

# A scaling model for severity of operational losses using generalized additive models for location scale and shape (GAMLSS)

Amandha Ganegoda\*

*School of Actuarial Studies, Australian School of Business, Sydney, Australia*

John Evans

*School of Actuarial Studies, Australian School of Business, Sydney, Australia*

## Abstract

In this paper, we investigate the problem of how to combine operational losses collected from various banks of different sizes and loss reporting thresholds in order to estimate the distribution of operational loss severities for a bank of a given size. We model the severity of operational losses by using the extreme value theory to account for the reporting bias of the external data, and a regression analysis based on the GAMLSS framework to model the scaling properties of operational losses. In contrast to previous studies on the scaling problem, our analysis gives particular emphasis to the scaling properties of the tail of the loss distribution. Contrary to existing knowledge, we find that the size of a bank is an important determinant of the severity of operational losses and that the tail index of the distribution is negatively correlated with the size of the bank. The results indicate that for very large banks, distribution of the operational loss severity can be extremely heavy tailed (i.e. tail index less than 1), a finding which has significant implications for capital calculation as well as for risk management. Furthermore, we also demonstrate that the capital estimates provided by our model is consistent with the industry standards and the model can be used by individual banks to simulate data to complement their internal data.

## JEL classification

G21; G28; C30; C35; C38

## Keywords

Operational risk; Generalized Additive Models for Location Scale And Shape (GAMLSS); Scaling models; Prudential Regulation of Banks

## 1 Introduction

---

The minimum capital requirement for banks set by the first accord of the Basel Committee on Banking Supervision in 1988 primarily focused on the credit risk of banks (Basel, 1988). In 1997 the agreement was amended to incorporate market risk for capital calculations (Basel, 1996). Since then

\*Correspondence to: Amandha Ganegoda, School of Actuarial Studies, Australian School of Business, UNSW, Sydney NSW 2052, Australia. E-mail: amandha@unswalumni.com

much research has been done on measuring and managing market and credit risk for banks but very little attention has been given to other types of risk such as operational risk. However, large scale operational loss events such as the unauthorized trading at Barings Bank in the UK which led to its collapse in 1995, the \$2.2 billion loss at National Australia bank in 2001 due to product issues and model errors in their HomeSide mortgage servicing subsidiary, the \$2.2 billion settlement with investors of Enron in 2005 by JP Morgan Chase for their role in helping the failed energy company to provide misleading financial results, and more recently, the \$7.2 billion loss due to fraudulent trading at Société Générale have made both industry and regulators realise that merely focusing on credit and market risk is insufficient to ensure the solvency of a bank. It has been found that operational risk in a bank is often much larger than the market risk (De Fontnouvelle *et al.*, 2006). In response, the Basel committee included operational risk as one of the risks that need to be quantified in their Basel II accord in 2004 (Basel, 2006).

However, there are definite challenges for risk managers in quantifying operational risk. Firstly, there is no widely accepted methodology to quantify operational risk. Secondly, the heavy-tailed nature of the operational loss distributions makes it difficult to estimate the tails, especially when data is inadequate. Thirdly, most of the Low-frequency/high-severity (LF/HS) operational losses are high degree unique events, which makes it even more difficult to model the tail. Given these circumstances the amount of unknown and the unknowable is far greater than the known in the area of operational risk. As a result there are considerable uncertainties regarding the economic capital estimates for operational risk. For example, using bank-level data of six large internationally active banks, De Fontnouvelle *et al.* (2007) report a wide variation in the shape parameter (a low of 0.498 to a high of 0.859) of the fitted Generalized Pareto distribution (GPD) for the losses from each bank. In a comment on their paper Kuritzkes points out that this wide variation has a significant impact on the capital charge for individual banks. Using six hypothetical banks with identical exposures, Kuritzkes demonstrates that the capital charge can vary from \$208 mill (if the shape parameter is 0.498) to \$4,320 mill (if the shape parameter is 0.859). This significant variation raises the important question as to what extent banks and regulators can rely on bank level data to model the shape of the tail of the loss distribution. Given that banks are required to hold enough capital to cover one in a thousand year adverse event, it is obvious that it is impossible to observe the relevant operational loss data for capital estimation within any one firm. As a solution to this Basel II requires banks to use external data to complement their internal data. However, the process of combining internal data with external data is an exercise which requires due diligence due to biases inherent in the external data. If not properly combined, external data may distort parameter estimates and eventually lead to perverse estimates of economic capital (Baud *et al.*, 2002). Wilson (2007) points out there are three types of biases inherent to external data which may lead to poor estimates:

- i) Reporting bias – occurs when different thresholds have been used by institutions to report losses to the external database
- ii) Control bias – occurs when data is collected from institutions with different risk control systems
- iii) Scale bias – occurs when data is collected from institutions with different sizes

In this paper we demonstrate how an external database can be used to model the severity of operational losses by correcting for reporting bias and scale bias.

The severity model developed in this study makes use of a method based on Extreme value theory proposed by De Fontnouvelle *et al.* (2006) to account for the reporting bias of the external data, and

a regression analysis based on the GAMLSS framework introduced by Rigby & Stasinopoulos (2001, 2005) to model the scaling properties of operational losses. The model can simulate losses for a bank given the size of the bank and other explanatory variables that might affect the severity of operational losses. The model is calibrated using the industry data and therefore can be used by banks with very little data of their own or by banks whose operational environment has changed significantly due to a merger or an acquisition. The next section provides a brief review of scaling models reported in the literature. The section 3 formulates the model and provides details of the research methodology, section 4 presents the results of the model after fitting it to data from the US banking industry, section 5 provides a demonstration of the severity model at work by computing capital estimates for several hypothetical banks as well as for two real banks by making simplified assumptions about the frequency distribution, and section 6 concludes.

## 2 Literature Review

It is natural to hypothesise that the severity of operational losses will have a relationship to the size of the bank. To the best of our knowledge, the first to test this hypothesis for financial firms were Shih *et al.* (2000). They tested the correlation of the firm size indicators of revenue, assets, and the number of employees with the operational loss severities using the OpVar database (which is an external database consisting of publicly reported losses). They reported all three variables to be correlated with loss severity, with the revenue showing the strongest relationship. Furthermore, authors found that the relationship between the loss severity and the size of the firm is nonlinear. They reported that “a firm that is twice as large as another does not, on average, suffer a loss that is twice the size of the loss experienced by the other firm”. They hypothesised that the relationship between the loss severity and the size of a firm can be modelled as

$$\log(\text{Loss}_i) = \beta_0 + \beta_1 \log(\text{revenue}_i)$$

An ordinary least squares (OLS) regression of the model confirmed a linear relationship between the log-loss severity and the log-revenue. However, authors reported the presence of heteroskedacity in the error term and corrected for this by running a weighted least squares (WLS) regression. They reported an adjusted  $R^2$  value of 9% for their WLS regression and concluded that the size only accounts for around 5% of the variability in the loss severity. Furthermore, they reported that “there is clear evidence of diminishing relationship between the size of a firm and loss magnitude”. A major limitation of this study is that authors did not correct for reporting bias in the external data. Later, De Fontnouvelle *et al.* (2006) showed that reporting bias in the OpVar database is too significant to be ignored. Therefore the results of the study should be taken with a grain of salt since parameter estimates are likely to be distorted as the model has been fitted using a disproportionate number of large losses.

Another early study to consider the scaling problem is Na *et al.* (2006). Their study mainly analysed the scaling of aggregate losses as opposed to individual losses. Authors assumed aggregate operational losses can be broken down into two components, firstly a common component, which is assumed to be stochastic, capturing the influence on the loss severity by factors such as macroeconomic, geopolitical, cultural and business environment common to all banks and secondly an idiosyncratic component, which is deterministic, capturing the influence of the size of the bank on the loss severity. Their analysis revealed that the mean and the standard deviation of the aggregate losses scale similarly. Moreover, they found that the ratio of aggregate loss between two business lines (or banks) is simply equal to the ratio between their gross income. Authors also briefly

analyzed whether their scaling model can be applied to individual loss severities. However, they reported that the results did not turn out to be convincing.

A study which took a slightly different approach to analyze the scaling properties of operational losses is that of Cope & Labbi (2008). The authors used quantile regression techniques to analyse the relationship between different quantiles of the operational loss distribution and the size of a bank. Their study revealed that large losses scale differently from small losses and also confirmed a diminishing relationship between the loss size and the firm size for certain categories of operational losses. An important finding of their study is that not all operational losses are positively correlated with the size of a bank. They reported that for certain lines of business and event types, operational loss severity may decrease as the size of a bank increases. A limitation of the study is that in order to apply the quantile regression techniques proposed by the authors one needs to be able to identify the corresponding percentile level of each loss. However, it is not possible to achieve this when data is collected from more than one bank subject to a reporting threshold. To overcome this issue authors proposed a novel quantile matching algorithm. The algorithm requires the loss reporting threshold to be known and fixed. For some databases such as the ORX database, which the authors used in their study, this is not a problem. However, for most other external databases the loss reporting threshold of individual banks is unknown. In such instances one would require different techniques or a modification to the authors' quantile matching algorithm.

A more recent paper to look at the scaling problem is by Dahan & Dionne (2010). They extended the work of Shih *et al.* (2000) by introducing scaling factors other than the firm size into the regression analysis. Also, in contrast to Shih *et al.* (2000) the authors used the Heckman method to correct for selection bias in the external data. The form of the regression model they employed is

$$\log(\text{Loss}_i) = \beta_0 + \beta_1 \log(\text{assets}_i) + \sum_{\text{all } j} \beta_j \text{BL}_{ij} + \sum_{\text{all } k} \beta_k \text{ET}_{ik} + e_i$$

where,  $\text{BL}_{ij}$  is a business line indicator,  $\text{ET}_{ik}$  is an operational loss event type indicator,  $e_i$  is an error term assumed to be normal, and  $\beta$ s are the parameters to be estimated. The model was fitted using OLS and an adjusted  $R^2$  value of 29.58% was reported as an overall fit to the data. Although the value of  $R^2$  is low, it was a significant improvement on the  $R^2$  value reported by Shih *et al.* (2000). However, given the low value of  $R^2$  it is reasonable to question the appropriateness of the distributional assumption made by the authors that log-losses are Normally distributed. If the log-losses are distributed as a Normal distribution, then the losses should follow a Lognormal distribution. Although Lognormal distribution is a fairly heavy-tailed distribution, it is not as heavy-tailed as Pareto or Log-gamma distribution. Various previous studies have found evidence that operational losses are (extremely) heavy-tailed (see e.g., De Fontnouvelle *et al.*, 2007; Moscadelli, 2004). The low value of the adjusted  $R^2$  found by Dahan & Dionne (2010) could be possibly explained by the inappropriateness of the distributional assumption.

Another interesting study of the scaling problem is found in Wei (2007) where he tested the suitability of using the peak over threshold (POT) method from extreme value theory over the traditional actuarial techniques to model severity of losses. The author reported that the capital charge based on the Generalized Pareto distribution from the POT method was too conservative and too costly for the banks, whereas traditional actuarial models which assume losses arise from a known parametric family of distribution with scale parameter related to bank assets, provided a much more reasonable yet conservative capital estimates.

As discussed so far most of the previous work has focused on how the mean of the log-loss severity may scale in response to the size of a bank. However, there are several issues in this approach. Firstly, mean of the log-losses does not relate to any meaningful statistic of the (raw) loss distribution. Hence, results are difficult to interpret in most circumstances. Secondly, the operational risk capital charge for a bank is determined by the 99.9<sup>th</sup> percentile of the loss distribution. The information provided by a scaling model fitted to the mean of the log-losses provides very little information on how the 99<sup>th</sup> percentile may scale. To illustrate this point consider two banks denoted as Bank I and Bank II with assets of \$100 bill and \$200 bill respectively. Let's assume that the operational loss distribution of both banks follows a Log-gamma distribution with parameter values for Bank I being  $\alpha_I = 6.25$ , and  $p_I = 25$ , and parameter values for Bank II being  $\alpha_{II} = 1.3889$ , and  $p_{II} = 2.778$  (the density function and a discussion of Log-gamma distribution will be given later in subsection 4.3.1). The mean and the 99.9<sup>th</sup> percentile loss for each bank under these distributional assumptions are given in table 1. One can observe from the table that the mean operational loss (as well as the mean log-loss) at the larger bank is less than half the mean loss at the smaller bank. However, the 99.9<sup>th</sup> percentile loss in the large bank is more than twice the loss in the small bank. In this particular example, the mean of the log-losses (as well as the mean of the raw-losses) and the 99.9<sup>th</sup> percentile of the log-losses scale in the opposite direction. In this situation OLS regression techniques (which have been used by many of the previous studies) would lead to a misleading conclusion that the size of a bank is negatively related to the loss severity. Although this may in fact be true of the mean of the loss distribution, in terms of capital calculation the result will be misleading.

In general, mean of a loss distribution provides very little information when losses are heavy-tailed. Rootzén & Klüppelberg (1999) points out that “for catastrophic risk means don't mean anything”. To illustrate this we shall provide two slightly altered examples taken from their paper.

