Combining Underreported Internal and External Data For Operational Risk Measurement

Montserrat Guillen* Jim Gustafsson† Jens Perch Nielsen‡

7th November 2008

Abstract

Operational risk data sets have two types of sample selection problems: truncation below a given threshold due to data that are not recorded and random censoring above that level caused by data that are not reported. This paper proposes a model for operational losses that improves the internal loss distribution modelling by combining internal and external operational risk data. It also considers the possibility that internal and external data have been collected with a different truncation threshold. Moreover, the model is able to cope with unreported losses by means of an estimated underreporting function.

keywords: Operational Risk, Mixing Model, Underreporting

JEL classification: C14, C15, G32

1 Introduction

In this paper we combine the analysis of three well known challenges to the investigation of operational risk data. All these three problems stem from the difficulty of collecting operational risk data;
the data are underreported, they are truncated and, besides that, external data is often needed to incorporate more information. We present an empirical application with real data from the insurance industry. However, in order to show the methodology in an intuitive way, we first exemplify every calculation with simulated data. In the first step, the simulated internal data set and the simulated external data set have not been truncated and all operational risk losses can be assumed reported. Then, we go through the analysis of such data sets. Afterwards, we introduce into our simulation the underreporting and truncation of the internal and external data sets and we show how to overcome the consequences of these sampling restrictions in the statistical analysis. We first consider a purely parametric approach to operational loss data estimation. We then introduce a nonparametric correction to the sampling scheme of operational losses considered in this paper, because it is well known that a nonparametric correction is capable of robustifying the final result. Finally, we discuss the empirical application to real data.

According to the latest proposal of the Basel Committee, in addition to market and credit risk, operational risk is a determinant of the new capital requirements as from 2007. This has paved the way for the banking industry as frontrunner in model development for operational risk assessment, and several large European banks are allotted with AMA qualification. Reflecting different levels of sophistication and risk sensitivity, AMA is the most flexible approach for operational risk quantification and if banks accomplish detailed requirements, AMA recognizes internal systems. At present, more free hands are given to insurers in their ambition towards a Solvency II implementation in 2012. The Financial Services Authority (FSA) will judge the accuracy and sophistication of the proposed internal model by the insurer, as a result, a penalty in form of ’capital add-on’ could be executed on the company solvency capital if FSA notices weaknesses in the methodology. As with banks, the Directive for insurers will converge to a set of quantitative and qualitative standards that should be fulfilled, instead of prescribing a particular type of model. The classical risk measure is Value-at-Risk (VaR) at a specific confidence level with a holding period of one year. The FSA requires that the measurement of operational risk should be integrated in the daily operations. Also the measurement should be independent of risk control functions. Unfortunately, one of the major problems with modelling
operational risk is the lack of suitable loss data. To overcome this problem, the general perception in the industry is that a company needs to complement its internal loss data with some more information.

The grounds for loss distribution modelling can be found in Klugman, Panjer and Willmot (1998), McNeil, Frey and Embrechts (2005), Cizek, Härdle and Weron (2005) and Panjer (2006). The past few years have provided the literature with alternative models on how one can incorporate external information into internal loss distribution modelling. The papers by Shevchenko and Wüthrich (2006), Bühlmann, Shevchenko and Wüthrich (2007) and Lambrigger, Shevchenko and Wüthrich (2007) make use of Bayesian inference to combine several sources of data. Dahen and Dionne (2007) present methods to scale severities and frequencies in the context of banks, because evidence has been found of enormous heterogeneity due to diversity in size of the firm and location operations. Figini, Guidici, Uberti and Sanyal (2008) develop a method to mix data by applying a truncated external data for internal loss distribution modelling. Another procedure was proposed in Wei (2007), where Bayesian credibility is used to combine external and internal data. Gustafsson and Nielsen (2008), and the extended version by Gustafsson (2008), develops a systematic approach that incorporates external information into internal loss distribution modelling. The standard statistical model resembles Bayesian methodology and credibility theory in the sense that prior knowledge (external data) has more weight when internal data is scarce than when internal data is abundant.

Many different issues arise in practice when insurers analyse their operational risk data. We will focus on three of them:

i) Internal information is scarce.

ii) External data can be shared by all insurers under the auspices of a consortium, but pooling the data requires a mixing model that combines internal and external information.

iii) Data used to assess operational risk are sometimes underreported, which means that losses do happen, but either they are unknown or hidden. Data on operational losses are sometimes unrecorded below a given threshold due to existing habits or organizational reasons, which
means that information on losses below a given level is not collected. One should not confuse underreporting with unrecording, because the latter is a volunteer action not to collect all those losses below a given level.

When estimating an internal operational risk loss distribution a major obstacle is that not all losses are observed. Of course, the prior knowledge incorporated will ensure exposure is present in some intervals, while these losses are in general large in size, some intervals will be underreported in the loss distribution modelling. To estimate such an underreporting function from the internal data (and/or the external data) requires a complicated mathematical procedure. A simplified approach was introduced by Guillen, Gustafsson, Nielsen and Pritchard (2007) and an extended version by Buch-Kromann, Englund, Gustafsson, Nielsen and Thuring (2007). They developed an underreporting function in a way that the theoretical problem is simplified and they proposed a solution that is closely related to what one would have gotten if all internal losses would have been observed without underreporting.

