistical point of view, very challenging. This occurs for a number of reasons, most relevant of which are:

1. In the "conventional" actuarial approach, the severity and frequency components are treated as separate disjoint estimation problems; the aggregate distribution is derived as a proper combination of its estimated components. However, as the expression for the aggregated loss is only in rare cases analytically derivable from the frequency and severity distributions, approximations or simulations are usually called for (e.g. the Monte Carlo procedure). These methods require a large number of scenarios to be generated to get reliable figures of the highest percentiles of the loss distribution;
2. "Less conventional" approaches, inherited from the engineering field, as the point process, allow to address in a jointly exclusive manner the problem of estimating the parameters of the frequency and severity of the operational losses. Differently from the conventional approach, they take into consideration in the estimation procedure the (unknown) relationship between the frequency and the severity of large losses up to the end of the distribution, hence reducing the computational cost and the error related to a not analytical representation of the aggregated losses. These techniques, however, require specific conditions to be fulfilled in order to be workable;
3. Operational losses are often recorded in banks' databases starting from a threshold of a specific amount (usually \$10,000 or €5,000). This phenomenon makes the inferential procedures more complicated, and, if not properly addressed, may create unwanted biases of the aggregated loss based statistics.

While the challenges in carrying out "conventional" and "less conventional" approaches for determining the aggregated loss distributions have more recently been a subject of intense study[4], the inferential problem of dealing with incomplete data has been less investigated.
This paper moves from an initial study by Chernobai et al ("A note on the estimation of the frequency and severity distributions of operational losses", 2005), to cover, under a theoretical and practical point of view, all the issues related to the estimation of aggregated loss distribution in the presence of incomplete data available at hand. The main objective of the paper is to analytically measure the extent of the bias on the *EL* and *UL* figures when incorrect statistical approaches are used to treat incomplete data. The paper is organized as follows. Section 2 deals with the relevant definitional aspects related to incomplete data, while Sections 3 and 4 illustrate, respectively, the theoretical approaches and a specific MLE

---

[1] Banking Supervision Department, Bank of Italy
[2] Department of Statistics and Applied Probability, University of California, Santa Barbara, USA
[3] Institut für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, Germany, and Department of Statistics and Applied Probability, University of California, Santa Barbara, USA. Rachev gratefully acknowledges research support by grants from Division of Mathematical, Life and Physical Sciences, College of Letters and Science, University of California, Santa Barbara, the Deutschen Forschungsgemeinschaft and the Deutscher Akademischer Austausch Dienst.
[4] For a theoretical and practical comparison of the two approaches, see the Bank of Italy working paper ("*The Modelling of Operational Risk: Experience with the Analysis of the Data Collected by the Basel Committee*", Moscadelli, 2004).

algorithm that may be carried out to estimate loss distributions in the presence of missing data. Sections 5 and 6 focus on a typical operational risk model, the Poisson-LogNormal model; for this model the effect of adopting correct and incorrect approaches on the estimate of the capital charge relevant figures (frequency and severity parameters, *EL*, *VaR* and Expected Shortfall) are analytically computed and measured. Section 7 gives some final remarks for a general aggregated loss model and concludes.

2. Incomplete data: definitional aspects

In general, data being incomplete means that specific observations are either lost or are not recorded exactly. Based on the definitions adopted from the insurance context (see Klugman et al. 2004, "Loss Models: From Data to Decisions"), there are two main ways in which the data can be incomplete:

a. *Censored data: Data are said to be censored when the number of observations that fall in a given set is known, but the specific values of the observations are unknown; data are said to be censored from below (or left-censored) when the set is all numbers less than a specific value.*
b. *Truncated data: Data are said to be truncated when observations that fall in a given set are excluded; data are said to be truncated from below (or left-truncated) when the set is all numbers less than a specific value.*

While the "left-censored" definition would point out that only the number of observations under the threshold has been recorded (the frequency), the "left-truncated" definition would point out that neither the number (the frequency) nor the amounts (the severity) of such observations have been recorded.

In fact, in the operational risk field the second scenario is the most common. The truncated data refer to the recorded observations all of which fall above a positive threshold of a specific amount, while the missing data identify the unrecorded observations falling below the known threshold. The latter are usually called "non-randomly missing data", to distinguish them from the "randomly missing data" that may instead affect the observations that fall over the entire range of the data and can be caused, for example, by an inadequate loss data collection process.

3. Approaches with incomplete data

All statistical approaches become somewhat ad hoc in the presence of incomplete data. That is because the estimation process must account for the specific nature of the modifications. From the definition above, it is clear that the presence of censored data does affect the process of estimation of the severity distribution, but not that of the frequency; the presence of truncated data instead affects the process of estimation of both the frequency and severity distributions.

In light of the operational risk peculiarities, in the subsequent study, we address the worst situation, the truncated data problem, meaning that the number and the amounts of the observations below the set threshold are unknown.

In general, we identify four possible approaches that banks may undertake to estimate the parameters of the frequency and severity distributions in the presence of missing data. As we will see, only the last approach (Approach 4) is correct, or, to be more precise, is the best we can do under the given conditions on the data.

- *Approach 1 ("naïve"): Fit "unconditional" severity and frequency distributions to the data over the threshold.*

The term "unconditional" means that the missing observations are ignored and the observed data are treated as a complete data set during the process of fitting the frequency and severity distributions. We refer to it as the "naïve" approach, because no account is given to the missing data in the estimations of both distributions.

- *<u>Approach 2:</u> Fit "unconditional" severity and frequency distributions to the data over the threshold, and adjust the frequency parameter(s).*

The first step of such approach is identical to the previous one: unconditional distributions are fitted to the severity and frequency of the observed data.

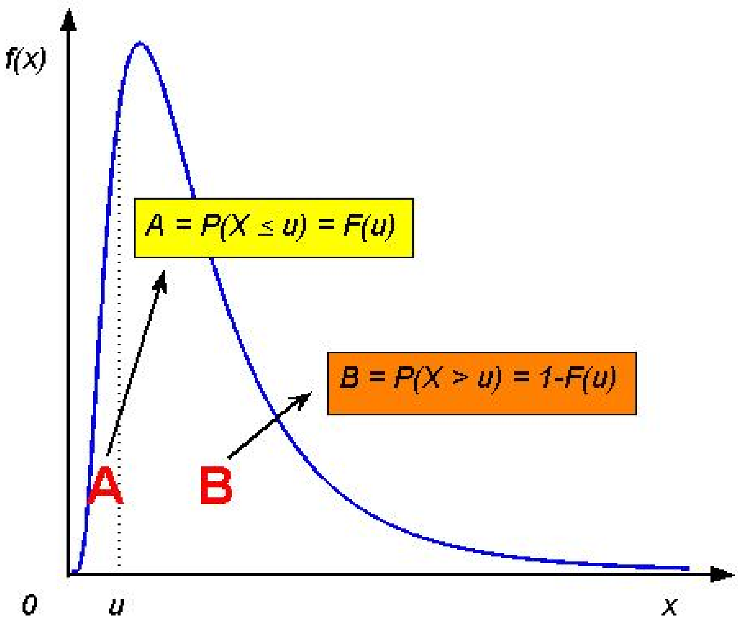
In the second step the incompleteness of the number of data is recognized and the frequency parameter(s) is adjusted according to the estimated fraction of the data over the threshold, which is obtained using the parameters of the information provided by the severity distribution.