Consider that a bank is exposed to two types of risk. Under Risk Type I a bank may lose 1 with a probability of 50% or lose 1.5 with a probability of 50% for each day of 365 days. In contrast, under Risk Type II a bank may lose 0.2 with a probability of 99.85% or lose 700.2 with a probability of 0.15% for each day of 365 days. In this example the expected loss for both risk types is 456.25. However, the distribution of losses is quite different for the two types of risk, as illustrated in figure 1. One can observe that the mean is an important summary statistic for Risk Type I. However, for Risk Type II, mean has little significance in explaining the overall risk. Under risk type II, a bank may lose 73 with a probability of 58% or lose at least 773 with a probability of 42%. Clearly the second risk type is more risky than the first, but mean alone cannot differentiate this.

The second issue Rootzén & Klüppelberg (1999) raised is that mean may be quite unstable for losses which follow a heavy-tailed distribution. To illustrate this, consider two banks, Bank I and Bank II,

**Table 1.** Mean and the 99.9<sup>th</sup> percentile loss for the hypothetical Bank I and Bank II.

|                               | Raw Loss (\$ mill)               |                                   | Log Loss (\$ mill)               |                                   |
|-------------------------------|----------------------------------|-----------------------------------|----------------------------------|-----------------------------------|
|                               | Bank I<br>(\$100 bill in assets) | Bank II<br>(\$200 bill in assets) | Bank I<br>(\$100 bill in assets) | Bank II<br>(\$200 bill in assets) |
| Mean                          | 78.17                            | 34.33                             | 4                                | 2                                 |
| 99.9 <sup>th</sup> percentile | 1025.43                          | 2387.28                           | 6.933                            | 7.778                             |

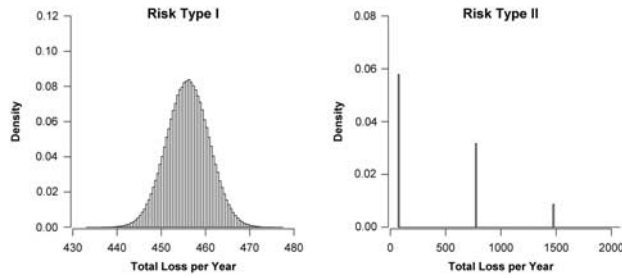


Figure 1. The distribution of total Losses for two risk types with same mean.

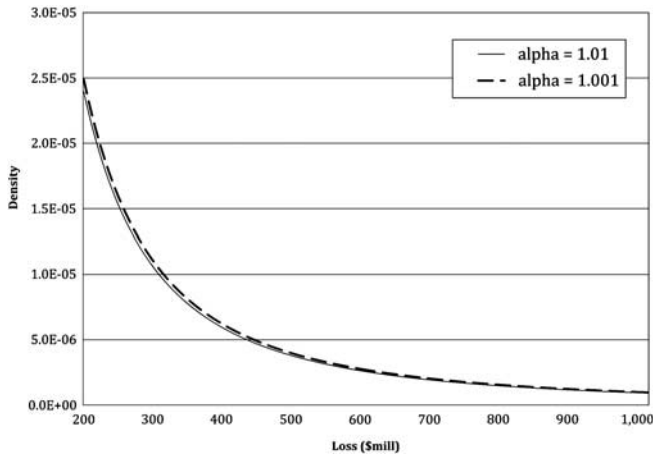


Figure 2. The density of two different Pareto risks.

in which operational risk follows a Pareto distribution;  $Pr(X > x) = x^{-\alpha}$ , for  $x > 1$  with  $\alpha = 1.01$  for bank I and  $\alpha = 1.001$  for bank II. Figure 2 provides the plots of the density function of the loss distribution for the two banks. Clearly, both distributions are very similar. In fact, the loss level at the 99.9% percentile is 933 for Bank I and 993 for Bank II. Hence, the capital amount that the two banks need to hold should be very similar. However, the mean operational loss at Bank I is 101, whereas the mean loss at Bank II is 1001. The means differ by about a factor of 10. Again, this particular example demonstrates how the mean may be misleading for heavy-tailed risks. If one analysed the mean of the loss distribution of each bank, to find a relationship between the size of a bank and the severity of losses, the results could be misleading. In this instance considering the logarithm of losses will be useful. But as we demonstrated earlier, mean of log-losses does not necessarily need to follow the same scaling mechanism as the tail of the distribution. Hence, different techniques need to be considered.

The preceding three examples illustrated why it is inappropriate to scale the severity of losses by using regression techniques which only try to fit the mean of the response variable. Such models may lead to spurious results about the relationship between the size of a firm and the loss severity even under correct distributional assumptions. A better approach to analyse the scaling properties would be to model the tail index of the loss distribution. In this study we explicitly model the tail

index of the loss distribution by using size and other explanatory variables. We considered several distributional assumptions for operational loss data and found that a Log-gamma distribution provides the best fit. Our results suggest that the tail index of the operational loss distribution and the size of a bank have a negative relationship. In other words, the tail of the operational loss distribution in large banks is heavier than in small banks. In particular, we find that the 99<sup>th</sup> percentile loss can increase by up to 25% to 35% when the size of a bank doubles. Hence, our results suggest that the size of a bank has a more significant impact on the capital requirement than thought previously.

### **3 The Methodology**

---

As discussed earlier, previous literature has revealed evidence of heteroskedasticity in operational loss data. A one possible explanation for this could be the level of sophistication in banking activities between small banks and large banks. Small banks tend to carry out less sophisticated standard banking activities such as taking deposits, providing personal loans and facilitating transactions. The severity of operational losses which can arise from such activities tends to be lower. In contrast, in addition to standard banking activities, larger banks engage in more sophisticated banking activities such as securitization, trading of exotic securities, providing syndicated loans, assisting in mergers and acquisitions, and market making. The severity of operational losses from such complex banking activities is usually higher than those losses which can arise from standard banking activities. Since large banks engage in both standard as well as complex banking activities, it is reasonable to hypothesise that the variation in the severity of operational losses in a large bank is higher than in a small bank. Hence, a scaling model for operational losses should be able to account for this variation.

Furthermore, another issue which needs to be considered when developing a scaling model is the differences in severities across different operational risk events and business lines. The loss collection exercise by the Basel Committee in 2002 and 2008 has revealed that certain operational loss events and business lines have distributions with longer tails than the others (Basel, 2003; 2009). Hence, a model which only accounts for the variation in the mean of the severities for different categories of operational risk would not be able to capture the tail behaviour properly. In other words a good scaling model should be able to account for the variation in location, scale and shape of the distribution across different business lines and event types.

A convenient framework to model location, scale and shape of a distribution is the recently introduced 'generalized additive models for location scale and shape' (GAMLSS) (Rigby & Stasinopoulos, 2001; 2005). GAMLSS is a very general class of regression model that incorporates popular GLM<sup>1</sup>, GAM<sup>2</sup>, GLMM<sup>3</sup> and GAMM<sup>4</sup> together. However, GAMLSS is more general than the above mentioned, since firstly it relaxes the assumption that the response variable belongs to the natural exponential family. Hence, GAMLSS is a convenient framework to test various distributional assumptions such as Gumbel, Weibull and Student-*t*, in addition to the standard natural exponential family distributions.

The second advantage of the GAMLSS framework is that it does not limit the modelling to the location of the distribution as in GLM and the other similar frameworks. The standard GLM setup (similarly OLS) cannot model distributional parameters other than the location parameter explicitly

<sup>1</sup> GLM – Generalized linear models

<sup>2</sup> GAM – Generalized Additive models

<sup>3</sup> GLMM – Generalized Linear Mixed Models

<sup>4</sup> GAMM – Generalized Additive Mixed Models

by using explanatory variables. The ability of the GAMLSS framework to model all the distributional parameters allows one to estimate models with more flexible scaling properties. For example, if we assume log losses are distributed normally and if we model the standard deviation using the size of the bank as an explanatory variable, we can obtain a model that allows losses in a large bank to have a larger standard deviation compared to the losses in a smaller bank, but with both banks having the same mean (and median) logarithm of loss. In other words we can implicitly allow different quantiles of the distribution to scale differently. On the other hand, if we assume losses follow a Weibull distribution with shape parameter related to some covariates, then we could obtain a model which allows loss distribution to have either negative, positive or symmetric skewness depending on the values of the covariates.

In the GAMLSS framework all distributional parameters can be explicitly modelled using both fixed and random effects (Rigby & Stasinopoulos, 2005). Furthermore, each distributional parameter can be modelled as linear and/or non-linear, parametric and/or smooth non-parametric functions of explanatory variables and/or random effects. In this study we limit our analysis to *parametric linear* GAMLSS models.

### 3.1 The Parametric Linear GAMLSS Framework

Consider log losses  $y = (y_1, y_2, \dots, y_n)^T$  are a random sample of independent observations. Let  $f(y_i; \boldsymbol{\vartheta}_i)$  be the density function conditional on the parameter vector  $\boldsymbol{\vartheta}_i$ . The  $\boldsymbol{\vartheta}_i$  can have any number of distributional parameters each of which can be modelled by explanatory variables. However, for the sake of simplicity in notation for the following discussion we shall assume  $\boldsymbol{\vartheta}_i = (\mu_i, \sigma_i)^T$  is a vector of two distributional parameters.

We define a set of link functions that specifies the relationship between the linear predictor and the distributional parameters of each component distribution as

$$\left. \begin{aligned} g_1(\mu_i) &= \eta_{i1} = \beta_{11} + \beta_{12}X_{i12} + \beta_{13}X_{i13} + \dots + \beta_{1p} X_{i1p} \\ g_2(\sigma_i) &= \eta_{i2} = \beta_{21} + \beta_{22}X_{i22} + \beta_{23}X_{i23} + \dots + \beta_{2p}X_{i2p} \end{aligned} \right\} i = 1, \dots, n,$$

where,  $X_{ijp}$  is the value of the  $p^{th}$  explanatory variable relating to the observation  $y_i$  in the  $j^{th}$  distributional parameter, and  $\beta_{jp}$  is the parameter corresponding to  $X_{ijp}$ . This set of equations can be simplified with the help of matrix notation as

$$\begin{aligned} g_1(\boldsymbol{\mu}_i) &= \mathbf{X}_1 \boldsymbol{\beta}_1 \\ g_2(\boldsymbol{\sigma}_i) &= \mathbf{X}_2 \boldsymbol{\beta}_2 \end{aligned}$$

where,  $\mathbf{X}_j$  are the design matrix of the  $j^{th}$  distributional parameter, and  $\boldsymbol{\beta}_j$  are the corresponding parameter vectors.

The maximum likelihood estimates of  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$  can be obtained by solving

$$\text{Max}_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \sum_{i=1}^n \log f(y_i; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \tag{1}$$

The optimization of eq. (1) can be carried out using the RS algorithm of Rigby and Stasinopoulos (2010, p. 92). We used the packages gamlss (Stasinopoulos & Rigby, 2007) in software R (R Development Core Team, 2010) for parameter estimation.



### 3.2 Diagnostics and Model Selection

Rigby & Stasinopoulos (2005) recommend using normalized quantile residuals of Dunn & Smyth (1996) to verify the adequacy of the fitted GAMLSS models. For a response variable  $Y$  having a continuous cumulative distribution function  $F(y_i; \theta_i)$ , Dunn & Smyth (1996) define normalized quantile residuals as  $\hat{r}_i = \Phi^{-1}[F(y_i; \hat{\theta}_i)]$ , where  $\Phi^{-1}(\cdot)$  is the inverse cumulative distribution function of the standard Normal distribution. If the model is adequate, then the error  $r_i$  should be standard Normally distributed. Any departure of the estimated residuals  $\hat{r}_i$  from the standard Normal distribution can be seen as an inadequacy of the model. Tests for normality of the estimated residuals can easily be carried out using a QQ plot of estimated residuals against the theoretical quantiles of the standard Normal distribution.

If there is more than one competing model, an information criterion can be used to test between the models. The formula to compute the information criterion is

$$CR = -2 \log(\hat{L}) + c \times (\text{degrees of freedom}) \quad (2)$$

where,  $c$  is the penalty term and the degrees of freedom is the number of fitted parameters in the model. When  $c = 2$  we have the Akaike Information Criterion (AIC), and  $c = \log(n)$  we get the Schwartz Bayesian Criterion (SBC).

### 3.3 Accounting for Reporting Bias

Most external databases suffer from reporting bias. The reporting bias is quite significant in databases maintained by vendors who collect publicly reported operational losses since not all operational losses are publicly disclosed. De Fontnouvelle *et al.* (2006) find that the reporting bias in external data is significant and losses as large as \$50 mill only have around 30% probability of being publicly disclosed. Thus, if the reporting bias is not properly taken into account parameter estimates will be biased and the model will overestimate the losses since it has been fitted using disproportionate number of large losses.