In this article, we address the quantification of operational risk when accounting for the three problems indicated above. In fact, parts of the methodology have already been proposed before for every different problem independently. Here we want to present a unified approach that can cope with all of them at the same time. The most difficult part is trying to find ways to combine information from several sources (i.e. internal and external data), each one coming from institutions having different collection thresholds and possibly different reporting behaviours. Our contribution is to present a model that corrects for underreporting and combines internal and external data. We merge the theory developed by Gustafsson and Nielsen (2008) with the two underreporting papers Guillen et al. (2007) and Buch-Kromann et al. (2007), and generalize the methodology to incorporating collection thresholds as well. We proceed as follows: Section 2 lays out data availability and discusses the issues a company could visualize when sophisticated methodologies should be utilized for operational risk assessment. In Section 3 the underreporting model is presented. Section 4 presents the model by Gustafsson and Nielsen (2008), i.e. when two sources of data can be utilized and an unrecording problem exists. The proposed model is extended to allow for underreporting both for the internal and the external data. In Section 5, an application on real operational risk data is provided that illustrates
the interpretation and usefulness of the proposed estimation framework. For each model the total loss distributions with holding period of one year for specific risk tolerance levels is estimated and its quantiles are derived. We conclude and discuss future extensions that could improve operational risk capital assessment in the insurance industry if operational data are shared among companies.

2 Data Availability

In general, the collection period of internal operational risk losses is very short for most insurers, which itself generates a very scarce sample to estimate and validate on. Therefore, nobody doubts that complementing internal data with more abundant external data is desirable. A suitable framework to overcome the lack of data is getting data from a consortium. Individual insurers come together to form a consortium, where the population from which the data is drawn is assumed to be similar for every member of the group.

Let \((X_i)_{1 \leq i \leq M_I}\) be occurred internal losses from a specific event risk category (e.g. internal fraud, system failures, etc.) with total occurred number of losses denoted \(M_I\) over some time period. Let \((Y_j)_{1 \leq j \leq M_E}\) be occurred external losses (e.g. consortium data or publicly reported losses) from the same event risk category over a similar time period.

To illustrate our presentation we assume that the internal and external losses are lognormally distributed with the same location parameter and different scale parameter. More precisely, we have assumed \(X_i \sim \log N(0, 3/2)\) and \(Y_j \sim \log N(0, 5/2)\). In our example, we let \(M_I = 50\) and \(M_E = 500\), and usually the relation \(M_I < M_E\) holds. The assumption on similar location parameter between the internal and external data sources follows the assumption that similarity exists for every member of the consortium. We simulated occurred losses, and in our simulated sample the maximum internal occurred loss is 96 and the maximum loss for the external sample turned out to be 919.

\(^1\)We know that this can be argued, because not all insurers are the same and one can raise questions about whether consortium data can be considered to have been generated by the same process, or whether pooled data do reflect the size of each insurer transaction volume. We do not discuss these issues explicitly in this paper and we assume no scaling is needed.
Figure 1 shows the simulated data sets for the internal and external distribution. Initially we will consider all occurred losses, but we have used different symbols to identify the cases that will not be included in the truncated sample and later in the underreporting sample. All losses below the threshold (denoted by crosses) are not reported at all, so they will disappear in the truncated sample. The underreported losses over the collection threshold are marked with grey circles. There are 4 internal losses and 14 external losses that are not reported, even though their size exceeds the collection threshold. Our sample with underreporting will exclude those cases that are not reported and will work only with the reported losses (black circles), which are losses above the collection threshold.

*** Figure 1 About Here ***

2.1 Minimum Collection Threshold

Companies usually fix a minimum threshold level so that all losses below this level are not reported. Define the internal collection threshold by \( H_I \) and let \( H_E \) be the external collection threshold\(^2\). Then, we define a truncated sample by \((X^I_i)_{1 \leq i \leq M^I_i} = (X_i \mid X_i \geq H_I)_{1 \leq i \leq M^I_i}\) as occurred internal losses above the minimum collection threshold \( H_I \), and \((Y^I_j)_{1 \leq j \leq M^E_j} = (Y_j \mid Y_j \geq H_E)_{1 \leq i \leq M^E_j}\) as occurred external losses above the minimum collection threshold \( H_E \). Here \( M^I_i \) and \( M^E_E \) are the total number of occurred losses larger than the internal and external thresholds, respectively.

In our illustration we assume that the internal collection threshold \( H_I = 1 \) and the external collection threshold \( H_E = 10 \). Then, the occurred total number of losses above the thresholds turns out to be \( M^I_i = 21 \) and \( M^E_E = 89 \).

\(^2\)We assume that consortium data refer to losses above a given \( H_E \) level.
2.2 Modelling Each Data Set Separately

If all occurred losses \( (X_i)_{1\leq i\leq M_I} \) and \( (Y_j)_{1\leq j\leq M_E} \) were available and a lognormal distribution was assumed for the operational risk size (severity) distribution, the target density would be:

\[
h(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log(x) - \mu)^2}{2\sigma^2}}, \mu \in \mathbb{R}, \sigma > 0, x \geq 0.
\]  

(2.1)

The location and the scale parameters \( \{\mu, \sigma\} \) in the lognormal density would easily be estimated on each data set using maximum likelihood.

When only data above a given threshold are observed, the truncated lognormal density should be used (Johnson, Kotz and Balkrishan (1994)). Its density is defined depending on the truncation level \( H_I \) or \( H_E \) as:

\[
  f_I(x) = \frac{h_I(x)}{\int_{H_I}^{\infty} h_I(w)dw} \text{ for } x \geq H_I \quad \text{and} \quad f_E(x) = \frac{h_E(x)}{\int_{H_E}^{\infty} h_E(w)dw} \text{ for } x \geq H_E,
\]  

(2.2)

where \( h_I(\cdot) \) and \( h_E(\cdot) \) correspond to (2.1) for the internal and the external severity distribution respectively. The parameters for each truncated density \( f_I(\cdot) \) and \( f_E(\cdot) \) are estimated by a conditional maximum likelihood estimation.