In general, if all data were duly recorded (i.e. if the data set was complete), fitting unconditional severity distributions to such data would provide a correct estimate of the parameters of the severity distribution. Each data (or, rather, each range of data since we are dealing with continuous distributions) generated from such distribution would have a probability of:

- falling under a fixed threshold *u* (area denoted by *A* in Figure 1) equal to the distribution function computed at *u*, *F(u)*;
- falling over the threshold *u* (area denoted by *B* in Figure 1) equal to the complement of the distribution function computed at *u*, *1- F(u)*.

The areas A and B, as pointed out by the unconditional severity distribution, correspond to the fraction of missing and observed data, under the Approach 2.

Figure 1: Fraction of missing data (A) and observed data (B)



In light of that, the frequency parameter estimate(s), based on the observed data, must be adjusted for the probability of these data to occur, that is *1-F(u).* The frequency parameter(s) adjustment formula under Approach 2 may be then expressed by the following:

$$\hat{\theta}_{adj} = \frac{\hat{\theta}_{obs}}{1-\hat{F}_{uncond}(u)}, \quad \textbf{(1)}$$

where $\hat{\theta}_{adj}$ represents the adjusted frequency parameter(s) estimate and indicates the estimate of the intensity rate of the complete data, $\hat{\theta}_{obs}$ represents the unconditional (observed) frequency parameter(s) estimate, and $\hat{F}_{uncond}(u)$ represents the estimated unconditional severity distribution computed at the threshold *u*.

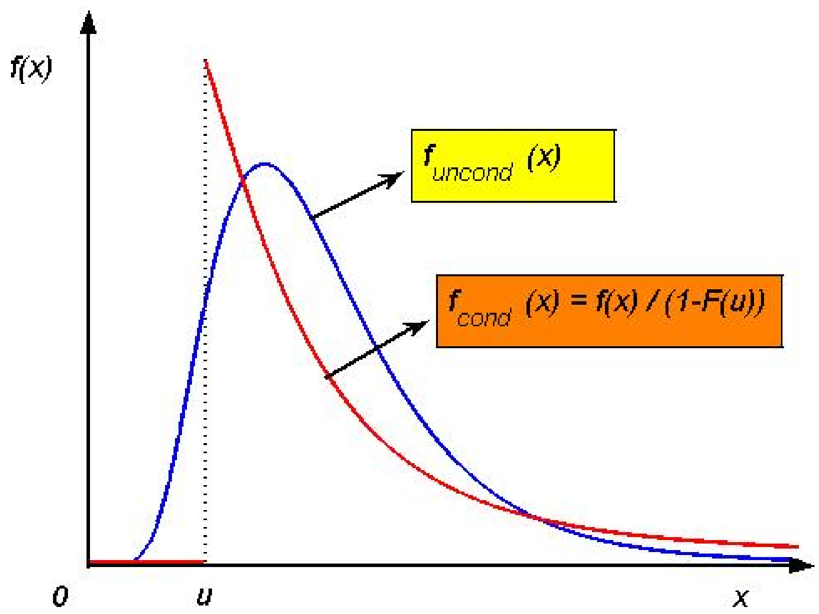
- *Approach 3:* *Fit "conditional" severity and "unconditional" frequency distributions to the data over the threshold.*

Differently from approaches 1 and 2, in this approach, the incompleteness of data is explicitly taken into account in the estimation of the severity distribution. The latter is indeed estimated "conditionally" on the fact that the observed data are now recognised as actually truncated data set and no longer a complete data set. Under the reasoning, the truncated loss severity distribution is fitted to the observed data, with the density expressed as follows:

$$f_{cond}(x) = \begin{cases} f(\mathrm{x})/(1-\mathrm{F(u)}) & \text{for } \mathrm{x} > u \\ 0 & \text{for } \mathrm{x} \le u \end{cases} \quad \textbf{(2)}$$

According to this approach, the unconditional frequency distribution is fitted, analogously to Approach 1, to the observed data in order to estimate the unconditional frequency parameter(s). No further adjustments are made. In Figure 2 the density functions for the unconditional and conditional severities are illustrated.

Figure 2: unconditional and conditional severity densities



In this approach, it is assumed that no losses under the threshold *u* have occurred, and the aggregated loss distribution is derived solely from the losses that are observed (above *u*).

- *Approach 4:* *Fit "conditional" severity and unconditional frequency distributions to the data over the threshold, and adjust the frequency parameter(s).*

The incompleteness of data is explicitly taken into account in the estimates of both the severity and frequency distributions under this approach. These distributions are indeed estimated "conditionally" on the fact that the observed data are now recognised as actually truncated data set and no longer a complete data set. As in Approach 3, the estimated severity distribution is the "conditional" one, and as in Approach 2 the frequency parameter(s) adjustment formula may be expressed by the following:

$$\hat{\theta}_{adj} = \frac{\hat{\theta}_{obs}}{1-\hat{F}_{cond}(u)} \quad , \qquad \textbf{(3)}$$

where $\hat{\theta}_{adj}$ represents the adjusted (complete-data) frequency parameter estimate(s), $\hat{\theta}_{obs}$ the unconditional (observed) frequency parameter estimate(s), and $\hat{F}_{cond}(u)$ represents the estimated conditional severity distribution computed at the threshold *u*. In the framework of operational risk modelling, this is the only relevant and correct approach, out of the four proposed. Since *all* loss data, both observed and missing, are essential for the aggregated loss derivation and subsequent estimation of *EL* and *UL*, both severity and frequency distributions are estimated in such a way so that the complete loss data comes into play.

4. Parameters estimation procedure

In general, different statistical methods may be carried out to estimate the parameters of the frequency and severity distributions. The method that has the majority of attractive properties and for this reason the one most widely used in practice, is the Maximum Likelihood Estimation (MLE), that is based on two steps: finding the functional form of the likelihood function of the data and finding the parameter value that maximizes it.

Unfortunately the MLE is more complex if the available data set is incomplete, either censored or truncated. The incompleteness of data is reflected in a restricted ability to both identify the expression for the likelihood function and maximize it. In particular, as analytical differentiation is often impossible in estimating the parameters, the computationally heavy numerical differentiation and the usual gradient-based algorithms (Newton-Raphson, Scoring, etc.) can be used for these purposes.

Additionally, a specific algorithm has been designed for MLE with incomplete data. The Expectation-Maximization algorithm (EM), developed by Dempster in 1977, is particularly convenient in cases when the range of the missing data is known, and when the MLE estimates have a closed-form solution. The algorithm has been used in a variety of applications such as probability density mixture models, hidden Markov models, cluster analysis, factor analysis and survival analysis.