A possible method to account for the reporting bias is to use inverse probability weights  $w_i$  for each observation when computing the log-likelihood in eq. (1). Thus, the maximization problem becomes

$$\text{Max}_{\beta_1, \beta_2} \sum_{i=1}^n w_i \log f(y_i; \beta_1, \beta_2) \quad (3)$$

In order to estimate the weights we use a method based on the Extreme value theory proposed by De Fontnouvelle *et al.* (2006). The following presentation draws on their paper.

Let us assume a (log) loss  $y_i$  is publicly reported only if it exceeds some unobservable random truncation point<sup>5</sup>  $t_i$ . By assuming  $y_i$  and  $t_i$  are independent, De Fontnouvelle *et al.* (2006) show that the distribution of the observed log losses in the external database is

$$f(y_i | y_i > t_i) = \frac{f(y_i)G(y_i)}{\int_{y_i} f(y)G(y) dy} \quad (4)$$

<sup>5</sup> Here the ‘truncation point’ refers to an unobserved, observation specific random variable that determines whether a loss event is observed by the data collector or not. The reader should not confuse the ‘truncation point’ with the data collection ‘threshold’, which is known and fixed. If a loss is greater than the random ‘truncation point’ but less than the data collection ‘threshold’ the data collector will observe the loss but will not include it in the database. In contrast, if a loss amount is higher than the collection ‘threshold’ but lower than the random ‘truncation point’ then the data collector will not observe the loss, hence it will not be included in the database. Only the losses that are higher than the random ‘truncation point’ as well as the fixed data collection ‘threshold’ are included in the database.

where  $f(y_i)$  is the marginal densities of  $y_i$  and  $G(\cdot)$  is the cumulative distribution function of random truncation point  $t_i$ .

In order to estimate eq. (4) one needs to define the functions for  $f(\cdot)$  and  $G(\cdot)$ .

De Fontnouvelle *et al.* (2006) state that there are many random factors which affect whether a loss will be publicly reported or not, such as the type of loss, the business line, whether there have been any legal proceedings, and the idiosyncrasies of executives and reporters who decide whether to report the loss or not. Hence, using the central limit theorem, authors argue that the truncation point should be Normally distributed. However, they find that in practice the Normality assumption for  $G(\cdot)$  leads to non-convergence of the numerical optimization of the maximum likelihood<sup>6</sup>. Therefore, they recommend using a Logistic distribution instead, which closely approximates the Normal distribution. The cdf of the Logistic distribution is given by

$$G(t_i; \tau, a) = \frac{1}{1 + \exp\left[\frac{-(t_i - \tau)}{a}\right]}$$

where,  $\tau$  is the location parameter which indicates the log loss with a 50% probability of being reported in the database and  $a$  is the scale parameter which dictates the rate at which probability of reporting increases with the loss magnitude.

In most practical situations, the functional form of  $f(\cdot)$  remains unknown. However, if we assume  $f(\cdot)$  is in the maximum domain of attraction of the Generalized extreme value distribution<sup>7</sup>, then the Pickands–Balkema–de Haan theorem (see McNeil *et al.*, 2005, p. 275 for details) tells us that for a high enough threshold  $u$ , the conditional excess distribution function  $F_u(z_i)$  where,  $z_i = (y_i - u)$  can be approximated by a Generalized Pareto distribution (GPD), which has a cdf of the form

$$GPD(z_i; \xi, b) = \begin{cases} 1 - (1 + \frac{\xi z_i}{b})^{-1/\xi} & \xi \neq 0 \\ 1 - \exp(-\frac{z_i}{b}) & \xi = 0. \end{cases}$$

Furthermore, if  $y_i$  belongs to a heavy-tailed class of distribution (i.e. the distribution of log-losses lies within the maximum domain of attraction of the Fréchet distribution), then the distribution of  $z_i$  will converge to a GPD with  $\xi > 0$  and if  $y_i$  belongs to a medium-to-light-tailed class of distribution (i.e. the distribution of log-losses lies within the maximum domain of attraction of the Gumbel distribution), then  $z_i$  will converge to a GPD with  $\xi = 0$ , which is the cdf of an exponential distribution. It is well known that the operational losses are heavy tailed (see De Fontnouvelle *et al.*, 2007; Moscadelli, 2004). Hence, the logarithm of operational losses should belong to a medium-to-light-tailed distribution (see results 3.3.33 and 3.3.34 of Embrechts *et al.*, 1997). In other words  $z_i$  can be approximated using an exponential distribution for a high enough threshold  $u$ . Thus, for a large enough threshold  $u$  we can obtain the likelihood equation

$$L(b, \tau, a) = \prod_{i=1}^n \frac{h(z_i; b) G(z_i; \tau', a)}{\int_{\mathbb{R}} h(z; b) G(z; \tau', a) dz} \tag{5}$$

where,  $h(z_i; b) = \frac{1}{b} \exp(-\frac{z_i}{b})$  and,  $\tau' = (\tau - u)$ . Then the parameters  $b$ ,  $\tau'$ , and  $a$  can be estimated by maximising the likelihood by numerical techniques. Once the parameters are estimated one could compute the weights by

$$w_i = \frac{1}{G(y_i | \tau, a)}$$

<sup>6</sup> We also experienced the same problem with the Normality assumption.

<sup>7</sup> Essentially all the common continuous distributions of statistics are in the maximum domain attraction of the Generalized extreme value distribution. See p. 267 of McNeil *et al.* (2005) for further details.

It is also possible to normalize the weights by using the following formula, such that the sum of weights will add up to the number of observations.

$$w_i^j = \frac{nw_i}{\sum_{i=1}^n w_i} \quad (6)$$

where,  $n$  is the total number of observations.

## 4 Application of the Model

We applied the model described in section 3 to Algorithmics's 'Algo Opdata' database. The database contains publicly reported operational losses above USD \$1 mill from financial firms around the world. In order to obtain a fairly homogeneous sample of losses we restricted our analysis to data collected from the US banking industry during the 15 year time period from 1-Jan-1994 to 31-Dec-2008 from banks with assets greater than or equal to \$100 bill. There are several reasons for excluding banks with assets less than \$100 bill from our analysis. Firstly, most of the records for banks with assets less than \$100 bill were incomplete as they did not have values for total deposits, total equity, and number of employees. We made an attempt to recover the missing values but were unsuccessful because documents such as annual reports were not publicly available as most of these small banks are not publicly traded. Secondly, since Algo OpData is collected from public sources such as newspapers, court documents, SEC filings etc., we suspected that the reporting bias in small banks could be different from the reporting bias in medium to large scale banks for the following reasons: 1) usually media tend to treat losses in large banks as more newsworthy than losses in small local banks, therefore the probability of data vendor picking up a loss from a media source is greater for a medium to large scale bank than for a small local bank, 2) regulatory disclosure requirements of small banks are different from medium to large scale banks<sup>8</sup>, therefore probability of a loss being publicly disclosed is lower for small banks, 3) disclosures requirements of publicly listed entities are much more greater than for non-listed entities. Given these circumstances we decided to limit our analysis to banks with assets greater than or equal to \$100 bill. However, it should be pointed out that limiting the analysis to banks with assets greater than or equal to \$100 bill does not discount the significance of the study, since our sample contains 855 data points, with a fair range of banks from \$100 bill to \$2000 bill in assets as shown in table 2.

In the following section we shall present the preliminary analysis of the chosen sample and a brief description of the potential explanatory variables for model fitting and their descriptive statistics.

### 4.1 Preliminary Analysis of the Data and the Explanatory Variables

#### 4.1.1 Size

As discussed earlier much of the previous work has suggested an increasing relationship between loss magnitude and the size of the firm. Algo OpData contains several possible quantities that can be taken as a measurement of firm size, namely total assets, total deposits, total equity and the number of employees. We considered each of these variables in the regression analysis to find the one with the highest explanatory power.

<sup>8</sup> For example, small banks are not subject to the Section 32 of FIDCIA which requires banks to establish an independent audit committee and report on the effectiveness of the entity's internal control. Altamuro & Beatty (2010) find evidence that the quality of financial reporting of banks subject to Section 32 of FIDCIA is better than the small banks which are exempted.

**Table 2.** The Distribution of the Selected Sample by Total Assets.

| Assets (\$ bill) | Number of Obs. | Assets (\$ bill) | Number of Obs. | Assets (\$ bill) | Number of Obs. |
|------------------|----------------|------------------|----------------|------------------|----------------|
| 100 to 200       | 83             | 800 to 900       | 84             | 1500 to 1600     | 25             |
| 200 to 300       | 103            | 900 to 1000      | 38             | 1600 to 1700     | 33             |
| 300 to 400       | 74             | 1000 to 1100     | 67             | 1700 to 1800     | 22             |
| 400 to 500       | 70             | 1100 to 1200     | 36             | 1800 to 1900     | 3              |
| 500 to 600       | 44             | 1200 to 1300     | 26             | 1900 to 2000     | 11             |
| 600 to 700       | 10             | 1300 to 1400     | 13             | 2000 to 2100     | 10             |
| 700 to 800       | 61             | 1400 to 1500     | 30             | 2100+            | 12             |

**Table 3.** The Summary Statistics by Business Line.

| Business Line           | Business Line Index | Number of Losses | % of Losses | Percentile       |                  |                  |
|-------------------------|---------------------|------------------|-------------|------------------|------------------|------------------|
|                         |                     |                  |             | 50 <sup>th</sup> | 75 <sup>th</sup> | 99 <sup>th</sup> |
| Corporate Finance (CF)  | 1                   | 117              | 13.7%       | 26.8             | 93.7             | 4133.0           |
| Trading and Sales (TS)  | 2                   | 176              | 20.6%       | 15.9             | 63.7             | 1905.9           |
| Retail Banking (RB)     | 3                   | 149              | 17.4%       | 7.2              | 21.1             | 1222.6           |
| Commercial Banking (CB) | 4                   | 127              | 14.9%       | 12.0             | 33.4             | 662.0            |
| Retail Brokerage (RBr)  | 5                   | 179              | 20.9%       | 5.0              | 19.8             | 626.3            |
| PS, AS, AM              | 6                   | 107              | 12.5%       | 17.2             | 56.0             | 731.6            |

Note 1: Basel II classifies operational losses into eight business lines. Due to the limited amount of data 'Payment and Settlement', 'Agency Services', and 'Asset Management' were grouped together and have been labelled as 'PS, AS, AM'.

Note 2: All losses are expressed in 2008 dollar terms.

#### 4.1.2 Business Lines (BL)

Certain lines of business can be perceived to be more risky than others. Thus, severity of operational losses may depend on the business line. As a preliminary analysis we classified business activity into six business lines and computed the median, 75<sup>th</sup>, and 99<sup>th</sup> percentiles of the losses for each business line. The results are given in table 3.

Results show that the highest number of losses have occurred in 'retail brokerage' followed by 'trading and sales'. 'Corporate finance' had the highest losses in all three percentiles indicating that this particular business line is relatively riskier than the others. In contrast, retail brokerage had lower severities at all three percentiles.

#### 4.1.3 Event Type (ET)

The type of operational loss is another potential explanatory variable. We categorised losses into five different event types as outlined in table 4. Preliminary analysis showed that more than half of the operational losses were due to 'clients, products & business practices' (CPBP). CPBP had large losses at higher percentiles indicating particular event category to be riskier than the others.

#### 4.1.4 Equity Ratio

Ideally when modelling operational risk one needs to take into account the internal controls and risk culture of a firm. Unfortunately the Algo OpData does not provide any such information.

**Table 4.** The Summary Statistics by Event Type.

| Event Type  | ET Index | Number of Losses | % of Losses | Percentile       |                  |                  |
|---|----------|------------------|-------------|------------------|------------------|------------------|
|   |          |                  |             | 50 <sup>th</sup> | 75 <sup>th</sup> | 99 <sup>th</sup> |
| Internal fraud (IF)                               | 1        | 133              | 15.6%       | 10.1             | 31.7             | 637.2            |
| External fraud (EF)                               | 2        | 118              | 13.8%       | 7.8              | 24.3             | 250.7            |
| Clients, Products and Business Practices (CPPB)   | 3        | 457              | 53.5%       | 13.8             | 62.4             | 2966.8           |
| Execution, Delivery and Process Management (EDPM) | 4        | 66               | 7.7%        | 9.6              | 36.9             | 662.0            |
| EPWS, DPA, BDSF                                   | 5        | 81               | 9.5%        | 5.4              | 26.2             | 856.4            |

Note 1: Basel II classifies operational losses into seven event types. Due to the limited amount of data the 'Employment Practices and Workplace Safety', 'Damage to Physical Assets', and 'Business Disruption and System Failures' were grouped together and have been labelled as 'EPWS, DPA, BDSF'.

Note 2: All losses are expressed in 2008 dollar terms.

Therefore, we turned to other publicly available variables which can be used as proxies for risk culture and internal controls.

The first variable we considered was a governance index created by Gompers *et al.* (2003). This index measures shareholder rights in comparison to management power. The index can be considered as a proxy of the quality of governance, assuming that the greater the power held by management, the more likely they are to breach fiduciary duties and adopt *ad hoc* market practices. However, we found little use of the index since index values were not available for the full sample duration as well as for many small banks in the sample.