The first two rows in Table 1 present the quantiles with return periods 1 in 20 years loss (95%), 1 in 100 years loss (99%) and 1 in 200 years loss (99.5%), where \( \hat{H}_d(\cdot) \) and \( \hat{F}_d(\cdot) \), with \( d = I \) or \( E \), denotes estimated quantile values for the lognormal and the truncated lognormal model, originating from models (2.1) and (2.2) when the original and the truncated data are used for estimation, respectively. The large values of the external truncated distribution can be explained by the large probability mass below the threshold \( H_E \), so it shows a big truncation effect, and possibly some instability due to a small sample size. Column 2 and 7 presents the total number of losses utilized to estimate the parameters presented in column 3 and 8.
<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
<th>${\hat{\mu}, \hat{\sigma}}$</th>
<th>Internal Data</th>
<th>External Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$q = 0.95$</td>
<td>$q = 0.99$</td>
<td>$q = 0.995$</td>
</tr>
<tr>
<td>$H_d^-(x_q)$</td>
<td>50</td>
<td>{0, 1.5}</td>
<td>11.79</td>
<td>32.77</td>
</tr>
<tr>
<td>$\hat{F}_d^-(x_q)$</td>
<td>21</td>
<td>{-0.46, 2.15}</td>
<td>50.46</td>
<td>183.74</td>
</tr>
<tr>
<td>$\hat{F}_d^+(x_q)$</td>
<td>17</td>
<td>{-0.34, 2.26}</td>
<td>44.59</td>
<td>172.84</td>
</tr>
<tr>
<td>$\hat{G}_d^+(x_q)$</td>
<td>17</td>
<td>{1.45, 1.19}</td>
<td>19.53</td>
<td>48.75</td>
</tr>
</tbody>
</table>
3 Underreporting Losses

The concept of underreporting was introduced in the context of operational risk in the purely para-
metric case by Guillen et al. (2007) and an extended semi-parametric version by Buch-Kromann et al. (2007). These developments provided an analysis of the quantitative impact of the failure to report all operational risk losses. When estimating operational risk losses, we know that not all losses are reported. Losses below the collection period will not be recorded at all. Those losses above the collection threshold will mostly be recorded but sometimes will be lost, ignored or hidden. In this section we will correct risk measurement by introducing an assumption on how underreporting behaves. In other words, we can say that real operational risk data combine two types of sample selection problems: truncation below a given threshold corresponding to lost data and random censoring above the level.

Since the reporting process in an organization is central and underreporting operational risk events is also an operational risk, this should be incorporated in the model. In this paper we will not discuss the opposite to underreporting, i.e the concept of overreporting. Some may argue that the losses not included in the operational risk calculation will be captured somewhere else in the organization and included in the overall capital figure. However, either if an operational risk loss is reported somewhere else, or not reported at all, we believe that it is essential to adjust for underreporting to obtain a correct operational risk severity distribution, for a better risk management.

In the papers by Guillen et al. (2007) and Buch-Kromann et al. (2007) they argued that estimating such an underreporting function from the data itself is a complicated mathematical deconvolution problem, and that the rate of convergence of the deconvoluted estimators is often very poor. After extensive interviews with risk expert practioners and a qualitative decision process, they deduced a best guess of an underreporting function, but they did not consider a collection threshold. In this paper a collection threshold is also included and therefore we are only interested in the reporting probabilities above the collection thresholds $H_I$ and $H_E$.  


Let us define two indicator functions $I_I(\cdot)$ and $I_E(\cdot)$ as follows: $I_I(i) = 1$ if $X_i$ is reported and zero otherwise, and $I_E(j) = 1$ if $Y_j$ is reported and zero otherwise. Now, including the minimum collection thresholds, and considering only the truncated samples $(X^t_i)_{1 \leq i \leq M_I^t}$ and $(Y^t_j)_{1 \leq j \leq M_E^t}$, we obtain different indicator functions that only evaluate losses reported above these thresholds.

$$I_I^r(i) = \begin{cases} 1 & \text{if } X^t_i \text{ is reported and zero otherwise,} \\ 0 & \text{otherwise} \end{cases}$$

$$I_E^r(j) = \begin{cases} 1 & \text{if } Y^t_j \text{ is reported and zero otherwise.} \\ 0 & \text{otherwise} \end{cases}$$

If the collection thresholds $H_I$ and $H_E$ are taken into account, it follows that the number of reported losses above the thresholds equals:

$$N_I^r = \sum_{i=1}^{M_I^t} I_I^r(i), \text{ reported number of internal losses above } H_I,$$

$$N_E^r = \sum_{j=1}^{M_E^t} I_E^r(j), \text{ reported number of external losses above } H_E.$$

The number of available data is reduced even further when underreporting is taken into account. Let

$$(X^t_{ir})_{1 \leq i \leq N_I^r} = (X_i | X_i \geq H_I \cap I_I(i) = 1)_{1 \leq i \leq M_I^t} = (X^t_i | I_I^r(i) = 1)_{1 \leq i \leq M_I^t}$$

be internal losses over the minimum collection threshold $H_I$ that are reported, and let

$$(Y^t_{jr})_{1 \leq j \leq N_E^r} = (Y_j | Y_j \geq H_E \cap I_E(j) = 1)_{1 \leq j \leq M_E^t} = (Y^t_j | I_E^r(j) = 1)_{1 \leq j \leq M_E^t}$$

be external losses over the minimum collection threshold $H_E$ that are reported. In our example illustration, $\{N_I^r, N_E^r\} = \{17, 75\}$.