A detailed explanation of the theoretical and practical elements of the EM is outside the purpose assigned to this paper. Nevertheless, two fundamental aspects of the EM are to be mentioned:

- The intuition behind EM consists in maximizing a hypothetical likelihood function, called complete likelihood, instead of the likelihood based on the observed data (either if censored or truncated), and is based on the combination of the expectations of the observed data likelihood and the missing data likelihood functions. The EM is a two-step iterative procedure that, starting from initial assigned values to the parameters to be estimated, computes and maximizes at each step the conditional expectation of the complete (log)likelihood function;

- Regardless of whether applied to truncated or censored data, the EM has some desirable properties: (a) it is simple to apply even when the form of the likelihood function is complicated, (b) it increases the likelihood at each step, and most importantly, (c) it is much less sensitive to the choice of the starting values than the direct numerical integration methods applied to Equation (2) (this means that the EM converges even for very bad initial choice values of the parameters).

5. Impact of using incorrect approaches on parameter estimates

The correct estimation of the parameters of the frequency and severity distributions is the key to determine accurate capital charge figures. Any density misspecification and/or incorrect estimation procedure would lead to biased estimates of the distribution parameters, which, in turn, would result in misleading figures of both *EL* and *UL*.

As observed in Section 3, four possible approaches may be used by banks to deal with incomplete data. Only Approach 4 appears to be correct.

The first approach - ignoring the missing data and treating the observed data as a complete data set - determines the highest biases in the estimates of the parameters of both the severity and frequency distributions; unfortunately this is the approach most followed by practitioners.

The second approach - fitting unconditional severity and frequency distributions and adjusting the frequency parameter(s) – even though it improves the previous situation, it produces biases in the parameter estimates of the severity distribution, unconditionally estimated. Consequently, these biases are reflected in the adjusted frequency parameter estimate, as it becomes incorrectly adjusted.

The third approach - fitting conditional severity and unconditional frequency distributions produces a smaller bias which comes from the unadjusted estimate of the parameters of the frequency distribution.

The fourth approach results in the minimum bias of the capital charge, and may only be due to the fact that the number of missing data is estimated rather than explicitly available. Still, the bias is expected to be at zero under the approach.

In order to fully appreciate the effect of incorrect approaches on the estimate of the parameters, we here consider a typical situation and derive an analytic expression of the biases in the parameters. We focus on two approaches: Approach 1 ("naive") and Approach 4.

The typical situation is represented by a Poisson(λ) - LogNormal(μ,σ) model for the frequency and severity distributions, respectively. We are aware that such model may not be the best one in depicting the actual behaviour of the operational risk data, as coming from the analysis of the QIS3 loss data (the cited Bank of Italy working paper puts in evidence that the operational risk losses usually follow a Binomial Negative – Lognormal model for the *EL*, and a heavy-tailed point process model for the *UL*). However, as it will be clearer later, the outcomes from such model may be easily generalized also to different (heavier-tail) operational risk models.

Given the Poisson-LogNormal model, it is then possible to express analytically the bias for the frequency and severity parameters when Approach 1 (as mentioned, the one most commonly followed by practitioners) is adopted. Using the relation between the fractions of observed and missing data to derive the true λ parameter, and using the closed-form expressions for the MLE estimates of μ and σ, it is possible to get the following expressions for the biases of the three parameters (in the ideal case, the estimates of μ and σ would correspond to the true parameters):

$$bias(\hat{\lambda}_{obs}) = -\lambda \, \Phi\left(\frac{\log u - \mu}{\sigma}\right) \qquad \textbf{< 0} \qquad \textbf{(4)}$$

$$bias(\hat{\mu}_{obs}) = \sigma \, \frac{\varphi\left(\frac{\log u - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{\log u - \mu}{\sigma}\right)} \qquad \textbf{> 0} \qquad \textbf{(5)}$$

$$bias(\hat{\sigma}^2_{obs}) = \sigma^2 \left(\frac{\log u - \mu}{\sigma} \, \frac{\varphi\left(\frac{\log u - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{\log u - \mu}{\sigma}\right)} - \left(\frac{\varphi\left(\frac{\log u - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{\log u - \mu}{\sigma}\right)} \right)^2 \right) \qquad \textbf{< 0} \text{ (since usually } \log u < \mu \text{)} \qquad \textbf{(6)}$$

where φ and Φ denote the density and distribution function of the standard Normal law and *u* is the threshold.

What is important to note is that the "naïve" Approach 1 would lead to an under-estimation of the Poisson frequency parameter λ, an over-estimation and an under-estimation of, respectively, the location (μ) and the scale (σ)parameters of the Lognormal law. While the magnitude of this effect depends on the threshold level and on the values of the true (unknown) parameters, it is interesting to examine how using the fourth approach would reduce such biases.

We illustrate such idea for the severity parameters, using true hypothetical values for μ and σ (μ_0 and σ_0) in the range 4-6.5 and 1.5-2.7, respectively, and use the threshold *u* to truncate the initially complete data set (the threshold is assumed to be equal to 50). The true fractions of missing data under such specifications are stated in Table 1.

Table 1: True fractions of missing data for various combinations of μ_0 and σ_0

σ_0 / μ_0	4	5	6.5
1.5	**0.48**	**0.23**	**0.04**
2	**0.48**	**0.29**	**0.10**
2.7	**0.49**	**0.34**	**0.17**

Figures 3 and 4 demonstrate the ratios of the estimated parameters of the truncated data under Approach 1 (left) and Approach 4 (right), for different μ_0 and σ_0 combinations. The ratio being closer to one indicates more accurate parameter estimate corresponding to the complete data, and a smaller bias.