Secondly, as an alternative we considered equity ratio as a proxy for the risk taking tendency of the senior management. The Equity ratio of a company is defined as the ratio between shareholders' equity and total assets. The ratio indicates the proportion of equity used to finance a company's total assets. It is a measure of leverage used as an alternative to the popular debt-to-equity ratio. Since Algo OpData provides information on both assets as well as the shareholder equity, the ratio was readily computable. We hypothesized that a high leverage is an indicator of the overall riskiness of the company and can therefore provide insight into risk culture within the company. If our hypothesis is correct the regression analysis should yield that the variable is positively correlated with the losses.

## 4.2. Estimating the Weights

The first step in developing the scaling model is to compute weights for the observations by using the techniques discussed in section 3.3. First, we plot a sample mean excess plot of the losses to confirm that the losses belong to a heavy-tailed class of distribution. We find that the mean excess plot has a linear upward trend as expected (see figure 3), which confirms that our data belongs to a distribution that lies in the maximum domain attraction of Fréchet distribution. In other words, distribution of the logarithm of losses should lie in the maximum domain attraction of Gumbel distribution, and therefore, for a high enough threshold one can approximate the excess log-loss distribution to an exponential distribution as described in section 3.3.

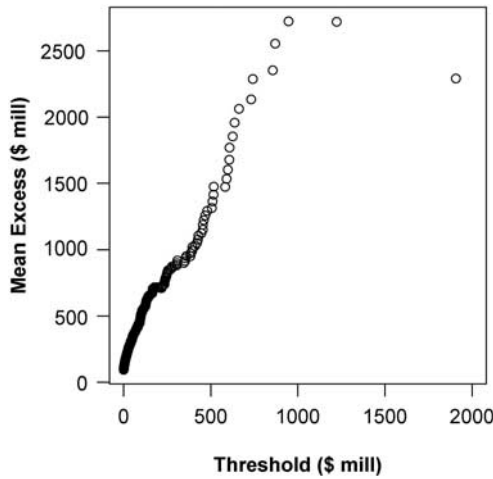


Figure 3. Mean Excess Plot of Operational Losses.

Table 5. Estimation Results  $b$ ,  $\tau$ , and  $a$  for Various Thresholds.

| Parameter    | Threshold $u$ |      |      |      |      |      |       |       |       |       |
|--------------|---------------|------|------|------|------|------|-------|-------|-------|-------|
|              | \$1M          | \$2M | \$3M | \$4M | \$5M | \$7M | \$10M | \$12M | \$15M | \$20M |
| $\hat{a}$    | 0.92          | 0.93 | 0.88 | 0.89 | 0.96 | 0.96 | 0.96  | 0.98  | 0.96  | 0.96  |
| $\hat{\tau}$ | 4.91          | 4.15 | 2.91 | 2.67 | 3.27 | 3.17 | 2.60  | 2.69  | 2.30  | 2.06  |
| $\hat{b}$    | 0.86          | 0.87 | 1.00 | 0.99 | 0.88 | 0.85 | 0.87  | 0.84  | 0.86  | 0.85  |

Note: We carried out likelihood ratio tests of the restriction that the reporting probabilities are constant across all losses for each threshold level (i.e. there is no reporting bias in the data). The p-values of the likelihood ratio tests for all the threshold values were less than 0.01, which confirmed the existence of reporting bias.

It is well known that the GPD approximation works well only if the threshold  $u$  is selected properly. If the threshold  $u$  is chosen correctly, then for any other threshold  $v (>u)$  the estimated parameter values for  $b$  should remain equal to the initial estimates. To verify this we estimated the parameters  $b$ ,  $\tau$ , and  $a$  by maximizing the likelihood equation given in eq. (5) for various thresholds. The estimated parameter values are given in table 5. We find that the parameter values for  $b$  and  $a$  stabilizes after the \$5 mill threshold. Hence, weights were computed using eq. (6) where  $\hat{a} = 0.96$ , and  $\hat{\tau} = (3.27 + \log(5))$ . These weights have been used to estimate all the GAMLSS models presented in the subsequent sections.

### 4.3 Model Selection

#### 4.3.1 Choosing a Base Model

The initial step in model selection is to find a base model having no explanatory variables which closely follows the data. In other words we are seeking an appropriate probability distribution assumption for model fitting. As discussed earlier in section 3.3 operational losses are heavy tailed, hence logarithm of operational losses should belong to a distribution which lies in the maximum domain of attraction of the Gumbel distribution. Distributions within this class exhibit tails with

exponential decay. Many thin to medium tailed distributions fall under this class of distributions including the Normal, Lognormal, Exponential, Gamma, Weibull, and Gumbel among others (McNeil *et al.*, 2005, p. 269). We fitted several of these distributions to the data and computed the AIC values using eq. (2). A list of the distributions we fitted together with their density functions, formulas for mean and variance are given in table 6. The estimated parameter values of the fitted models and their AIC values are given in table 7.

**Table 6.** The List of Fitted Distributions.

| Name                | Density Function, $f(y)$   | $E(y)$                                       | $Var(y)$  |
|---------------------|--|--|---|
| Exponential         | $\frac{1}{\mu} \exp\left(-\frac{y}{\mu}\right)$  | $\mu$  | $\mu^2$   |
| Gamma <sup>a</sup>  | $\frac{\mu^\sigma}{\Gamma(\sigma)} y^{\sigma-1} \exp(-y\mu)$   | $\frac{\sigma}{\mu}$                         | $\frac{\sigma}{\mu^2}$  |
| Inverse Gaussian    | $\frac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left[-\frac{1}{2\mu^2\sigma^2 y}(y-\mu)^2\right]$                                      | $\mu$  | $\sigma^2 \mu^3$  |
| Log-Normal          | $\frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{y} \exp\left[-\frac{[\log(y)-\mu]^2}{2\sigma^2}\right]$                                | $e^{\left(\mu+\frac{\sigma^2}{2}\right)}$    | $\left(e^{\sigma^2}-1\right)e^{(2\mu+\sigma^2)}$  |
| Gumbel <sup>b</sup> | $\frac{1}{\sigma} \exp\left[\left(\frac{y-\mu}{\sigma}\right)-\exp\left(\frac{y-\mu}{\sigma}\right)\right]$                    | $\mu-\gamma\sigma$                           | $\frac{\pi^2\sigma^2}{6}$   |
| Normal              | $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$  | $\mu$  | $\sigma^2$  |
| Three param- $t^c$  | $\frac{1}{\sigma B\left(\frac{1}{2}, \frac{v}{2}\right)^{1/2}} \left[1 + \frac{(y-\mu)^2}{\sigma^2 v}\right]^{-\frac{v+1}{2}}$ | $\mu$  | $\frac{\sigma^2 v}{v-2}$  |
| Weibull             | $\frac{\sigma y^{\sigma-1}}{\mu^\sigma} \exp\left[-\left(\frac{y}{\mu}\right)^\sigma\right]$                                   | $\mu\Gamma\left(\frac{1}{\sigma} + 1\right)$ | $\mu^2 \left\{ \Gamma\left(\frac{2}{\sigma} + 1\right) - \left[\Gamma\left(\frac{1}{\sigma} + 1\right)\right]^2 \right\}$ |

a -  $\Gamma(\cdot)$  is the gamma function.

b -  $\gamma$  is the Euler–Mascheroni constant  $\approx 0.5772$ .

c -  $B(\cdot)$  is the beta function.

**Table 7.** The Competing Base Models.

| Distribution                  | Parameter Estimates |                | AIC    |
|-------------------------------|---------------------|----------------|--------|
| Gamma                         | $\mu$               | 1.259 (0.055)  | 1716.8 |
|                               | $\sigma$            | 1.284 (0.059)  |        |
| Weibull                       | $\mu$               | 1.074 (0.034)  | 1720.3 |
|                               | $\sigma$            | 1.153 (0.030)  |        |
| Truncated three parameter $t$ | $\mu$               | 0.144 (0.195)  | 1726.4 |
|                               | $\sigma$            | 0.985 (0.117)  |        |
|                               | $v$                 | 4.041 (1.112)  |        |
| Truncated Normal              | $\mu$               | -4.476 (2.009) | 1737.6 |
|                               | $\sigma$            | 2.539 (0.410)  |        |
| Truncated Gumbel              | $\mu$               | -8.067 (1.910) | 1743.4 |
|                               | $\sigma$            | 5.405 (0.710)  |        |
| Exponential                   | $\mu$               | 1.020 (0.035)  | 1745.8 |
| Log-Normal                    | $\mu$               | -0.417 (0.037) | 1844.8 |
|                               | $\sigma$            | 1.078 (0.026)  |        |
| Inverse Gaussian              | $\mu$               | 1.020 (0.061)  | 2287.3 |
|                               | $\sigma$            | 1.720 (0.042)  |        |

Note: Since only the losses which are above the \$1 mill threshold are included in the database the range of the log-losses is  $y_1 \in [0, \infty)$ . Hence, truncated versions of Normal, Gumbel, and  $t$ -distribution were fitted to the data.

As reported in table 7, the model fitted with the Gamma distribution provided the lowest AIC value. Hence, we chose the Gamma distribution for model fitting.

It is worth pointing out that when the logarithm of a random variable  $X$  follows a Gamma distribution,  $X$  itself is said to follow a Log-gamma distribution. Log-gamma distribution is a heavy-tailed distribution with a Pareto-type tail behaviour. That is, there exists a *slowly varying*<sup>9</sup> function  $h(\cdot)$  such that

$$Pr(X > x) = x^{-\alpha} h(x) \tag{7}$$

The parameter  $\alpha$  in the above equation is called the tail index of the distribution. The value of  $\alpha$  determines the thickness of the tail. The smaller the  $\alpha$ , the thicker the tail of the distribution. In other words, extreme observations become even more likely. This particular characteristic makes Pareto-type distributions (which includes Log-gamma) an attractive framework to model operational risk.

The density function of the Log-gamma distribution is given by

$$f(x) = \frac{\alpha^p}{\Gamma(p)} x^{-\alpha-1} [\log(x)]^{p-1} \tag{8}$$

where,  $\Gamma(\cdot)$  is the gamma function, and  $p$  and  $\alpha$  are the lower and upper shape parameters. The distribution function of the Log-gamma distribution is given as

$$F(x) = \frac{\gamma(p, \alpha \log(x))}{\Gamma(p)} \tag{9}$$

where,  $\gamma(\cdot)$  is the lower incomplete gamma function. The expected value of the Log-gamma random variable is given by

$$E(x) = \left(\frac{\alpha}{\alpha-1}\right)^p \quad \text{for } \alpha > 1 \tag{10}$$

and, the variance by

$$Var(x) = \left(\frac{\alpha}{\alpha-2}\right)^p - \left(\frac{\alpha}{\alpha-1}\right)^{2p} \quad \text{for } \alpha > 2 \tag{11}$$

The coordinates  $(F, L)$  of the Lorenz curve is given by

$$\left( F = \frac{\gamma(p, \alpha \log(x))}{\Gamma(p)}, L = \frac{\gamma(p, (\alpha-1) \log(x))}{\Gamma(p)} \right) \tag{12}$$

Note that if  $X \sim \text{Log-gamma}(\alpha, p)$  as defined in eq. (8) then a random variable  $Y = \log(X)$  follows a gamma distribution as defined in table 7 with  $\mu = \alpha$  and  $\sigma = p$ , where  $\mu$  is called the rate parameter and  $\sigma$  is called the shape parameter.

The Log-gamma distribution has a finite variance only if  $\alpha > 2$ . Moreover, distribution has a finite mean only if  $\alpha > 1$ . A distribution with  $\alpha < 1$  is referred to as an *extremely heavy-tailed* distribution or as an *infinite mean model*. Such distributions can have important implications for capital calculations as have been pointed out by Nešlehová *et al.* (2006). We shall discuss this issue in detail later in section 4.6.1.

<sup>9</sup> see Embrechts *et al.* (1997, p. 564) for the definition of slowly varying functions.



### 4.3.2 Step-wise Selection of Explanatory Variable

Once the distributional assumption has been established, we carried out a step-wise selection of the explanatory variables discussed in section 4.1 for the parameters  $\mu$  and  $\sigma$  of the Gamma distribution. The strategy which we used to select the explanatory variables is presented schematically in figure 4.

The choice of link functions is an important consideration in fitting GAMLSS models. For the Gamma distribution both parameters  $\mu$  and  $\sigma$  are defined as greater than zero. Therefore, log link function appears to be a suitable candidate for both parameters. However, inverse link function is the canonical link function for Gamma distribution in the GLM framework<sup>10</sup>. Hence, we tested models with log link function as well as inverse link function by comparing the AIC values and the normalized quantile residual plots (discussed earlier in subsection 4.4). We found that models fitted using a log link function to  $\mu$  and  $\sigma$  provided lower AIC values as well as better residual plots. Therefore, log link functions were chosen for both parameters.