The third row in Table 1 shows what the quantile estimates denoted by $\hat{F}_{d}^{t-r}(\cdot)$, with $d = I$ or $E$, would be when using the underreported sample rather than the truncated one but no correction is introduced, so that underreporting is ignored.
3.1 Correct for Underreporting and Collection Threshold

If we continue to model each data set separately, one should correct each model for underreporting independently. For this we will need to define the underreporting functions by:

\[
    u_I(x) = P(I'_I(x) = 1), \quad x \geq H_I \\
    u_E(x) = P(I'_E(x) = 1), \quad x \geq H_E
\]

If \( g_I(\cdot) \) and \( g_E(\cdot) \) are the internal and external densities for the occurred losses, respectively, and by using the internal and external densities \( f_I(\cdot) \) and \( f_E(\cdot) \) for the reported losses above collection thresholds, the following relationship can be formulated:

\[
    g_I(x) = \frac{f'_I(x)/u_I(x)}{\int_{H_I}^{\infty} f'_I(w)/u_I(w)dw}, \quad x \geq H_I \quad \text{and} \quad g_E(x) = \frac{f'_E(x)/u_E(x)}{\int_{H_E}^{\infty} f'_E(w)/u_E(w)dw}, \quad x \geq H_E, \quad (3.1)
\]

where the denominator will serve as a normalizing factor. The probability of observing an internal and external operational risk loss above the respective threshold is given by

\[
    P_I = \int_{H_I}^{\infty} g_I(w)u_I(w)dw, \\
    P_E = \int_{H_E}^{\infty} g_E(w)u_E(w)dw. \quad (3.2)
\]

In practice it is quite hard, or sometimes impossible, to establish the external underreporting function. A solution to this would be to build one underreporting function that could be representative for both the internal and external exposure. Being member of a consortium, individuals come together to form this alliance, where members of the consortium are assumed to have similar behaviours. Therefore, assuming that there exists one underreporting function covering both sources intuitively makes sense.

A possible underreporting function for the internal and external exposure will then be defined as:

\[
    u(x) = \begin{cases} 
    u_I(x) & \text{if } H_I \leq x < H_E \\
    u_E(x) & \text{if } x \geq H_E 
    \end{cases}
\]
In order to estimate this function we couple the method proposed by Guillen et al. (2007) and Buch-Kromann et al. (2007), where numerical interpolation and extrapolation is used.

### 3.2 Including Underreporting in the Estimation Procedure

Figure 2 presents the estimated underreporting function for our example. Here the circles represent expert’s opinion on the likelihood of reporting losses of a particular size and the vertical grey lines represent the thresholds, the dashed line represent the values for the internal underreporting function, while the solid black line represent the common part.

*** Figure 2 About Here ***

With the underreporting function together with $f_I(\cdot)$ and $f_E(\cdot)$ available, the original densities $g_I(\cdot)$ and $g_E(\cdot)$ can be determined using maximum likelihood estimation, and the results are shown in Figure 3.

*** Figure 3 About Here ***

Interpreting Figure 3, the three top figures represent the estimated internal densities and the bottom row the external estimated densities. The domain has been seperated in three intervals to facilitate the graphical display. Also presented in Figure 3 are the losses and the collection thresholds for each data set. The black solid curves represent the lognormal density (2.1), estimated on the whole samples. The black and grey dashed curves represent the truncated lognormal density (2.2), where the black dashed curves represent the densities estimated on the truncated sample, i.e. occurred losses above the thresholds, $\{M^I_t, M^E_t\} = \{21, 89\}$. The grey dashed curves are estimated on the underreported samples with total number of reported losses $\{N^I_t, N^E_t\} = \{17, 75\}$. Here we establish that the internal density becomes higher close to the threshold, while the tail becomes slightly lighter compared to the black dashed curve. For the external graphs the same appearance is found on the grey dashed curve compared to the black dashed curve. However, since there exists more losses and the underreported losses are spread in size, the differences between the curves are not too large as in the internal situation. The grey solid curves corrects for underreporting and are defined by (3.1). Interpreting the outcomes one could see that adjusting for internal underreporting will provide a lighter tail for this
particular simulated sample, while the opposite is found for the external data.

With the estimated densities in (3.1), the probability of observing an operational risk event can be calculated using $P_I$ and $P_E$. From (3.2) we obtain that $\hat{P}_I \approx 44\%$ and $\hat{P}_E \approx 96\%$. These estimated values seem reasonable, if we take the mean of the reported expert opinion’s we obtain 68% on the support $[H_I, \infty)$ and 93% on the support $[H_E, \infty)$. Of course, the differences 44% to 68% and 96% to 93% depends on the data availability when estimating the densities in (3.1). The upper quantiles are presented in Table 1. The last row presents the models correcting for underreporting, represented by $\hat{G}_d^{-}(\cdot)$ with $d = I$ or $E$.

4 A Mixing Model in a Truncation Framework

The general perception to a more sophisticated approach is that a company should combine internal and external data for internal loss distribution modelling. Here we generalize the idea described by Gustafsson and Nielsen (2008) and Gustafsson (2008) to take into account both minimum collection threshold as well as underreporting. The mixing model should utilize the external data as prior knowledge. Then, a non-parametric smoothing technique adjustment should be applied on the prior knowledge according to the internal data availability.