Figure 3: Effects of using Approach 1 ("naïve") and Approach 4 on the μ estimate

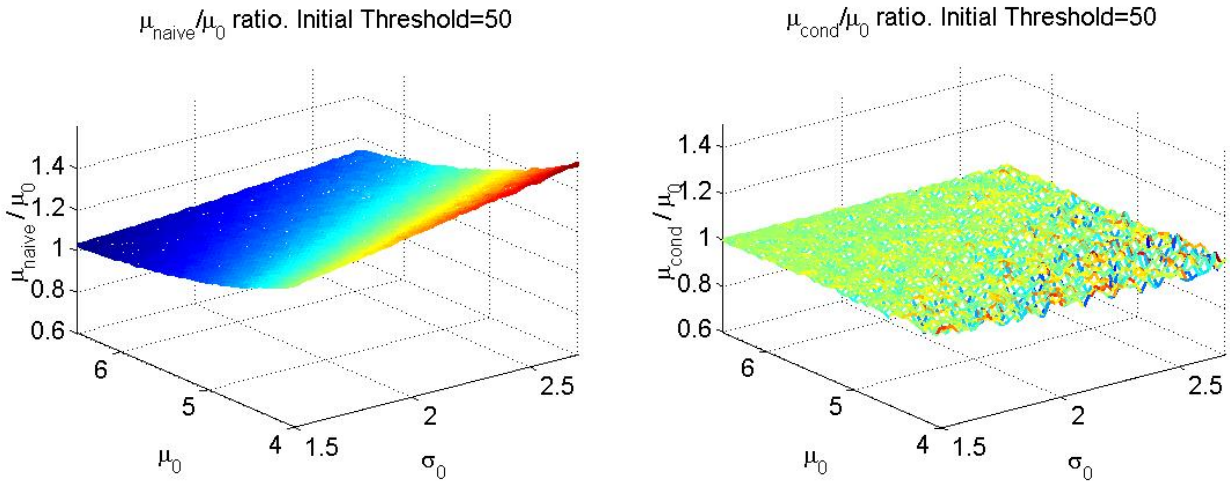
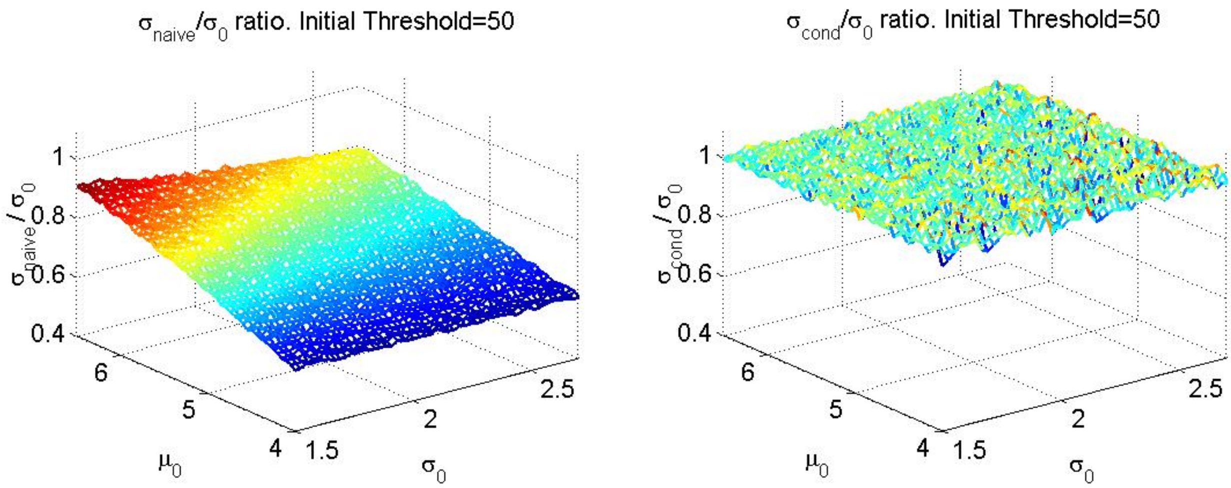


Figure 4: Effects of using Approach 1 ("naïve") and Approach 4 on the σ estimate



What it is remarkable is that, while, in general, the extent of the bias increases for lower values of the μ_0 parameter and higher values of the σ_0 parameter, Approach 1 produces significant biases in most of the possible combinations of the (true) parameters of the LogNormal law; in the worst case, it over-estimates μ and under-estimates σ by nearly 40-50%. If, instead, Approach 4 is adopted, highly accurate figures for μ and σ are obtained for the majority of the combinations of the true LogNormal parameters (that is the ratio is 1); when the bias occurs, over-estimation of μ and under-estimation of σ is roughly by at most 5%.

The results illustrated in the above figures would by all means hold in different models other than the Poisson-LogNormal. In fact, the pattern was observed when other loss distributions were considered (the results are omitted from this paper): over-estimated location parameter and under-estimated scale parameter, with the effect being more severe for heavier-tailed distributions (see the studies by Chernobai et al 2005 "Estimation of operational Value-at-Risk in the presence of minimum collection thresholds", and Chernobai et al 2005 "Modelling catastrophe claims with left-truncated severity distributions"). Applying Approach 4, in turn, would result in highly accurate estimates and would closely reflect the true nature of the complete data.

6. Impact of the biased parameter estimates on the *EL, VaR* and Expected Shortfall figures

The question that arises now is the impact – namely, magnitude and sign - the described biases would have on the *EL* and *UL* figures, which represent at the end of the day the ultimate target of the estimation process. In particular, we will explore the effects on *EL*, *VaR* and Expected Shortfall figure.

Impact on *EL*.
The expression for *EL* of an aggregated loss process is easily obtainable in the case when the conditions of homogeneity of the actuarial risk model are fulfilled, i.e. with the loss severity variable i.i.d. and independent from the loss frequency variable. In such case, *EL* is obtainable as the simple product of the Expected Frequency (*EF*) and the Expected Severity (*ES*). The problem is thus to assess the impact of the biases on the *EF* and the *ES* figures. If we use the arithmetic mean as an estimator of *EF* and *ES*[5], the expressions for *EF* and *ES* in the Poisson-LogNormal model for a unit time interval (usually, one year) with *λ, μ* and *σ* being the true model parameters, are the following:

$$EF_{Poi} = \lambda \qquad \textbf{(7)}$$

$$ES_{LogN} = \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \qquad \textbf{(8)}$$

Hence the EL becomes:

$$EL_{Poi-LogN} = \lambda \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \qquad \textbf{(9)}$$

If Approach 1 is adopted for the estimation of the parameters, each parameter enters the *EL* formula together with its bias, as discussed earlier. Therefore, the expression for the estimated *EL* will read:

$$\hat{E}L_{Poi-LogN} = \left(\lambda + bias\left(\hat{\lambda}_{obs}\right)\right)\exp\left\{\mu + bias\left(\hat{\mu}_{obs}\right) + \frac{\sigma^2 + bias\left(\hat{\sigma}^2_{obs}\right)}{2}\right\}, \qquad \textbf{(10)}$$

where the biases may be expressed in terms of the true parameters λ, μ and σ and the threshold *u*, as in Equations **(4)**, **(5)** and **(6)**.

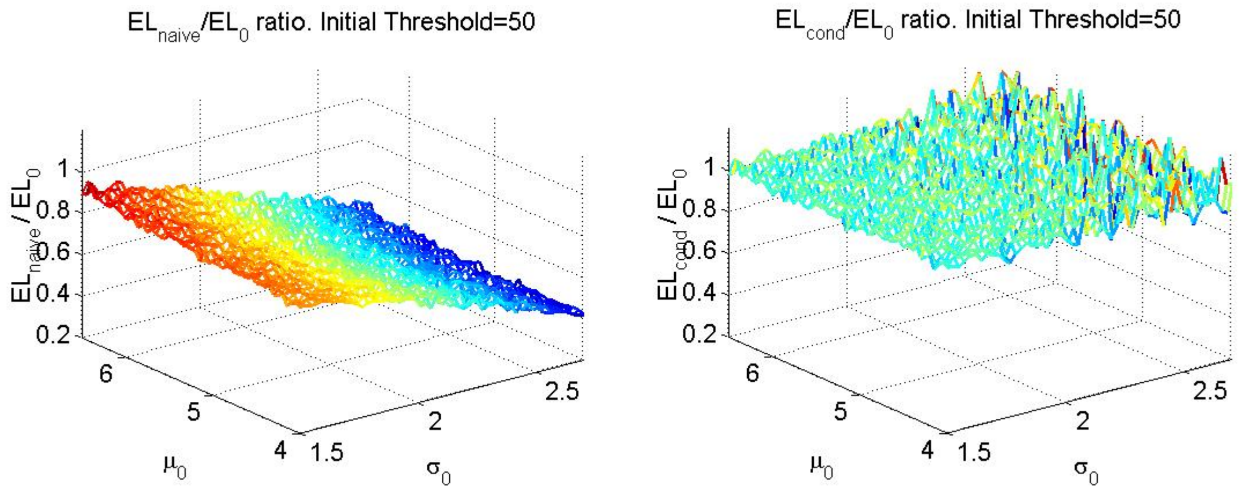
Given the estimate **(10)**, it is important to evaluate the sign and the extent of the bias, that is whether $\hat{E}L$ over-estimates or under-estimates the true EL, and the magnitude of such eventual bias. For this purpose, we compare the $\hat{E}L$ estimates under Approach 1 and 4 with the true EL values for a variety of simulated scenarios that involve different combinations of the parameters *μ* and *σ*. In particular, *μ* and *σ* are assumed to vary almost continuously in plausible ranges (50x50 combinations of μ_0 and σ_0, considered within the range of 4-6.5 and 1.5-2.7, respectively). The threshold is fixed at 50, as earlier. The following Figure 5 compares the ratios of the unconditional (Approach 1, "naïve") and conditional (Approach 4)