As mentioned earlier Algo Opdata provides several possible quantities, namely total assets, total deposits, total equity and number of employees, which could be used to measure the size of a bank. We considered logarithm of each of these variables in the regression analysis and compared the AIC values and the residual plots to find the variables with the highest explanatory power. Both log-assets as well as log-equity provided models with slightly better fit than the models based on log(number of employees) and log-deposits. Given that all these variables are highly correlated it is no surprise that all of them were found to be significant. Since total assets is usually a better measure of the size of a bank and the magnitude of its operations than total equity, and also because previous studies have used total assets for their analysis, we decided to use log-assets in model fitting over log-equity.

The parameter estimates of the final model given by the step-wise selection method is shown in table 8 (denoted as model G). The model has been fitted using retail brokerage as the baseline category for business lines and the event variable 'EPWS, DPA, BDSF' as the baseline category for event types.

Based on the step-wise selection method outlined earlier we find log-assets, equity ratio, corporate finance, commercial banking, 'PS, AS, AM' business lines, and the clients, products and business practices event type are significant explanatory variables to the rate parameter  $\mu$ . As for the shape parameter  $\sigma$ , log-Assets and commercial banking were found to be significant explanatory variables.

We find that the size of a firm is a significant explanatory variable for both shape as well as the rate parameter of the distribution. In particular, we find a negative correlation between the size of a bank and the rate parameter  $\mu$ . This implies that the tail of the operational loss distribution becomes heavier with the size variable. This is an important finding since previous studies have mostly focused on the effect of the size of a bank on the mean of the log-loss distribution and had come to the conclusion "...size effect is of little significance in explaining the level of losses" (Dahen & Dionne, 2010, p. 1491). A detailed analysis of this important result as well as an interpretation of the estimated parameters will be provided later in subsection 4.6. In the subsequent sections we will present the robustness analysis for the parameter estimates and the model diagnostics.

<sup>10</sup> However, note that our parameterization of the Gamma distribution is different from the GLM parameterization.

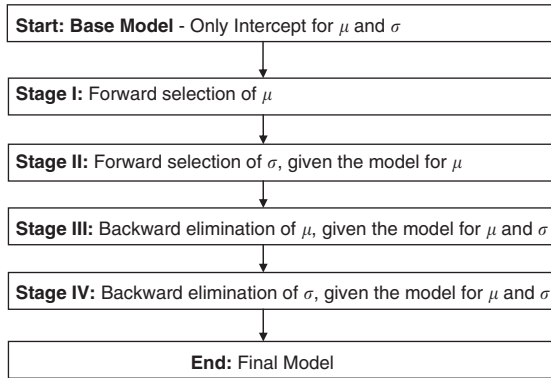


Figure 4. The Step-Wise Selection Strategy of Explanatory Variables.

Table 8. Estimated Parameter Values for Model G.

| Explanatory Variable                       | $\mu$    |          | $\sigma$ |          |
|--|----------|----------|----------|----------|
|  | Estimate | St. Err. | Estimate | St. Err. |
| Intercept                                  | 2.69     | (0.909)  | 4.801    | (0.857)  |
| Log-assets                                 | -0.182   | (0.067)  | -0.331   | (0.065)  |
| Equity Ratio                               | 3.983    | (1.1)    | -        | -        |
| Corporate Finance                          | -0.37    | (0.11)   | -        | -        |
| Trading and Sales                          | -        | -        | -        | -        |
| Retail Banking                             | -        | -        | -        | -        |
| Commercial Banking                         | -0.438   | (0.139)  | -0.279   | (0.159)  |
| PS, AS, AM                                 | -0.270   | (0.103)  | -        | -        |
| Internal fraud                             | -        | -        | -        | -        |
| External fraud                             | -        | -        | -        | -        |
| Clients, Products and Business Practices   | -0.148   | (0.048)  | -        | -        |
| Execution, Delivery and Process Management | -        | -        | -        | -        |

Note: the model is fitted using retail brokerage as the baseline line category for business lines and ‘EPWS, DPA, BDSF’ as the baseline category for event types.

## 4.4 Model Diagnostics and Robustness Analysis

### 4.4.1 Diagnostic Plots

The diagnostics for the fitted model were carried out using a QQ plot of normalized quantile residuals as described in section 3.2. The resulting QQ plot is given in figure 5. The QQ plot indicates that the estimated residuals closely follow a standard Normal distribution. Hence, we concluded that the model provides a reasonable fit to the data.

### 4.4.2 Robustness of Parameter Estimates

In order to test whether the estimated parameters are robust to model specification and the underlying data we analyzed the following two cases:

- 1) Model G.1 – refit model G after removing losses above \$1 bill in banks with assets greater than \$500 bill.

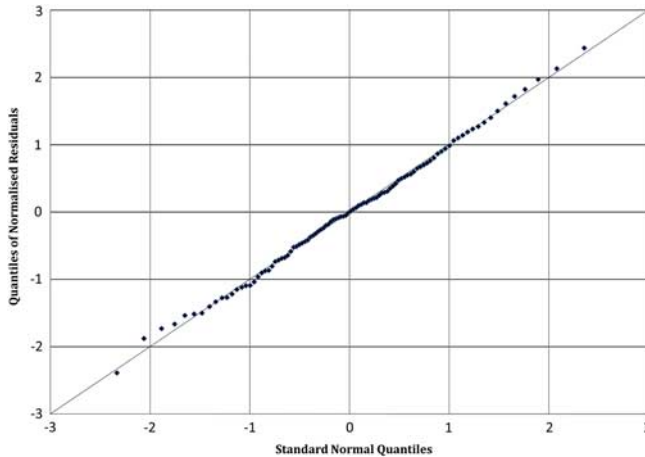


Figure 5. Normalized Quantile Residual Plot of the Model G.

- 2) Model G.2 to G.11 – sequentially add explanatory variable to a model with only intercept and log-assets as explanatory variables.

The Case 1 test the robustness of the parameter estimates against few extreme losses occurred in large banks. There were 5 losses above \$1 billion in banks with assets greater than \$500 billion. The refitted model after removing those data points are given in table 9 under the column heading model G.1. We find that the parameter estimates do not vary much from initial estimates. Especially, the coefficient of the size variable remains remarkably stable even after removing the large losses.

Since size is an important variable for the scaling model, we analyzed the stability of the coefficient of the size variable with respect to model specification in Case 2, by fitting a series of models with different specifications. We started with model G.2 which only has the log-assets as explanatory variables. Then we sequentially added other explanatory variables to the simple model and tested the stability of the coefficients of the size variable. The results show that the size variable remained relatively stable as well as statistically significant, across all models.

## 4.5 Validation of the Model by Simulation

As a further validation of the model we simulated losses for the US banking industry, and two individual US banks, and compared the simulated losses against the observed losses. For individual banks we chose two banks with sufficient loss data for statistical estimations. To protect anonymity we shall refer to the two banks as Bank A and Bank B. Bank A had incurred 141 losses and Bank B had incurred 92 losses during the sample period. Both banks were operating continually throughout the sample period. The results of the simulations are presented in the following sections.

### 4.5.1 Simulating Losses for the US Banking Industry

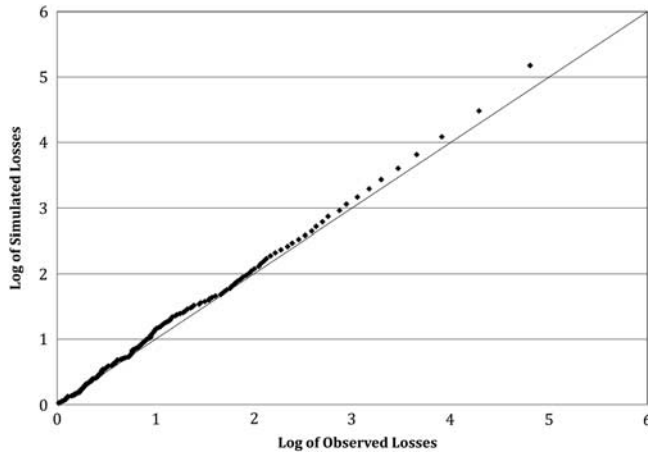
We simulated 1,000,000 random losses for the US banking industry by using the observed values of the explanatory variables in the fitted model repeatedly. Then we tested the goodness-of-fit by comparing the quantiles of the simulated losses with the observed losses using a QQ plot. Furthermore, we constructed distribution free 95% confidence intervals for the quantiles of the

**Table 9.** The Robustness Analysis.

| Exp. Variable     | Model G.1         |                   | Model G.2         |                   | Model G.3         |                   | Model G.4         |                   | Model G.5         |                   |
|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
|                   | $\mu$             | $\sigma$          | $\mu$             | $\sigma$          | $\mu$             | $\sigma$          | $\mu$             | $\sigma$          | $\mu$             | $\sigma$          |
| Intercept         | 2.647<br>(0.872)  | 4.783<br>(0.746)  | 2.786<br>(0.935)  | 4.133<br>(0.949)  | 2.824<br>(0.867)  | 4.209<br>(0.816)  | 2.021<br>(0.924)  | 4.157<br>(0.892)  | 2.029<br>(0.941)  | 4.192<br>(0.860)  |
| Log-assets        | -0.179<br>(0.064) | -0.33<br>(0.057)  | -0.188<br>(0.069) | -0.288<br>(0.070) | -0.191<br>(0.065) | -0.293<br>(0.062) | -0.151<br>(0.068) | -0.288<br>(0.068) | -0.147<br>(0.069) | -0.29<br>(0.065)  |
| Equity Ratio      | 4.042<br>(1.100)  | -                 | -                 | -                 | -                 | -                 | 4.260<br>(1.052)  | -                 | 4.075<br>(1.048)  | -                 |
| Corporate Finance | -0.362<br>(0.114) | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -0.344<br>(0.115) | -                 |
| Trading & Sales   | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 |
| Retail Banking    | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 |
| Comm. Banking     | -0.439<br>(0.124) | -0.281<br>(0.150) | -                 | -                 | -                 | -0.058<br>(0.102) | -                 | -0.037<br>(0.103) | -                 | -0.017<br>(0.105) |
| PS, AS, AM        | -0.270<br>(0.106) | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 |
| Internal fraud    | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 |
| External fraud    | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 |
| CLBP              | -0.146<br>(0.053) | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 |
| EDPM              | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 | -                 |

Table 9. *Continued*

| Exp. Variable     | Model G.6         |                   | Model G.7         |                   | Model G.8         |                   |
|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
|                   | $\mu$             | $\sigma$          | $\mu$             | $\sigma$          | $\mu$             | $\sigma$          |
| Intercept         | 2.389<br>(0.832)  | 4.496<br>(0.756)  | 2.548<br>(0.953)  | 4.675<br>(0.892)  | 3.216<br>(0.998)  | 5.107<br>(1.008)  |
| Log-assets        | -0.169<br>(0.061) | -0.31<br>(0.058)  | -0.178<br>(0.069) | -0.322<br>(0.068) | -0.190<br>(0.070) | -0.324<br>(0.071) |
| Equity Ratio      | 4.043<br>(1.143)  | -                 | 4.208<br>(1.145)  | -                 | 1.506<br>(2.673)  | -2.681<br>(2.913) |
| Corporate Finance | -0.364<br>(0.114) | -                 | -0.402<br>(0.118) | -                 | -0.595<br>(0.158) | -0.192<br>(0.149) |
| Trading & Sales   | -                 | -                 | -                 | -                 | -0.125<br>(0.173) | -0.005<br>(0.189) |
| Retail Banking    | -                 | -                 | -                 | -                 | -0.138<br>(0.132) | -0.026<br>(0.136) |
| Comm. Banking     | -0.368<br>(0.138) | -0.269<br>(0.155) | -0.421<br>(0.126) | -0.284<br>(0.150) | -0.633<br>(0.152) | -0.374<br>(0.175) |
| PS, AS, AM        | -                 | -                 | -0.295<br>(0.105) | -                 | -0.577<br>(0.170) | -0.237<br>(0.163) |
| Internal fraud    | -                 | -                 | -                 | -                 | -0.273<br>(0.157) | -0.294<br>(0.151) |
| External fraud    | -                 | -                 | -                 | -                 | 0.046<br>(0.185)  | -0.054<br>(0.181) |
| CLBP              | -                 | -                 | -                 | -                 | -0.319<br>(0.143) | -0.212<br>(0.140) |
| EDPM              | -                 | -                 | -                 | -                 | -0.071<br>(0.252) | -0.033<br>(0.265) |



**Figure 6.** QQ Plot of Simulated Losses vs. Observed Losses for US Banking Industry. Note: quantiles for the observed losses have been computed using the weights estimated in subsection 4.2.