In our example, the mixing model is applied as a first step. Let us assume all occurred internal losses $(X_i)_{1 \leq i \leq M_I}$ were available. Then they could be transformed to bounded support by:

$$\left( \hat{H}_E(X_i) \right)_{1 \leq i \leq M_I} \in [0, 1],$$

where $\hat{H}_E(\cdot)$ denotes a lognormal cumulative distribution function, and its parameters have been estimated with the sample of occurred external losses (assuming all of them were also available). The mixing methodology continues by bring into play kernel density estimation on the bounded support, thereby a correction function will be estimated that depends on the internal data characteristics and its discrepancy versus prior knowledge. The underlying parametric transformation is guided by a local
constant kernel density estimator presented by:

\[ \hat{k}(\hat{H}_E(x)) = \frac{1}{n \cdot \alpha_01(\hat{H}_E(x), h)} \sum_{i=1}^{n} \mathcal{K}_h \left( \hat{H}_E(X_i) - \hat{H}_E(x) \right), \hat{H}_E(x) \in [0, 1], \]

with transformed losses \( \hat{H}_E(X_i) \), number of observations \( M_I = n \), symmetric Epanechnikov kernel \( \mathcal{K}_h \) with \( K_h(\cdot) = (1/h)K(\cdot/h) \), Silverman’s rule-of-thumb bandwidth \( h \) estimated on the internal sample, and

\[ \alpha_{ij}(u, h) = \int_{\max(-1,(u-1)/h)}^{\min(1,u/h)} (vh)^j \mathcal{K}(v)^j dv. \]

For a review of modern kernel smoothing techniques see Wand and Jones (1995), and for regression functions, see Fan and Gijbels (1996). The final step in the transformation process is to backtransform to the original axis. This is done by transforming the smoothed distribution \( \hat{k}(\cdot) \) of the transformed data with \( \hat{H}_E^{-1}(\cdot) \). The transformation methodology can be summarized into one explicit expression:

\[ \hat{h}_m(x) = \frac{\hat{h}_E(x)}{n \cdot \alpha_01(\hat{H}_E(x), h)} \sum_{i=1}^{n} \mathcal{K}_h \left( \hat{H}_E(X_i) - \hat{H}_E(x) \right) \]

\[ = \hat{h}_E(x) \cdot \hat{k}(\hat{H}_E(x)), x \in [0, \infty), \quad (4.1) \]

where subindex \( m \) indicates a mixing model. A deeper analysis of the transformation methodology is found in Gustafsson and Nielsen (2008).

When incorporating the thresholds on occurred data, i.e utilizing the data \((X_i^t)_{1 \leq i \leq M_I} \) and \((Y_j^t)_{1 \leq j \leq M_E} \), some adjustment is needed in the transformation process. In practice, as we should see in the application section, if \( H_I \) and \( H_E \) are equal, there would not be a transformation problem. However, in our simulated example \( H_I \neq H_E \), therefore the transformation function will be defined on the support \([H_E, \infty)\), while the occurred internal data \((X_i^t)_{1 \leq i \leq M_I} \) that should be incorporated in the transformation function are defined on the support \([H_I, \infty)\). Therefore, the losses in the support \([H_I, H_E]\) cannot be transformed correctly since the transformation function is not defined on this support. The procedure to solve this issue draws on extrapolation from the lower limit \((H_I)\) to the upper threshold (boundary point \( H_E \)).
Let us assume \( h_E(\cdot) \) is estimated using external data, then we will extrapolate and normalize to the interval \([H_I, \infty)\) to solve the transformation problem. By defining the extrapolated normalized external density as:

\[
\hat{f}^*_E(x) = \frac{\hat{h}_E(x)}{\int_{H_I}^\infty \hat{h}_E(w)dw}, \quad x \geq H_I, \tag{4.2}
\]

one could determine a truncated cumulative distribution function \( \hat{F}^*_E(\cdot) \) that should transform the internal occurred losses \((X^t_i)_{1 \leq i \leq M^t_I}\) to bounded support. That is, \( \forall i, \)

\[
\left( \hat{F}^*_E(X^t_i) \right)_{1 \leq i \leq M^t_I} \in [\hat{F}^*_E(H_I), 1]
\]

where \( \hat{F}^*_E(H_I) \) represent the estimated internal threshold on the transformed scale. The truncated mixing semi-parametric model will then follow the expression:

\[
\hat{f}_m(x) = \frac{\hat{f}_E(x)}{n \cdot \alpha_{01}(\hat{F}^*_E(x), h)} \sum_{i=1}^n K_h \left( \hat{F}^*_E(X^t_i) - \hat{F}^*_E(x) \right)
= \hat{f}_E(x) \cdot \hat{k}(\hat{F}^*_E(x)), \quad x \in [H_I, \infty), \tag{4.3}
\]

with \( M^t_I = n \). If collection thresholds were identical, \( H_I = H_E \), the distribution \( \hat{F}^*_E(\cdot) \) could be replaced with \( \hat{F}^*_E(\cdot) \) in the expression above.

The final step is to correct for underreporting. Following the same argumentation as above, we evaluate the data set \((X^t_ir)_{1 \leq i \leq N^t_I}\) instead of \((X^t_i)_{1 \leq i \leq M^t_I}\) in the bounded support, guided by the transformation function \( \hat{F}^{r,s}_E(\cdot) \), obtained by following a procedure that is similar to (4.2).

Hence, the final model becomes

\[
\hat{f}^*_r(x) = \frac{\hat{f}^{r,s}_E(x)}{n \cdot \alpha_{01}(\hat{F}^{r,s}_E(x), h)} \sum_{i=1}^n K_h \left( \hat{F}^{r,s}_E(X^t_ir) - \hat{F}^{r,s}_E(x) \right)
= \hat{f}^{r,s}_E(x) \cdot \hat{k}(\hat{F}^{r,s}_E(x)), \quad x \in [H_I, \infty), \tag{4.4}
\]

15
with \( N^r_I = n \) and where the kernel density estimation function is defined on \( \hat{k}(\cdot) \in [\hat{F}^r_E(H_I), 1] \). In Table 2, the three different mixing models (4.1), for the whole sample of occurred losses, (4.3) with the truncated sample and (4.4) with the truncated and underreported sample, are compared in the first three rows for different quantile values, denoted by \( \hat{H}^{-}_m(\cdot) \), \( \hat{F}^{-}_m(\cdot) \) and \( \hat{F}^{r^{-}}_m(\cdot) \).