---

[5] In the context of the Basel Accord, there is no definition of *EL*, hence of its components *EF* and *ES*. Even if in this study we use the arithmetic mean as a measure of *EL*, this does not necessarily mean that it represents the best candidate for *EL*. Depending on the shape and characteristics of the data, alternative, more robust, measures (as the median or the trimmed/winsorized means) could be called for to represent the typical loss experience of the bank.

$\hat{EL}$ estimates to the true EL value, for the wide range of true *μ* and *σ* of the initial complete data (and any value of *λ*, as it cancels out inside the ratio).

Figure 5: Effects of using Approach 1 ("naïve") and Approach 4 on the EL estimate



The exercise shows that Approach 1 always *under-estimates* the true value of EL : the bias is on average 35% and assumes its maximum (appr. 60%) in presence of the lowest considered values of *μ* and the highest considered values of *σ*. Given its role of scale in Equation **(10)**, the frequency does not affect the bias in relative terms, but only affects in absolute terms.

Impact on the *VaR*.

In regards to *VaR*, its expression is analytically derivable from the fact that the LogNormal distribution belongs to the class of sub-exponential distributions. Following the tail approximation of the compound Poisson process given in Embrechts et al 1997 "Modelling Extremal Events for Insurance and Finance", the following formula holds (for a unit of time):

$$VaR_{Poi-LogN,1-\alpha} \approx \exp\left\{ \mu + \sigma\, \Phi^{-1}\left(1 - \frac{\alpha}{\lambda}\right) \right\} \qquad \textbf{(11)}$$

where Φ^{-1} denotes the standard Normal quantile and *(1-α)x100%* the confidence level.

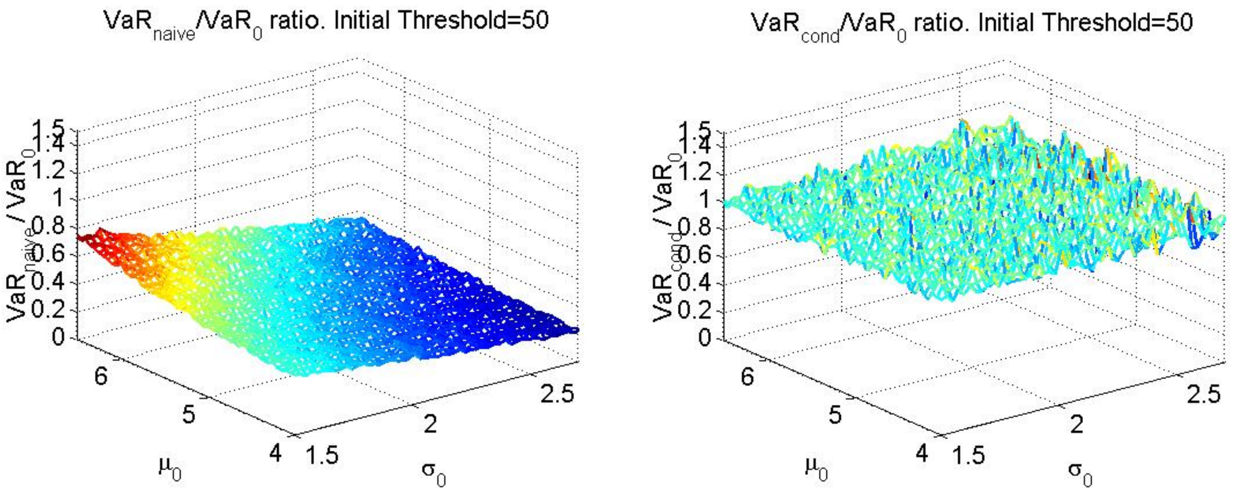
If Approach 1 is adopted for the estimation of the parameters, each parameter enters the *VaR* formula together with its bias, as computed in the expression **(3)**, **(4)** and **(5)** above. Therefore, the expression for the estimated *VaR* will turn to:

$$V\hat{a}R_{Poi-LogN,1-\alpha} \approx \exp\left\{ \mu + bias(\hat{\mu}_{obs}) + \left(\sigma + bias(\hat{\sigma}_{obs})\right) \Phi^{-1}\left(1 - \frac{\alpha}{\lambda + bias(\hat{\lambda}_{obs})}\right) \right\}, \qquad \textbf{(12)}$$

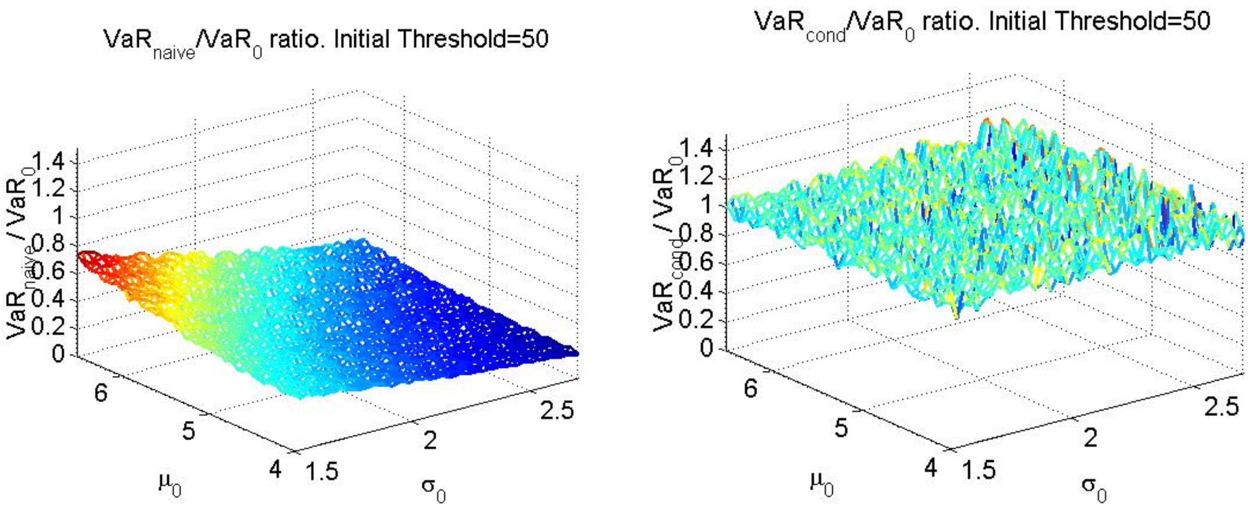
where the bias may be expressed in terms of the true parameters *λ, μ* and *σ* and the threshold *u*.