**Table 10.** Summary Statistics of the Simulation for the US Banking Industry.

| Quantile | Observed Losses (\$M)  | Simulated Losses (\$M) |
|----------|------------------------|------------------------|
|          | [95 % CI]              | [95 % CI]              |
| 25%      | 1.4 [1.3, 1.6]         | 1.5 [1.5, 1.5]         |
| 50%      | 2.1 [1.8, 2.3]         | 2.3 [2.3, 2.3]         |
| 75%      | 3.9 [3.2, 4.7]         | 4.4 [4.4, 4.4]         |
| 90%      | 9.2 [7.9, 11]          | 10.2 [10.1, 10.2]      |
| 95%      | 17.6 [14.7, 20.2]      | 19.3 [19.2, 19.5]      |
| 97%      | 26.9 [23.1, 32.4]      | 31.0 [30.7, 31.4]      |
| 99%      | 72.9 [57.4, 90.7]      | 88.5 [86.9, 90.3]      |
| 99.9%    | 711.19 [516.7, 1905.9] | 955.0 [891.5, 1027.6]  |

Note 1: percentiles for the observed losses have been computed using the weights estimated in subsection 4.2.

observed losses using the methodology outlined in Arnold *et al.* (1992). The resultant QQ plot and the table of some of the selected quantiles with their confidence intervals are given in figure 6 and table 10 respectively.

We find that the QQ plot follows a fairly straight line with a unit gradient confirming our model. The model fits particularly well in the lower percentiles of the loss distribution. However, the QQ plot indicates a possible deterioration in the fit towards the tail of the distribution. The reported values in table 10 shows that simulated loss is 9.7% higher than the observed loss at the 95<sup>th</sup> percentile. At the 99.9<sup>th</sup> percentile the difference between the simulated and the observed percentiles becomes 34.3%. However, a closer look at the table reveals that the percentiles of the simulated losses lie well within the 95% confidence intervals of the percentiles of the observed losses. Hence, there is not enough evidence to suggest that the model is overestimating the losses. One could observe from table 10 that the confidence intervals of the higher percentiles of the observed losses are fairly wide due to lack of data. This increases the uncertainty of the goodness-of-fit in the tail of the

distribution. In such situations one would need to take a conservative approach to assess the goodness-of-fit of the model. Under the given circumstances we believe that the seeming overestimation of the far tail of the distribution by the fitted model will reduce the uncertainty of underestimation of the tail. Thus, this apparent overestimation in the tail can be seen as a strength of the model rather than a weakness. Hence, we conclude that the model provides a reasonable, yet conservative fit, to the data.

#### 4.5.2 Simulating Losses for Two Individual Banks in the US

Our purpose is to build a model that can simulate losses to a bank when there is not enough internal data of their own to assess the tail of the loss distribution accurately. Thus, to verify whether our model can be used for this purpose we re-estimated the parameters of model G by removing the losses incurred in Bank A from the dataset and then simulated 1,000,000 random losses using the new parameter estimates. Similarly, 1,000,000 random losses were simulated for Bank B by re-estimating the parameters of model G by removing the losses incurred in Bank B from the dataset. The QQ plots of the observed losses against the simulated losses are given in figure 7 and figure 8. A summary table of the percentiles and their confidence intervals is given in table 11. The re-estimated parameter values for model G is given in table 12 and table 13.

The QQ plots for both banks followed a straight line with a unit gradient confirming our model. table 11 shows that the percentiles of the simulated losses are well within the 95% confidence intervals of the observed percentiles. Hence, the results confirm that the model can be used by individual banks to complement their internal data.

### 4.6 Interpretation of the Fitted Model

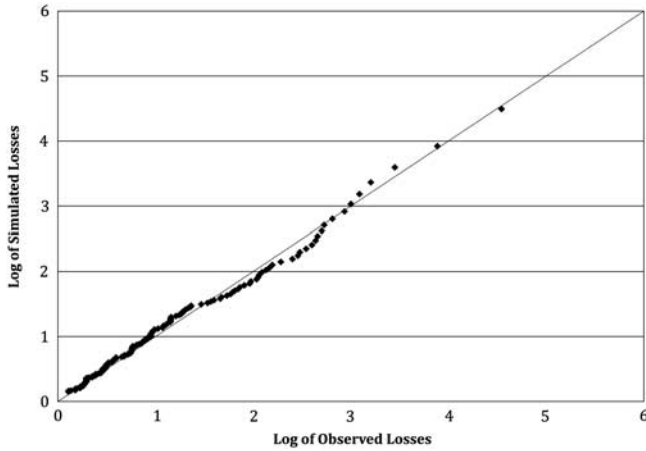
In this section we will analyze how various parameters in the fitted model affect the severity of the loss distribution. In particular, we are interested in how parameters affect the tail index of distribution, 99<sup>th</sup> percentile loss, and the average operational loss.

#### 4.6.1 The Effect of Size on Operational Loss Severity

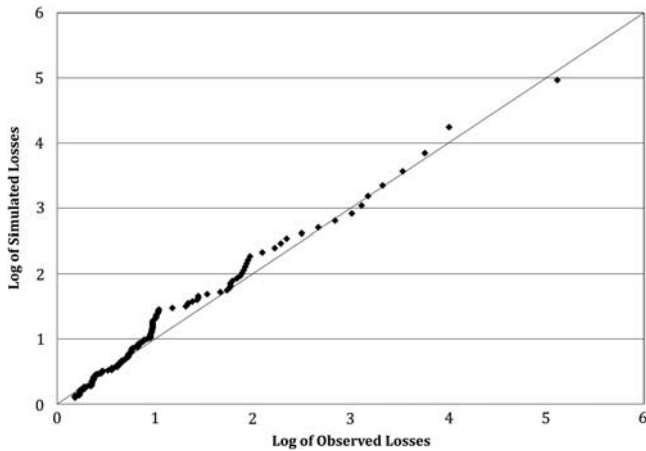
To analyze how the size of a bank may affect its operational loss distribution we consider a series of hypothetical banks with various sizes ranging from \$200 bill in assets to \$2000 bill in assets. The analysis is carried out by keeping the equity ratio of the hypothetical banks at 6.7%, which is the median equity ratio for US banks in our sample. The computed mean, 99<sup>th</sup> percentile loss, and the tail index for each bank is given in table 14, table 15, and table 17 respectively. Table 16 provides the percentage change of the 99<sup>th</sup> percentile of the loss distribution as the size of a bank doubles.

The results in table 14 show that mean loss severity for most BL/ET categories increases with the size of the bank. The baseline combination 'OtherB-OtherE' is the main exception<sup>11</sup>. It is somewhat puzzling why we may observe a decreasing relationship between severity of loss and the size of the bank in this particular BL/ET category. However, a previous study by Cope & Labbi (2008) had also reported similar findings. Given that 57% of the number of losses in this category were due to fraud (which accounted for 69.6% of the total loss in this category), our results are consistent with the findings reported by Cope & Labbi (2008) where they have reported a decreasing relationship

<sup>11</sup> The average loss in the 'OtherB-CPBP' category seems to slightly reduce at first when the size of the bank increases. However, the decrease does not seem to be significant and for very large banks average loss eventually starts to increase.



**Figure 7.** QQ plot of Simulated Losses vs. Observed Losses for Bank A. Note: quantiles for the observed losses have been computed using the weights estimated in subsection 4.2.



**Figure 8.** QQ plot of Simulated Losses vs. Observed Losses for Bank B. Note: quantiles for the observed losses have been computed using the weights estimated in subsection 4.2.

between fraud related loss severities and the size of a bank. One possible explanation for this phenomenon is that as banks become larger they will have access to more sophisticated IT systems, internal controls, and audit methods which can reduce fraud, leading to a decreasing relationship between average loss severity between fraud and the size of a bank.

However, although the mean loss may decrease for some BL/ET categories, we find loss at the 99<sup>th</sup> percentile increases for all BL/ET combinations as the size of a bank increases (see table 15). The results in table 16 show that for certain BL/ET combinations 99<sup>th</sup> percentile loss can increase by up to 25% to 35% when the size of a bank doubles. This is an important result as it proves (contrary to previous studies) that the size of a bank is in fact an important determinant of the severity of operational losses. In particular, the fitted model suggests that large banks are exposed to more tail risk than small banks.



**Table 11.** Summary Statistics of the Simulation for Bank A and B.

| Percentile       | Bank A                      |                              | Bank B                      |                              |
|------------------|-----------------------------|------------------------------|-----------------------------|------------------------------|
|                  | Observed (\$M)<br>[95 % CI] | Simulated (\$M)<br>[95 % CI] | Observed (\$M)<br>[95 % CI] | Simulated (\$M)<br>[95 % CI] |
| 25 <sup>th</sup> | 1.3<br>[1.1, 1.6]           | 1.4<br>[1.4, 1.4]            | 1.3<br>[1.1, 1.5]           | 1.3<br>[1.3, 1.3]            |
| 50 <sup>th</sup> | 2.0<br>[1.6, 2.4]           | 2.1<br>[2.1, 2.1]            | 1.9<br>[1.4, 2.4]           | 1.9<br>[1.9, 1.9]            |
| 75 <sup>th</sup> | 3.3<br>[2.5, 5.3]           | 3.7<br>[3.7, 3.7]            | 2.7<br>[2.1, 5.9]           | 3.5<br>[3.5, 3.5]            |
| 90 <sup>th</sup> | 8.7<br>[6.1, 13.2]          | 7.9<br>[7.8, 7.9]            | 6.7<br>[3.7, 15]            | 8.0<br>[7.9, 8]              |
| 95 <sup>th</sup> | 14.9<br>[11.8, 18.9]        | 13.9<br>[13.8, 13.9]         | 14.5<br>[6.6, 25.2]         | 15.1<br>[15, 15.2]           |
| 97 <sup>th</sup> | 20.1<br>[15, 27.1]          | 21.0<br>[20.8, 21.2]         | 24.0<br>[10.5, 37.9]        | 24.4<br>[24.1, 24.6]         |
| 99 <sup>th</sup> | 51.9<br>[26.2, 118.1]       | 52.2<br>[51.3, 53.1]         | 56.3<br>[37.9, 227.4]       | 70.3<br>[69, 71.7]           |

Note: percentiles for the observed losses have been computed using the weights estimated in subsection 4.2.

**Table 12.** Re-estimated Parameter Values of Model G after Removing Bank A.

| Explanatory Variable                       | $\mu$    |          | $\sigma$ |          |
|--|----------|----------|----------|----------|
|  | Estimate | St. Err. | Estimate | St. Err. |
| Intercept                                  | 2.012    | (0.901)  | 4.234    | (0.798)  |
| Log-assets                                 | -0.135   | (0.118)  | -0.290   | (0.061)  |
| Equity Ratio                               | 4.267    | (0.065)  | -        | -        |
| Corporate Finance                          | -0.330   | (1.194)  | -        | -        |
| Trading and Sales                          | -        | -        | -        | -        |
| Retail Banking                             | -        | -        | -        | -        |
| Commercial Banking                         | -0.369   | (0.113)  | -0.197   | (0.149)  |
| PS, AS, AM                                 | -0.268   | (0.052)  | -        | -        |
| Internal fraud                             | -        | -        | -        | -        |
| External fraud                             | -        | -        | -        | -        |
| Clients, Products and Business Practices   | -0.099   | (0.135)  | -        | -        |
| Execution, Delivery and Process Management | -        | -        | -        | -        |

Note: the model is fitted using retail brokerage as the baseline line category for business lines and ‘EPWS, DPA, BDSF’ as the baseline category for event types.

Another way of quantifying the size effect is to compute the tail index implied by the fitted model for banks of various sizes. The results presented in table 17 show that the tail index decreases for all BL/ET combinations as the firm size increases. For almost all BL/ET combinations the tail index is less than 2. That is, the variance of the loss distribution is infinite (which is usually considered as a characteristic of a heavy-tailed distribution). In addition, we find that for very large banks the tail index becomes less than 1 for certain BL/ET categories. That is, the losses follow an extremely heavy-tailed distribution and hence the mean is infinite. Losses which follow an

**Table 13.** Re-estimated Parameter Values of Model G after Removing Bank B.

| Explanatory Variable                       | $\mu$    |          | $\sigma$ |          |
|--|----------|----------|----------|----------|
|  | Estimate | St. Err. | Estimate | St. Err. |
| Intercept                                  | 3.194    | (0.893)  | 5.214    | (0.781)  |
| Log-assets                                 | -0.223   | (0.117)  | -0.365   | (0.06)   |
| Equity Ratio                               | 3.810    | (0.065)  | -        | -        |
| Corporate Finance                          | -0.326   | (1.135)  | -        | -        |
| Trading and Sales                          | -        | -        | -        | -        |
| Retail Banking                             | -        | -        | -        | -        |
| Commercial Banking                         | -0.410   | (0.112)  | -0.294   | (0.157)  |
| PS, AS, AM                                 | -0.247   | (0.053)  | -        | -        |
| Internal fraud                             | -        | -        | -        | -        |
| External fraud                             | -        | -        | -        | -        |
| Clients, Products and Business Practices   | -0.139   | (0.133)  | -        | -        |
| Execution, Delivery and Process Management | -        | -        | -        | -        |

Note: the model is fitted using retail brokerage as the baseline line category for business lines and 'EPWS, DPA, BDSF' as the baseline category for event types.