**TABLE 2**

Quantile values for several models based on pooled simulated data

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
<th>( q = 0.95 )</th>
<th>( q = 0.99 )</th>
<th>( q = 0.995 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{H}^{-}_m(x_q) )</td>
<td>550</td>
<td>36.26</td>
<td>213.46</td>
<td>410.50</td>
</tr>
<tr>
<td>( \hat{F}^{-}_m(x_q) )</td>
<td>110</td>
<td>115.54</td>
<td>485.77</td>
<td>830.36</td>
</tr>
<tr>
<td>( \hat{F}^{r^{-}}_m(x_q) )</td>
<td>92</td>
<td>96.27</td>
<td>472.95</td>
<td>873.05</td>
</tr>
<tr>
<td>( \hat{T}^{-}_m(x_q) )</td>
<td>92</td>
<td>65.28</td>
<td>279.43</td>
<td>304.75</td>
</tr>
</tbody>
</table>

To this point in the paper, the two methodologies, correction for underreporting and mixing internal and external data, have been considered separately. In the remaining of this section we determine a model that incorporates both these adjustments. The procedure begins by seeking the transformation function. This transformation function should correct for underreporting as well as using the internal threshold \( H_I \) as lower bound. We estimate a normalized density extrapolated for the extended support \([H_I, \infty)\) as:

\[
\hat{g}^*_E(x) = \frac{\hat{g}_E(x)}{\int^{\infty}_{H_I} \hat{g}_E(w)dw}, \quad x \in [H_I, \infty).
\]  

(4.5)

Using \( \hat{G}^*_E(\cdot) \), the cumulative distribution function derived from (4.5), we transform the losses \((X^{t,r}_i)_{1 \leq i \leq N^r_I}\) to bounded support. The final semi-parametric estimator is then

\[
\hat{t}_m(x) = \frac{\hat{g}^*_E(x)}{n \cdot \alpha_0(G^*_E(x), h)} \sum_{i=1}^{n} \hat{K}_h \left( \hat{G}^*_E(X^{t,r}_i) - \hat{G}^*_E(x) \right)
\]

(4.6)

\[
= \hat{g}^*_E(x) \cdot \hat{k}(\hat{G}^*_E(x)), \quad x \in [H_I, \infty),
\]

with \( N^r_I = n \) and where the kernel density estimation function \( \hat{k}(\cdot) \) is defined on \([\hat{G}^*_E(H_I), 1]\). The quantiles following from model (4.6) with the underreported sample are presented as the last row of
Table 2.

Interesting to see is the outcome of model (4.6) in the tail. Focus on the quantile 99.5%, the proposed model $\hat{T}_m(x_{\epsilon})$ shows the smallest value among the four mixing models. However, this is not the case for the lower quantiles 95% and 99%. The proposed model presents larger values on all quantiles compared to the pure internal models and smaller values compared to the pure external models.

*** Figure 4 About Here ***

Figure 4 plots the simulated severity distributions with the models that can be used when pooling together internal and external data. The grey and dashed grey line correspond to modelling internal and external data separately. They are plotted as a reference.

5 Operational Risk Application

This Section shows how the choice of methodology can significantly affect the estimation of required risk based capital. We perform an LDA on real operational losses and calculate the risk measures Value-at-Risk (VaR) and Tail-Value-at-Risk (TVaR) for different levels of risk tolerance $\alpha$. The operational loss data follows the event risk category Internal Fraud and the internal losses is defined by $(X_{i}^{r,n})_{1\leq i\leq N_{I}^{r}}$ and the external data $(Y_{j}^{r,n})_{1\leq j\leq N_{E}^{r}}$ is taken from a consortium. The collection threshold for the two samples are the same with the value $H_{I} = H_{E} = H = 10.000 £$. A summary statistic of the operational losses are presented in Table 3.

<table>
<thead>
<tr>
<th>Statistics for Event Risk Category Internal Fraud (£M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reported Losses ($n$)</td>
</tr>
<tr>
<td>$(X_{i}^{r,n})<em>{1\leq i\leq N</em>{I}^{r}}$</td>
</tr>
<tr>
<td>$(Y_{j}^{r,n})<em>{1\leq j\leq N</em>{E}^{r}}$</td>
</tr>
</tbody>
</table>
In total, eight models will be compared and evaluated. The developed model \((4.6)\), the truncated mixing model \((4.3)\), internal and external underreporting models \((3.1)\), truncated internal and truncated external models such as \((2.2)\) and pure parametric model without any adjustments. The underlying distribution is important both in the underreporting procedure as well as in the mixing methodology. In the recent years, new and more advanced parametric distributions for operational risk have emerged, with more parameters and more advanced methods of estimation. These more advanced parametric distributions present a better goodness of fit; however using such estimators might introduce high estimation variance. Mentionable ones are the \(\alpha\)-stable distribution and the \(g\) and \(h\) distribution. The \(\alpha\)-stable distribution is investigated in Giacometti, Rachev, Chernobai and Bertocchi (2008) where they test the distribution by transforming the data to better fit the parameter estimation. They show that in order to get a good fit of the operational risk data they are forced to model the body and the tail separately when using the \(\alpha\)-stable distribution. We find this troubling since we are interested in a parametric distribution that fits the entire data. There is another problem with the \(\alpha\)-stable distribution and it is that closed form densities do not exist for most cases. Hence, parameters estimations based on maximum likelihood becomes complicated, a strong reason for practitioners not to chose this distribution. The \(g\) and \(h\) distribution, on the other hand, has an appealing link to the common extreme value theory (EVT) methodology which is addressed in Degen, Embrechts and Lambringer (2007). However, the \(g\) and \(h\) distribution has a flaw in its slow convergence to the generalized Pareto distribution. Using it in connection with EVT may lead to inaccurate capital estimations, in particular for small levels of risk tolerance, which is shown in Dutta and Perry (2006). Instead we use the rather new Generalised Champernowne Distribution (GCD). The GCD has the same appealing property as the \(g\) and \(h\) distribution with its link to EVT. However, the tail’s convergence to the heavy tailed Pareto distribution is faster than the \(g\) and \(h\) distribution, which favours the GCD. Also a large flexibility is imbued in the distribution, see e.g Gustafsson, Pritchard and Thuring (2008). The prior density GCD that will be used in this study takes the form