Analogously to the *EL* case, an exercise is carried out to find the sign and the extent of the bias for $V\hat{a}R$, that is whether $V\hat{a}R$ over-estimates or under-estimates VaR and the magnitude of such, eventual, bias. The $V\hat{a}R$ estimates are thus compared with the true VaR values for different combinations of the complete-data parameters *λ, μ* and *σ*. The same scenario, combinations and ranges adopted in the *EL* case are here reproduced (50x50 combinations of *μ* and *σ* in the range of values, respectively, 4-6.5 and 1.5-2.7), for four cases of λ (λ = *50, 100, 150, 200)*. The threshold is fixed at 50. Figure 6 illustrates the effects of using the "naïve" (Approach 1) and conditional (Approach 4) models on the ratios of estimated $\hat{VaR}$ to the true *VaR*, for *a=0.05*.

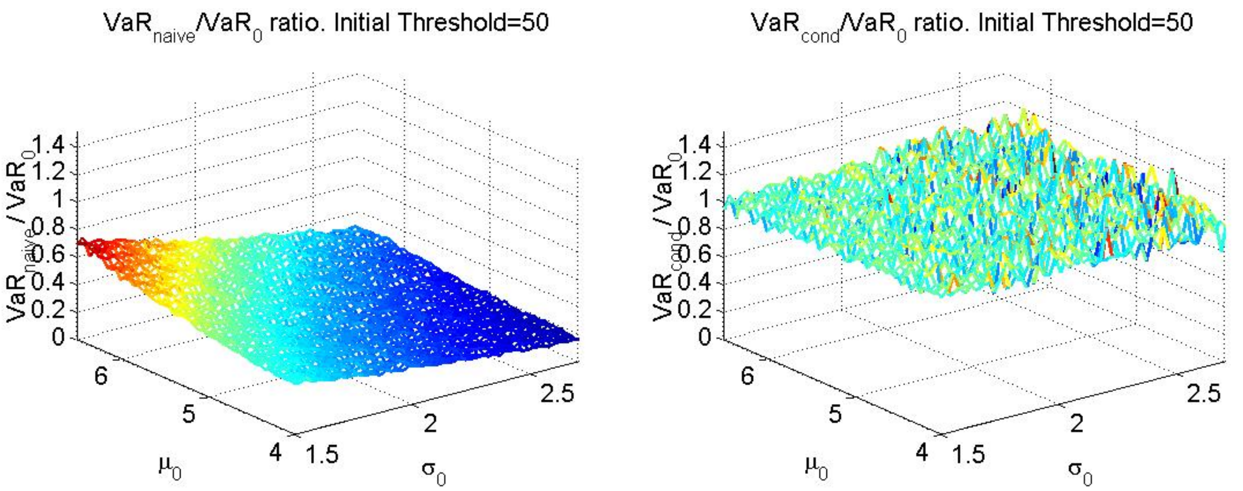
Figure 6: Effects of using Approach 1 ("naïve") and Approach 4 on the *VaR* estimates



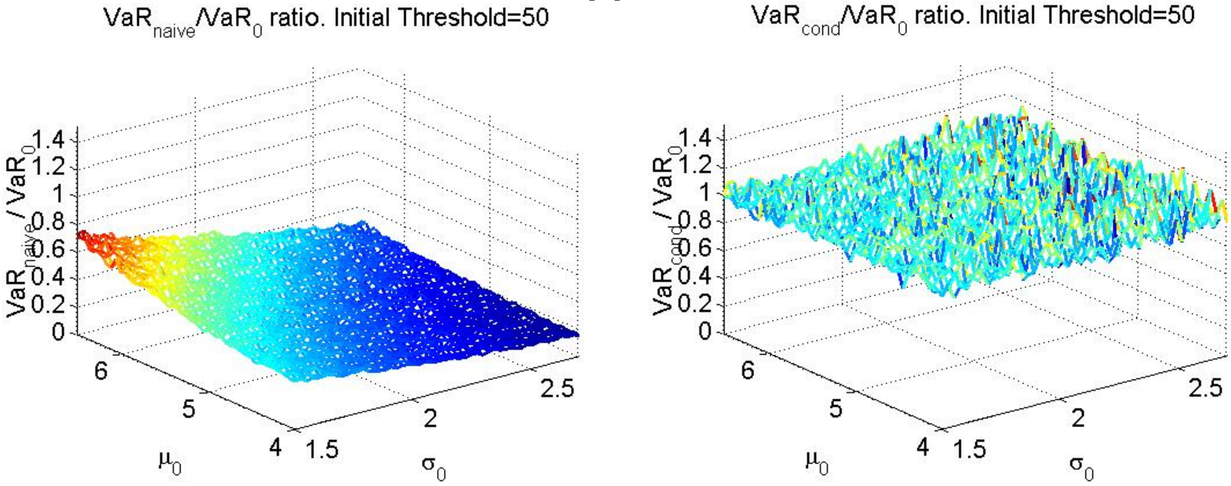
(a) λ = 50



(b) λ = 100



(c) λ = 150



(d) λ = 200

The exercise shows that Approach 1 (unconditional) always underestimates the true value of the *VaR*-based capital charge: the bias is on average 50% and attains its maximum (appr.80%) in the presence of the lowest considered value for *μ* and the highest for *σ*. The frequency has a limited impact on the bias: the highest frequency scenario (*λ* = 200) determines an increase of the bias of around 4% in comparison to the lowest frequency scenario (*λ* = 50).

Impact on the Expected Shortfall (or CVaR).