**Table 14.** Estimated Average Loss (\$ mill).

| Business Line | Event Type | Total Assets (\$ bill) |      |       |       |       |      |      |
|---------------|------------|------------------------|------|-------|-------|-------|------|------|
|               |            | 200                    | 300  | 400   | 500   | 600   | 700  | 2000 |
| CF            | CPBP       | 33.1                   | 43.7 | 65.1  | 122.8 | 513.0 | NA   | NA   |
| CF            | OtherE     | 12.6                   | 13.1 | 14.0  | 15.1  | 16.6  | 18.5 | NA   |
| CB            | CPBP       | 24.7                   | 42.1 | 146.7 | NA    | NA    | NA   | NA   |
| CB            | OtherE     | 9.0                    | 9.8  | 10.9  | 12.5  | 14.8  | 18.6 | NA   |
| PS, AS, AM    | CPBP       | 16.2                   | 17.6 | 19.7  | 22.6  | 26.7  | 33.0 | NA   |
| PS, AS, AM    | OtherE     | 8.3                    | 8.2  | 8.3   | 8.5   | 8.8   | 9.0  | 23.1 |
| OtherB        | CPBP       | 5.7                    | 5.5  | 5.4   | 5.4   | 5.4   | 5.4  | 6.5  |
| OtherB        | OtherE     | 4.0                    | 3.9  | 3.8   | 3.7   | 3.7   | 3.7  | 3.7  |

Note 1 : CF – 'Corporate finance', CB – 'Commercial banking', PS, AS, AM – grouped business lines of 'Payment and Settlement', 'Agency Services', and 'Asset Management', OtherB – all other business lines, CPBP – 'Clients products and business practices', OtherE – all other event types.

Note 2 : All values have been computed by assuming an equity ratio of 6.7%.

Note 3 : NA refers to situations with infinite means.

extremely heavy-tailed distribution can have serious implications on capital calculations as well as risk management for banks. Nešlehová *et al.* (2006) provide a detailed discussion on these issues and the challenges of modelling operational risk with extremely heavy tailed behaviour. The following discussion draws on their paper.

Nešlehová *et al.* (2006) point out that for loss distributions which follow a Pareto-type tail behaviour (i.e., as in eq. (7)), the Value-at-Risk (VaR) grows exponentially as  $\alpha$  decrease. Therefore, when  $\alpha$  tends to zero the capital charge computed by VaR could become ridiculously high. The fitted model suggests that use of  $\text{VaR}_\gamma$  can become problematic for large banks since losses in some BL/ET combinations may be extremely heavy-tailed. A one possible solution for this problem is to

**Table 15.** Estimated Loss at the 99<sup>th</sup> Percentile (\$ mill).

| Business Line | Event Type | Total Assets (\$ bill) |      |      |      |      |      |      |
|---------------|------------|------------------------|------|------|------|------|------|------|
|               |            | 200                    | 300  | 400  | 500  | 600  | 700  | 2000 |
| CF            | CPBP       | 2139                   | 2354 | 2547 | 2724 | 2890 | 3047 | 4694 |
| CF            | OtherE     | 747                    | 811  | 868  | 920  | 968  | 1013 | 1471 |
| CB            | CPBP       | 1396                   | 1578 | 1739 | 1886 | 2023 | 2153 | 3519 |
| CB            | OtherE     | 517                    | 574  | 624  | 670  | 712  | 751  | 1147 |
| PS, AS, AM    | CPBP       | 1030                   | 1123 | 1206 | 1281 | 1352 | 1418 | 2096 |
| PS, AS, AM    | OtherE     | 397                    | 428  | 455  | 480  | 503  | 524  | 734  |
| OtherB        | CPBP       | 200                    | 214  | 226  | 236  | 246  | 255  | 344  |
| OtherB        | OtherE     | 97                     | 102  | 107  | 112  | 116  | 119  | 154  |

Note 1 : CF – ‘Corporate finance’, CB – ‘Commercial banking’, PS, AS, AM – grouped business lines of ‘Payment and Settlement’, ‘Agency Services’, and ‘Asset Management’, OtherB – all other business lines, CPBP – ‘Clients products and business practices’, OtherE – all other event types.

Note 2 : All values have been computed by assuming an equity ratio of 6.7%.

**Table 16.** The increase in the 99<sup>th</sup> percentile loss when the size of the bank doubles.

| Business Line | Event Type | Increase in Total Assets |                          |                           |
|---------------|------------|--------------------------|--------------------------|---------------------------|
|               |            | \$200 bill to \$400 bill | \$400 bill to \$800 bill | \$800 bill to \$1600 bill |
| CF            | CPBP       | 19.1%                    | 25.5%                    | 32.5%                     |
| CF            | OtherE     | 16.2%                    | 21.7%                    | 27.5%                     |
| CB            | CPBP       | 24.5%                    | 30.9%                    | 37.8%                     |
| CB            | OtherE     | 20.8%                    | 26.2%                    | 31.9%                     |
| PS, AS, AM    | CPBP       | 17.1%                    | 22.8%                    | 29.0%                     |
| PS, AS, AM    | OtherE     | 14.6%                    | 19.4%                    | 24.6%                     |
| OtherB        | CPBP       | 12.8%                    | 17.0%                    | 21.5%                     |
| OtherB        | OtherE     | 11.0%                    | 14.5%                    | 18.3%                     |

Note 1 : CF – ‘Corporate finance’, CB – ‘Commercial banking’, PS, AS, AM – grouped business lines of ‘Payment and Settlement’, ‘Agency Services’, and ‘Asset Management’, OtherB – all other business lines, CPBP – ‘Clients products and business practices’, OtherE – all other event types.

Note 2 : All values have been computed by assuming an equity ratio of 6.7%.

implement a minimum probability threshold as suggested by Evans (2001). The author states that “a reasonable value judgment may be that events which have less than a 1 in 1,000,000 chance of happening should simply be ignored”. However, for the fitted model we find a higher probability limit of 1 in 10,000 would be required to obtain capital estimates consistent with the industry standards for very large banks. We shall revisit this issue in detail later in section 5 when we compute capital estimates.

Here, one might be tempted to argue that infinite mean models should not be used to model operational losses since the total capital existing is finite. Nešlehová *et al.* (2006) point out that one shouldn’t shy away from infinite mean models as financial literature is full of power law models which have some infinite moments, and yet they are used frequently. However, authors do warn that a transition from infinite second moment to infinite first moment requires careful handling.

Table 17. Estimated Tail Indices.

| Business Line | Event Type | Total Assets (\$ bill) |       |       |       |       |       |       |
|---------------|------------|------------------------|-------|-------|-------|-------|-------|-------|
|               |            | 200                    | 300   | 400   | 500   | 600   | 700   | 2000  |
| CF            | CPBP       | 1.240                  | 1.152 | 1.093 | 1.049 | 1.015 | 0.987 | 0.815 |
| CF            | OtherE     | 1.437                  | 1.335 | 1.267 | 1.216 | 1.177 | 1.144 | 0.945 |
| CB            | CPBP       | 1.158                  | 1.076 | 1.021 | 0.980 | 0.948 | 0.922 | 0.761 |
| CB            | OtherE     | 1.342                  | 1.247 | 1.183 | 1.136 | 1.099 | 1.068 | 0.882 |
| PS, AS, AM    | CPBP       | 1.371                  | 1.273 | 1.208 | 1.160 | 1.122 | 1.091 | 0.901 |
| PS, AS, AM    | OtherE     | 1.589                  | 1.476 | 1.400 | 1.344 | 1.301 | 1.264 | 1.044 |
| OtherB        | CPBP       | 1.795                  | 1.667 | 1.582 | 1.519 | 1.469 | 1.428 | 1.180 |
| OtherB        | OtherE     | 2.080                  | 1.932 | 1.833 | 1.760 | 1.703 | 1.656 | 1.367 |

Note 1 : CF – ‘Corporate finance’, CB – ‘Commercial banking’, PS, AS, AM – grouped business lines of ‘Payment and Settlement’, ‘Agency Services’, and ‘Asset Management’, OtherB – all other business lines, CPBP – ‘Clients products and business practices’, OtherE – all other event types.

Note 2 : All values have been computed by assuming an equity ratio of 6.7%.

The second issue of infinite mean models is related to diversification of risk. Often in practice the operational risk capital charge is determined by computing  $VaR_\gamma$  for each business line and adding them up as

$$VaR_{99.9\%}(L) = \sum_{k=1}^K VaR_{99.9\%}(L_k)$$

where,  $L$  is the total loss random variable of the bank, and  $L_k$  are the total loss random variables for business lines  $k = 1, \dots, K$ . Then, if the bank can demonstrate a reduction in risk due to correlation, a capital reduction would be allowed from the latter. As per our findings this particular method of capital calculation creates several problems. Firstly, we find operational losses are heavy-tailed (i.e.  $\alpha < 2$ ) for most BL/ET combinations as reported in table 17. In this context a diversification reduction based on correlation makes no sense since correlation coefficients are not defined when  $\alpha < 2$ . Secondly, for large banks we find  $\alpha$  can become less than 1 for certain BL/ET combinations. Under these conditions (i.e.  $\alpha < 1$ )  $VaR_\gamma$  is no longer a *coherent risk measure*<sup>12</sup> since the subadditivity property

$$VaR_\gamma(L) = VaR_\gamma\left(\sum_{k=1}^K L_k\right) \leq \sum_{k=1}^K VaR_\gamma(L_k)$$

will no longer hold. Rootzén & Klüppelberg (1999) and Nešlehová *et al.* (2006) provide several examples of situations in which subadditivity property is violated by  $VaR$  under extremely heavy-tailed distributions. In particular, Nešlehová *et al.* (2006) show that for independent business lines having Pareto-type risks (i.e. a tail behaviour of the form eq. (7)) with  $0 < \alpha < 1$

$$VaR_\gamma\left(\sum_{k=1}^K L_k\right) > \sum_{k=1}^K VaR_\gamma(L_k)$$

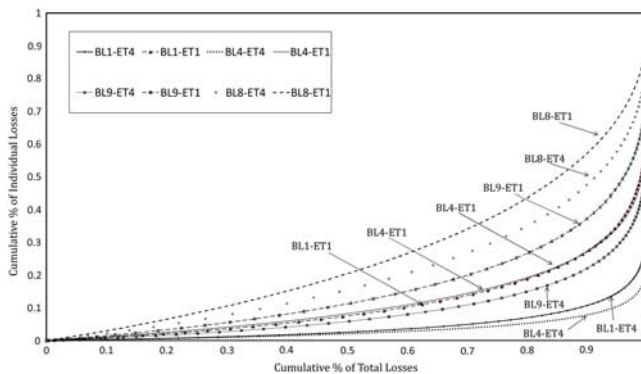
for sufficiently large  $\gamma$ . In other words, when losses are distributed as in an infinite mean model diversification no longer exists in the traditional sense under the  $VaR_\gamma$  measure. The authors state

<sup>12</sup> A risk measure is said to be coherent if it satisfies the properties of monotonicity, translational invariance, homogeneity, and subadditivity (Artzner *et al.*, 1999).

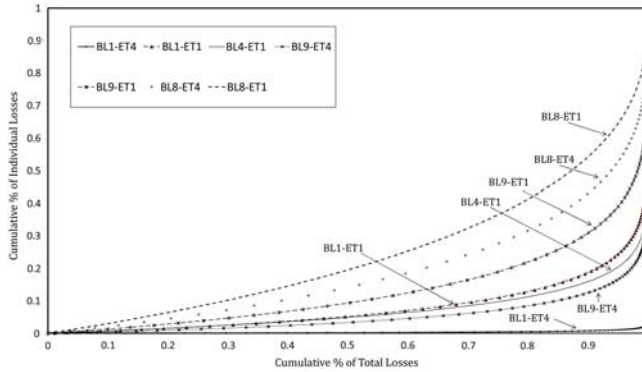
that "...in the presence of extremely heavy-tailed loss distributions, standard economic thinking may have to be reconsidered" and we "...have to accept that a VAR-based risk-capital calculation for operational risk is not yet fully understood". The tail index estimates given in table 17 indicates that the banks with assets greater than \$500 bill should take extra concern about these issues.

Usually subadditivity is considered an important property for a risk measure as it promotes diversification. However, as has been pointed out by Rootzén & Klüppelberg (1999) subadditivity does not necessarily need to hold for catastrophic risk. By using a simple example authors show that combining two catastrophic risks will increase the probability of ruin rather than decrease it. The authors state that "big is not always beautiful" and "catastrophic financial risks don't become smaller if collected in one big company". Hence, non-subadditivity of  $\text{VaR}_\alpha$  when  $\alpha < 1$  should not be considered as a limitation of the risk measure. However, risk managers should keep in mind that when operational losses follow extremely heavy-tailed distributions, simply adding up the capital estimates for each business lines would not provide a conservative estimate as it would provide in most other types of risk. This issue would be particularly important to banks with assets greater than \$500 bill in which tail index of operational losses can become less than 1 as seen from table 17.