\[
h(x; \alpha, M, c) = \frac{\alpha (x + c)^{\alpha - 1} ((M + c)^\alpha - c^\alpha)}{(x + c)^\alpha + (M + c)^\alpha - 2c^\alpha)^2}, \forall x \in \mathbb{R}_+, \alpha > 0, M > 0, c \geq 0. \quad (5.1)
\]
This density will then be estimated on both samples, used separate as models and incorporated both in the underreporting models as well as the mixing methodology. The underreporting function used in this study is estimated by expert’s opinion and described by Figure 5.

*** Figure 5 About Here ***

For the sake of simplicity, we introduce the abbreviations \( \hat{F}_1, \hat{F}_2, \ldots, \hat{F}_8 \) for the estimated severity models considered in the analysis, described in Table 4. The severity models \( \hat{F}_1 \) and \( \hat{F}_2 \) are pure parametric models estimated on the internal and external data respectively, using the parametric choice (5.1). The two models \( \hat{F}_3 \) and \( \hat{F}_4 \) are pure parametric as well, however they are defined by a truncated version of (5.1) outlined by formula (2.2). Further, the models \( \hat{F}_5 \) and \( \hat{F}_6 \) takes its offspring from the two previous models, but with the differences that the underreporting function defined by Figure 5 is taken into account by using formula (3.1). Model \( \hat{F}_7 \) is the first mixing model considered in the study with the prior start (5.1) and defined through equation (4.3). The final model, \( \hat{F}_8 \), is an extended version of \( \hat{F}_7 \) by also including underreporting in the estimation.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Internal Data</th>
<th>External Data</th>
<th>Pooled Internal and External Data</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{F}_1 )</td>
<td>(5.1)</td>
<td>·</td>
<td>·</td>
<td>Pure parametric</td>
</tr>
<tr>
<td>( \hat{F}_2 )</td>
<td>·</td>
<td>(5.1)</td>
<td>·</td>
<td>Pure parametric</td>
</tr>
<tr>
<td>( \hat{F}_3 )</td>
<td>(2.2)</td>
<td>·</td>
<td>·</td>
<td>Trunc. parametric</td>
</tr>
<tr>
<td>( \hat{F}_4 )</td>
<td>·</td>
<td>(2.2)</td>
<td>·</td>
<td>Trunc. parametric</td>
</tr>
<tr>
<td>( \hat{F}_5 )</td>
<td>(3.1)</td>
<td>·</td>
<td>·</td>
<td>Trunc. und.rep. parametric</td>
</tr>
<tr>
<td>( \hat{F}_6 )</td>
<td>·</td>
<td>(3.1)</td>
<td>·</td>
<td>Trunc. und.rep. parametric</td>
</tr>
<tr>
<td>( \hat{F}_7 )</td>
<td>·</td>
<td>·</td>
<td>(4.3)</td>
<td>Trunc. mixing model</td>
</tr>
<tr>
<td>( \hat{F}_8 )</td>
<td>·</td>
<td>·</td>
<td>(4.6)</td>
<td>Trunc. und.rep. mixing model</td>
</tr>
</tbody>
</table>

Three different estimation procedures are needed as a first step to estimate the eight models. The parameter \( M \) in the distribution (5.1) is estimated by the empirical median in all three situations.
Hence, $H(M; \alpha, M, c) = 1/2$, thereby indicates that the empirical median is an appropriate estimation choice. Also Lehmann (1991) pointed out that the median is a robust estimator especially for heavy tail distribution. The other parameters $\{\alpha, c\}$ are then obtained by maximizing the log-likelihood function corresponding to each model.

We are going to simulate the frequency of events. In the Monte Carlo simulation $N_f^i$ is assumed to follow an homogenous Poisson process with intensity $\lambda > 0$ and defined by

$$P(N_f^i = n) = e^{-\lambda} \left(\frac{(\lambda t)^n}{n!}\right).$$

We estimate the intensity for each model and use maximum likelihood. So, $\hat{\lambda} = n/T$, where $n$ is the reported number of losses and $T$ is the corresponding collection period presented in Table 3. This frequency estimation is then incorporated in five models, the ones without correcting for underreporting. Model 5, 6 and 8 correct for underreporting, hence the frequency needs to be corrected as well as the severity side. With underreporting, we assume that the reported losses $N_f^i$ follows a Poisson distribution with intensity equal to $\lambda P$, where $P$ is the probability of observing an operational risk loss, defined by (3.2). When accounting for underreporting we focus on $M_f^i$, the occurred internal losses above the threshold $H$. We assume $M_f^i$ follows a Poisson distribution with intensity $\lambda$. The correction is then obtained for each of the three underreporting models by dividing the estimated $\hat{\lambda}$ with the probability of observing a loss for each model. When we calculate these probabilities of observing an operational risk loss using (3.2), we get $\{\hat{P}_5, \hat{P}_6, \hat{P}_8\} = \{15.1\%, 28.6\%, 18.0\%\}$. Then, all the intensities used in this study will be $\{\hat{\lambda}, \hat{\lambda}_5, \hat{\lambda}_6, \hat{\lambda}_8\} = \{55, 366.7, 193.8, 308.3\}$.