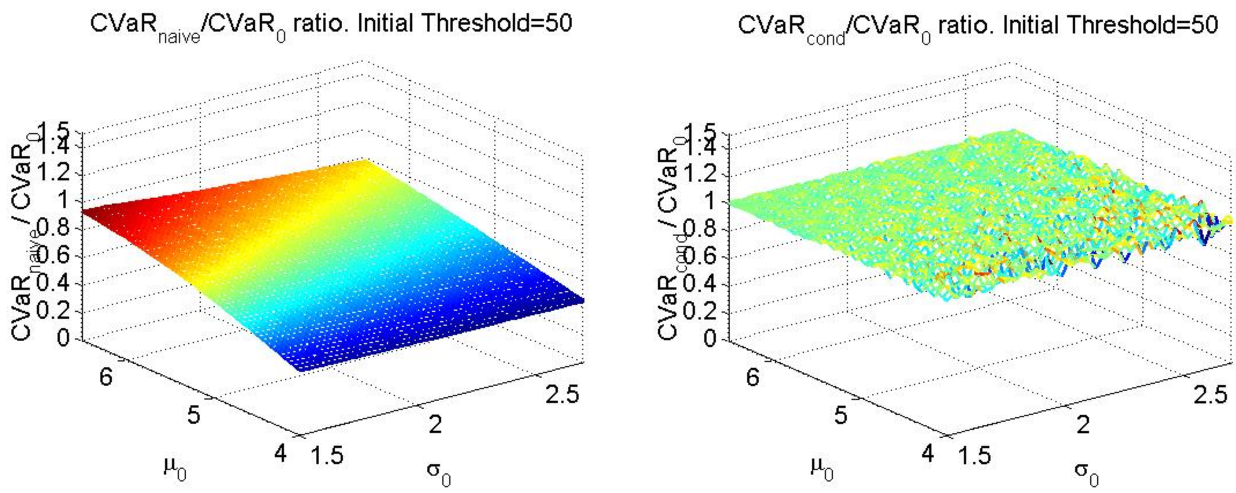
A great attention in recent literature has been given to the use of the Expected Shortfall (or the Conditional *VaR, CVaR)* as a measure of risk superior to *VaR*. As argued by Artzner et al 1997 "Thinking coherently" and 1999, "Coherent measures of risk", *CVaR* is a coherent measure of risk because it satisfies the sub-additivity property, while *VaR* can violate it. Even more importantly, *CVaR* is able to capture the tail behaviour of losses much better than *VaR*. The use of *CVaR* has been emphasized in financial models; recent references include Rachev et al 2005 "Fat-tailed and skewed asset return distributions: Implications for risk management, portfolio selection, and option pricing". *CVaR* is defined as the expected value of loss, given that the loss exceeds *VaR*. It is expressed as:

$$CVaR_{Poi-LogN,1-\alpha} = E\left[L \mid L \geq VaR_{Poi-LogN,1-\alpha}\right] = \frac{E\left[L; L \geq VaR_{Poi-LogN,1-\alpha}\right]}{\alpha} \quad \textbf{(13)}$$

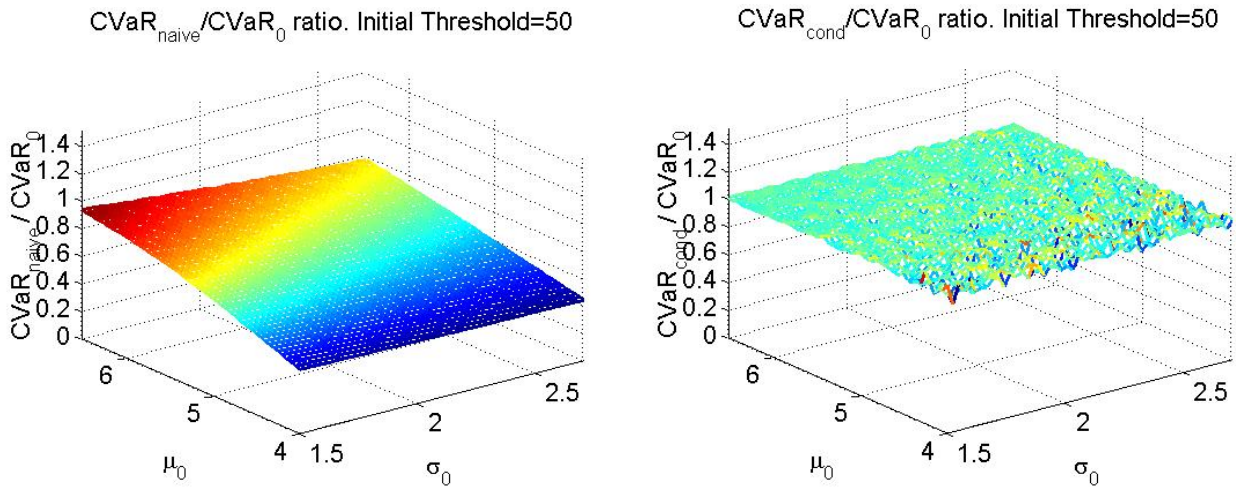
Analytical expression in a simple form exists only for the Normally distributed losses. For other cases, Monte Carlo simulations or other techniques must be used.

The $\hat{CVaR}$ estimates are further compared with the true $CVaR$ values for the same large number of combinations of the complete-data parameters *λ, μ* and *σ*. Figure 7 illustrates the effects of using the "naïve" (Approach 1) and conditional (Approach 4) models on the ratios of estimated *CVaR* to the true *CVaR*, for *α=0.05*.

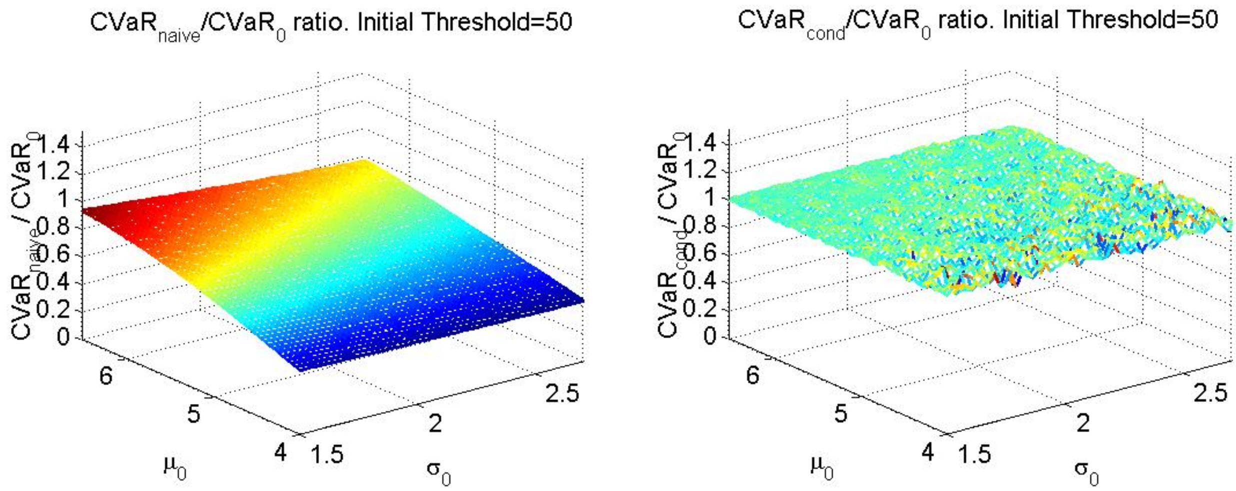
Figure 7: Effects of using Approach 1 ("naïve") and Approach 4 on the *CVaR* estimates



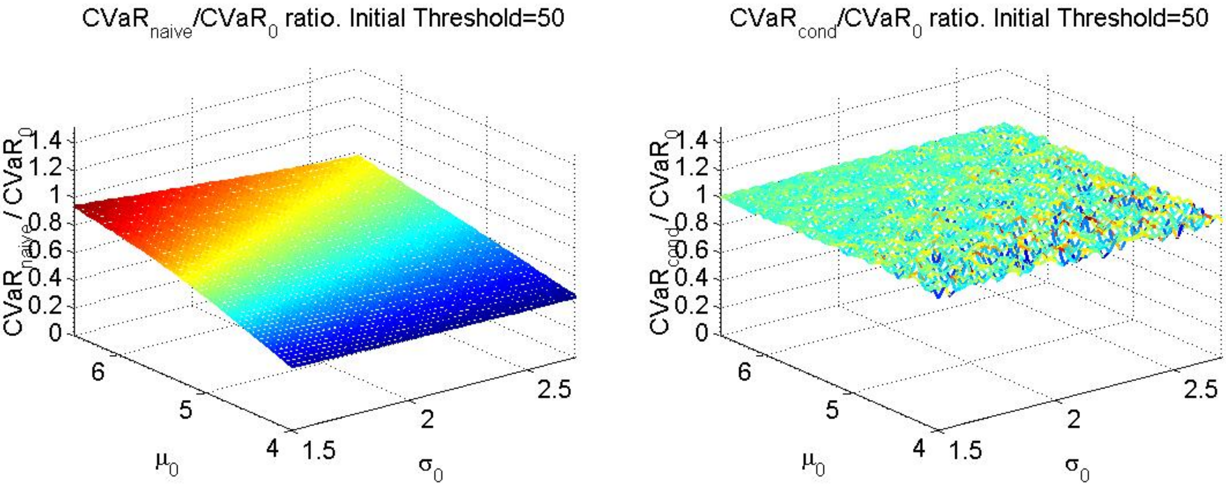
(a) λ = 50



(b) λ = 100



(c) λ = 150



(d) λ = 200

The exercise results in similar conclusions about the effect on *CVaR* to those said about the effect on *VaR*. The "naïve" approach often highly under-estimates the true *CVaR* while the conditional model captures the true *CVaR* remarkably well.

7. Conclusions

This paper deals with the problem of estimation of the aggregate operational loss distribution in the presence of incomplete data. The existence of data unrecorded under a seemingly low threshold (of, for instance, $10,000 or €5,000) has serious implications on the operational capital charge relevant figures, if not duly accounted for.

In the first part of the paper, some definitional aspects were addressed: a clear distinction between "non-randomly missing data" (for example, data that fall under a threshold of a specific amount) and "randomly missing data" (observations that are missing randomly over the entire range of the data, and are caused, for example, by an inadequate loss data collection process) was made. Within the first category, a clear line was also put between the categories of censored and truncated data, sometimes incorrectly treated as synonymous terms: while with the censored data the information loss refers only to the severity, in the truncated data the information loss occurs in both the frequency and severity.

For the truncated data (the worst case), the paper illustrated possible approaches that may be carried out to estimate the parameters of the frequency and severity distributions: four approaches were depicted, from the fully "unconditional" that ignores the missing observations and treats the observed data as a complete data set (the "naïve" approach), to the fully "conditional", where the incompleteness of data is explicitly taken into account in the estimation of both severity and frequency distributions.

A specific algorithm, the Expectation-Maximization (EM) algorithm, was then introduced as a robust iterative procedure designed for Maximum Likelihood estimation with incomplete data; in particular it was highlighted that the EM algorithm is easy to apply even when the form of the likelihood function is complicated, and results in an increased likelihood value at each iteration and, most importantly, convergence to the true parameter values is achieved even for very bad starting values.

In the second part of the paper, the effects of the use of the stated approaches on the correctness of the estimate of the capital charge relevant figures was described, analytically derived, and measured.

In particular it was stressed that the "naïve" approach – the most followed by practitioners - determines the highest biases in the estimates of the parameters of both severity and frequency distributions. This is confirmed by the Poisson (λ)-LogNormal (μ,σ) model, for which the sign and the extent of the bias were computed and then measured for a large number of the true λ-μ-σ combinations. The model demonstrates that the extent of the bias increases for lower values of the location parameter (μ) and higher values of the scale parameter (σ); in the worst case it over-estimates μ by 50% and under-estimates σ by 40%. On the other side, when the fully "conditional" approach is adopted, the estimates of μ and σ coincide with the true values for the majority of scenarios, and when the bias occurs, it stays under 5%.

Finally, the bias on the estimate of the *EL* and *VaR* (and Expected Shortfall) figures generated under the "naïve" approach in the case of the Poisson-LogNormal model was first analytically expressed and then measured. The exercise shows that this "naïve" approach always under-estimates the true values of *EL* and *VaR (*and Expected Shortfall*)*: the bias is on average 35% and 50%, respectively, and attains its maximum at roughly 60% and 80%. The frequency has a negligible impact on the bias in relative terms, but has an impact in the absolute ones. Equivalently to the conclusion made regarding the parameter estimation, correct (or minimally biased) figures for *EL* and *VaR* (and Expected Shortfall) would be obtained if the conditional distribution were fitted to the incomplete data, by adopting the Expectation-Maximization algorithm or, alternatively, direct numerical integration.

The exercise also shows that the underestimation of *EL* and *VaR* (and Expected Shortfall) figures rises when μ decreases and σ increases; this means that the bias is driven, other than by the fraction of missing data, by the asymmetry and the heavy-tailness of the model. Therefore, if instead of the Poisson-LogNormal case, models with a higher level of asymmetry and tail heaviness were used (for instance Negative Binomial - Generalized Pareto), the bias would significantly amplify, possibly up to figures bigger than 100%.

The recommendation for practitioners stemming from the depicted exercise is to fix the thresholds as low as possible and use a correct approach to estimate the parameters of the frequency and severity distributions in order to determine the *EL* and *VaR* (or Expected Shortfall) figures. By doing so, in addition to avoiding the information loss carried by the missing data, one would be able to produce accurate estimates of the operational capital charge.

References

Artzner P., F. Delbaen, J.-M. Eber, D. Heath, 1997, "*Thinking Coherently",* RISK, 10, 68-71

Artzner P., F. Delbaen, J.-M. Eber, D. Heath, 1999, "*Coherent Measures of Risk"*, Mathematical Finance, 9, no. 3, pp.203-228

Chernobai A., C. Menn, S. Trück, S. Rachev, 2005, *"A note on the estimation of the frequency and severity distributions of operational losses",* Mathematical Scientist, 30(2)

Chernobai A., C. Burnecki, S. Rachev, S. Trück, R. Weron, 2005, *"Modelling Catastrophe Claims with Left-Truncated Severity Distributions"*, submitted to Computational Statistics

Chernobai A., C. Menn, S. Trück, S. Rachev, 2005, *"Estimation of Operational Value-at-Risk in the Presence of Minimum Collection Thresholds"*, working paper

Dempster A.P., N. Laird, D. Rubin, 1977, *"Maximum Likelihood from Incomplete Data via the EM Algorithm"*, Journal of the Royal Statistical Society, 39(B), pp.1-38

Embrechts P., C. Klüppelberg, T. Mikosch, 1997, *"Modelling Extremal Events for Insurance and Finance",* Springer

Klugman S.A., H.H. Panjer, G.E. Willmot, 2004 *"Loss Models: from Data to Decisions"*, Wiley

Moscadelli M., 2004, *"The Modelling of Operational Risk: Experience with the Analysis of the Data Collected by the Basel Committee"*, Bank of Italy, working paper

Rachev S.T., C. Menn, F.J. Fabozzi, 2005, *"Fat-tailed and skewed asset return distributions: Implications for risk management, portfolio selection, and option pricing",* Wiley