The aggregated total loss $S_k$ is then defined as the stochastic process

$$S_k = \sum_{i=1}^{N_f^i} \hat{F}_k^{-}(u_i) \quad (5.2)$$

for severity model assumption $k$. Here $u_i$ is sampled from a uniform distribution guided by sampling the Poisson with intensity equal one of the four estimated models. Function $\hat{F}_k^{-}(\cdot)$ is the estimated
inverse function for one of the eight models described in Table 4. This procedure is then repeated 20,000 times for each model resulting in an aggregated empirical loss distribution with holding period one year. This is illustrated visually in Figure 6.

*** Figure 6 About Here ***

The developed model entitled 'A Mixing Truncated Compound Poisson Model That Corrects for Underreporting', corresponding to the eight model seems to place its outcome between the others. Interpreting the outcomes we find that $S_1$ and $S_5$ provide the lightest tails. It is interesting to see that $S_1$ becomes more heavy tailed when truncation is incorporated, see $S_3$, and lighter when adding underreporting, see $S_5$. The same effect, in some sense, holds when we study $S_2$, $S_4$ and $S_6$. The effect of adding underreporting to $S_7$, leads to obtain the proposed $S_8$. However, if we study the more extreme outcomes with the risk measures Value-at-Risk (VaR) and Tail-Value-at-Risk (TVaR) we find some effects.

In Table 5 some chosen risk tolerance levels are specified for the risk measures VaR and TVaR for each model assumption.
<table>
<thead>
<tr>
<th>Model</th>
<th>Value-at-Risk</th>
<th>Tail-Value-at-Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95%</td>
<td>99%</td>
</tr>
<tr>
<td>$S_1$</td>
<td>5.78</td>
<td>7.20</td>
</tr>
<tr>
<td>$S_2$</td>
<td>388.41</td>
<td>880.81</td>
</tr>
<tr>
<td>$S_3$</td>
<td>9.75</td>
<td>32.04</td>
</tr>
<tr>
<td>$S_4$</td>
<td>345.89</td>
<td>487.61</td>
</tr>
<tr>
<td>$S_5$</td>
<td>50.13</td>
<td>53.08</td>
</tr>
<tr>
<td>$S_6$</td>
<td>749.75</td>
<td>1004.13</td>
</tr>
<tr>
<td>$S_7$</td>
<td>44.01</td>
<td>91.25</td>
</tr>
<tr>
<td>$S_8$</td>
<td>160.27</td>
<td>229.45</td>
</tr>
</tbody>
</table>

6 Conclusions

This paper presented several systematic correction techniques, essential to quantify a company’s operational risk exposure. Starting with internal and external data sets, the importance of including collection thresholds was shown. Thereafter discussion and illustration on the difference between modelling observed or reported losses. The development continued to combine two data sources with a final result that both adjusted for collection threshold and underreporting.

The data application provided us with insight on the behaviour when adding more sophisticated corrections to the models. Starting from two parametric models $S_1$ and $S_2$, we see that the outcome on the risk measure VaR$^{99.9\%}$ was over 100 times larger value using the external data source instead of the internal data. By including truncation, in $S_3$ and $S_4$, the internal model provided a heavier tail while a lighter was found on the external data model. Adding underreporting as well, in $S_5$, placed result between models for $S_1$ and $S_3$, while $S_6$ became larger than both $S_2$ and $S_4$. By using the truncated mixing model, in $S_7$, the prior knowledge of the estimation brings a larger value than the
internal data models $S_1$, $S_3$ and $S_5$. However, we could see that the outcome is far away from the external data models. When taking underreporting into account, the outcome from the model in $S_8$ increases by 294% compared to $S_7$. 
References


Figure 1: Simulated internal and external loss data from a specific event risk category. The thresholds are represented with grey lines for the two samples, and assumed to be $H_I = 1$ and $H_E = 10$ for internal and external collection level respectively. The losses smaller than the thresholds are marked with crosses and the losses above the threshold that are not being reported are marked with grey circles.
Figure 2: The combined underreporting function.
Figure 3: Estimated lognormal densities on the samples \((X_i)_{1 \leq i \leq M_I}\) and \((Y_j)_{1 \leq j \leq M_E}\) represented by black curves, estimated truncated lognormal densities on the samples \((X_{t_i}^r)_{1 \leq i \leq M_I}^r\) and \((Y_{t_j}^r)_{1 \leq j \leq M_E}^r\) represented by dashed black curves, and estimated truncated lognormal densities on the samples \((X_{t_i}^r)_{1 \leq i \leq N_I}^r\) and \((Y_{t_j}^r)_{1 \leq j \leq N_E}^r\), represented by grey dashed curves. The densities represented by grey solid curves are estimated truncated lognormal densities on the samples \((X_{t_i}^r)_{1 \leq i \leq N_I}^r\) and \((Y_{t_j}^r)_{1 \leq j \leq N_E}^r\) with including underreporting in the estimation.
Figure 4: Model (4.6) is presented by a black curve, model (4.3) is represented by dashed black curve, and internal and external versions of model (3.1) are represented by grey- and dashed grey curve respectively.
Figure 5: Underreporting function for the event risk category: *Internal Fraud.*
Figure 6: The estimated total loss distributions $S_k$, $k = 1, \ldots, 8$, for the event risk category: Internal Fraud.