The final point we would like to raise is the *one-loss-causes-ruin* problem of operational risk. The one-loss-causes-ruin problem refers to the phenomena that a significant portion of total losses are caused by few individual losses. This phenomena can be illustrated by the use of a Lorenz curve. The concept of a Lorenz curve is used in economics to analyze inequality of income distribution (and wealth) in a population. A related concept to the Lorenz curve is Pareto's famous 80:20 law which states "80% of the all effects results from 20% of the causes". The 80:20 law has been observed in many economic and natural processes and is referred to as *law of the vital few*. (see Ultsch, 2002, to find many situations in which Pareto's 80:20-law is reported). The 80:20 law applies when outcomes of interest follow an exact Pareto distribution with  $\alpha \approx 1.16096$ . If operational losses follow the Pareto law, then 80% of the total loss would be caused by 20% of the individual losses. Hence, the threat to a bank's solvency would be mainly driven by those 20% large individual losses. To see whether such a pattern exists in operational losses, we computed the implied Lorenz curves by the fitted model for each BL/ET combination for two hypothetical banks with \$300 bill assets and \$600 bill assets respectively, using eq. (12). The plots of the resulting curves are given in figure 9 and figure 10. By comparing the two plots



**Figure 9.** The Lorenz curves of the loss distributions of a bank with 300 billion assets. Note: an equity ratio 6.7% has been assumed to compute the Lorenz curve.



**Figure 10.** The Lorenz curves of the loss distributions at a bank with 600 bill assets. Note: an equity ratio 6.7% has been assumed to compute the Lorenz curve.

**Table 18.** Cumulative % of individual losses as the % of total losses.

| Business Line | Event Type | % of Total Loss |        |        |
|---------------|------------|-----------------|--------|--------|
|               |            | 80%             | 90%    | 99%    |
| CF            | CPBP       | 7.00%           | 10.79% | 24.16% |
| CF            | OtherE     | 18.93%          | 27.45% | 51.89% |
| CB            | CPBP       | 5.26%           | 7.86%  | 16.80% |
| CB            | OtherE     | 19.31%          | 27.30% | 49.73% |
| PS, AS, AM    | CPBP       | 15.04%          | 22.22% | 44.09% |
| PS, AS, AM    | OtherE     | 26.77%          | 37.55% | 64.86% |
| OtherB        | CPBP       | 35.18%          | 47.73% | 75.62% |
| OtherB        | OtherE     | 43.70%          | 57.27% | 83.74% |

Note: an equity ratio of 6.7% has been assumed to compute the above values.

one could observe that the percentage of total loss caused by few extreme losses is high at the larger bank. To illustrate this further, we have reported values of the Lorenz curve for selected percentages in table 18 and table 19. The value given in the first row of the third column of table 18 indicates that for a bank with \$300 bill assets 80% of the total loss due to CF-CLBT is caused by 7% of the individual CF-CLBT losses. In contrast, the third row of the first column in table 19 indicates that for a bank with \$600 bill assets, 80% of the total loss due to CF-CLBT category is caused by only 0.5% of the individual CF-CLBT losses. Similar results hold for CB-CLBT combination as well. Moreover, the first row of the fifth column in table 19 indicates that for a bank with \$600 bill assets, 99% of the total loss due to CF-CLBT category is caused by only 1.83% of the individual CF-CLBT losses. This is an important result which should serve as a warning to large banks about the heavy-tailedness of the risk in these particular BL/ETs. Our results suggest that the one-loss-causes-ruin problem is relevant for small as well as for large banks, but it is much more severe for large banks. The complexity of banking activity at large banks can probably be the reason for this. Furthermore, results suggest that the problem seems to be rooted in the CPBP event category. Hence, more care should be given when modelling this type of risk.

**Table 19.** Cumulative % of individual losses as the % of total losses.

| Business Line | Event Type | % of Total Loss |        |        |
|---------------|------------|-----------------|--------|--------|
|               |            | 80%             | 90%    | 99%    |
| CF            | CPBP       | 0.50%           | 0.77%  | 1.83%  |
| CF            | OtherE     | 12.85%          | 18.70% | 36.80% |
| CB            | CPBP       | NA              | NA     | NA     |
| CB            | OtherE     | 11.23%          | 15.90% | 30.07% |
| PS, AS, AM    | CPBP       | 8.43%           | 12.48% | 25.94% |
| PS, AS, AM    | OtherE     | 21.93%          | 30.84% | 55.01% |
| OtherB        | CPBP       | 31.65%          | 43.00% | 69.76% |
| OtherB        | OtherE     | 41.25%          | 54.17% | 80.52% |

Note: an equity ratio of 6.7% has been assumed to compute the above values.

### 4.6.2 The Effect of Equity Ratio on Operational Loss Severity

In section 4.1.4 we hypothesized that the equity ratio can be used as a proxy to the risk culture within a bank. At this point, we shall test the validity of that hypothesis by analyzing how equity ratio affects the operational loss severity as suggested by the fitted model.

Table 20 and table 21 provides the fitted mean and the 99.9<sup>th</sup> percentile loss for banks with \$500 bill assets and an equity ratio of 5%, 6.7%, and 10%. Both tables indicate that the severity of operational losses are high for the bank with a low equity ratio. In particular, we find severity at the 99.9<sup>th</sup> percentile could increase by around 40% to 70% (for certain BL/ET up to 75%, e.g., CF-CPBP) when the equity ratio drops from the median ratio of 6.7% to 5%. Similarly, when the equity ratio increases from 6.7% to 10% losses at the 99.9<sup>th</sup> percentile drop by around 40% to 60%.

The above results suggest that our initial hypothesis that the equity ratio may serve as a proxy for the risk culture of a bank is valid. However, a word of caution is required here as the hypothesis may not hold universally for all banks, since the equity ratio is not a direct measure of the risk culture of a firm. Usually a low equity ratio does not necessarily mean a firm is too risky. It could well mean that the management is efficient in investing the equity capital. Similarly a too high equity ratio could be a sign of management wasting resources.

We conclude this section by providing a brief summary of the main inferences we gathered from the fitted model and the issues we have raised so far. As discussed earlier, the estimated parameters of the model suggest that the size of a bank is an important indicator of its operational loss severities. In particular, model indicate there is a considerable variation in the tail index of the loss distribution across banks of different sizes. This is an important result which may have a significant impact in terms of capital calculation as well as risk management. As we have pointed out, the tail index for certain BL/ET combinations can become less than 1 for very large banks. Although capital can be computed using the VaR method under these conditions, the computed capital can be too high. A possible solution for this is to implement a minimum probability threshold. The second issue one needs to be careful of is that VaR is no longer subadditive when  $\alpha < 1$ . Hence, one should not compute the total required capital by adding up the VaR estimates of each business line. We find these particular issues are more relevant to large banks than to small banks as the tail index for small banks tends to be in the region of  $\alpha > 2$ . Furthermore, the estimated parameters suggest that

**Table 20.** Estimated Average Loss (\$ mill).

| Business Line | Event Type | Equity Ratio |       |      |
|---------------|------------|--------------|-------|------|
|               |            | 5%           | 6.7%  | 10%  |
| CF            | CPBP       | NA           | 122.8 | 17.1 |
| CF            | OtherE     | 28.0         | 15.1  | 7.4  |
| CB            | CPBP       | NA           | NA    | 14.5 |
| CB            | OtherE     | 29.5         | 12.5  | 5.8  |
| PS, AS, AM    | CPBP       | 55.9         | 22.6  | 9.2  |
| PS, AS, AM    | OtherE     | 12.2         | 8.5   | 5.3  |
| OtherB        | CPBP       | 6.8          | 5.4   | 3.9  |
| OtherB        | OtherE     | 4.4          | 3.7   | 3.0  |

**Table 21.** Estimated Loss at the 99<sup>th</sup> Percentile (\$ mill).

| Business Line | Event Type | Equity Ratio |      |      |
|---------------|------------|--------------|------|------|
|               |            | 5%           | 6.7% | 10%  |
| CF            | CPBP       | 4742         | 2724 | 1028 |
| CF            | OtherE     | 1484         | 920  | 397  |
| CB            | CPBP       | 3199         | 1886 | 745  |
| CB            | OtherE     | 1056         | 670  | 301  |
| PS, AS, AM    | CPBP       | 2116         | 1281 | 531  |
| PS, AS, AM    | OtherE     | 740          | 480  | 224  |
| OtherB        | CPBP       | 347          | 236  | 121  |
| OtherB        | OtherE     | 155          | 112  | 62   |

both small as well as large banks are subject to the one-loss-causes-ruin problem, but the exposure is severe for large banks. Particularly, this raises the importance of scenario analysis as well as having a sound crisis management system in place to handle catastrophic events.

## 5 Capital Calculation

This section provides a demonstration of the severity model at work by computing capital estimates for several hypothetical banks as well as for two real banks by making simple assumptions about the frequency distribution. The results show that the capital estimates are reasonable and consistent with the industry standards.

### 5.1 Capital Estimates for Hypothetical Banks

In order to compute capital estimates one would need a frequency distribution for operational losses. Here, we make the simple assumption that the loss frequency in each BL/ET category is distributed as a Poisson random variables with a density function

$$Pr(N_j = n_j) = \frac{\lambda_j^{n_j} e^{-\lambda_j}}{n_j!}$$



where  $n_j$  is the number of losses above \$ 1million and  $\lambda_j$  is the Poisson parameter of the  $j^{th}$  BL/ET category for  $j = 1, \dots, 30$ . The maximum likelihood of parameter of  $\lambda_j$  is the sample mean of the loss frequency in BL/ET category  $j$ . This makes the Poisson assumption particularly attractive in the current context, since estimation becomes very easy as data on average frequency of losses is publicly available. However, a limitation of the Poisson distribution is that its variance is equal to its mean. Hence, other distributions such as Negative Binomial would be more suitable if one needs to allow for more variation in the frequency model. However, it is a well known fact that for compound distributions of the form  $S = \sum_{i=1}^N X_i$ , where,  $X_i$  are subexponential severities with distribution function  $F$  and the frequency distribution given by  $Pr(N = n) = p(n)$ , if there exist a constant  $\omega > 0$  such that

$$\sum_{n=0}^{\infty} (1 + \omega)^n p(n) < \infty \tag{13}$$

then the aggregate loss distribution  $G$  is subexponential with tail behaviour given by

$$\tilde{G}(s) = E(N) \tilde{F}(s) \tag{14}$$

In other words, for heavy tailed losses, the tail of the compound distribution is simply the expected number losses times the tail of the severity given that frequency distribution satisfies the condition given in eq. (13)<sup>13</sup>. This means capital requirement for heavy-tailed risk such as operational risk is mainly driven by the expected number of losses, and therefore over-dispersion is not a major concern. For this reason as well as for its simplicity the Poisson distribution remains a common distributional assumption for modelling operational loss frequencies (see for example De Fontnouvelle *et al.*, 2006; De Fontnouvelle *et al.*, 2007; Medova & Berg-Yuen, 2009). Thus, we believe the assumption is reasonable for the purpose of demonstrating the severity model.

In practice, banks are required to model the dependence structure when computing capital estimates. It is typical to assume dependence between frequencies while severities are assumed independent. Due to data limitations we are unable to model the dependence between frequencies of each business line-event type combination. Hence, we assume that the frequency distribution of each BL/ET combination is independent. Although this is a strong assumption, Basel (2009) reports that modelling of dependence at a typical bank results in a modest 8.3% increase in the economic capital relative to the assumption of full independence. Therefore, we argue that the independence assumption will not affect the results significantly.

Given our assumption that the frequency distribution of each BL/ET combination is independent, the distribution of total frequency  $N = \sum_{j=1}^{30} n_j$  has a Poisson distribution with a parameter  $\Lambda = \sum_{j=1}^{30} \lambda_j$ . This is a direct consequence of the property that the sum of independent Poisson random variables also has a Poisson distribution. To calibrate the parameter  $\Lambda$ , we used the results published by the Basel Committee on the 2008 Loss Data Collection Exercise (LDCE) (Basel, 2009). Table ILD9 in that paper reports that a typical bank experienced 0.013 losses per year of €1 mill (US \$1.58 mill) or more for each €1 bill (US \$1.58 bill) in assets. Hence, it is reasonable to approximate  $\hat{\Lambda} = 0.00823$  for each US \$1 bill in assets. Then, the parameter values of  $\lambda_j$  can be computed by

$$\hat{\lambda}_j = \frac{\sum w_j'}{n} \hat{\Lambda}$$

where,  $\sum w_j'$  is the sum of normalized weights given by eq. (6) for the  $j^{th}$  BL/ET combination and  $n$  is the total number of observations in the data sample. Table 22 provides the estimated values of  $\hat{\lambda}_j$  for a bank with \$100 bill assets.

<sup>13</sup> In fact it can be shown that the result given in (14) holds for both Poisson distribution as well as for Negative Binomial distribution (see examples 1.3.10 and 1.3.11 of Embrechts *et al.*, 1997).

**Table 22.** Estimates of  $\hat{\lambda}_j$  for a bank with \$100 bill assets.

|                    | Internal fraud | External fraud | Clients, Products and Business Practices | Execution, Delivery and Process Management | EPWS, DPA, BDSF |
|--------------------|----------------|----------------|--|--|-----------------|
| Corporate Finance  | 5.1E-03        | 1.9E-03        | 4.4E-02                                  | 9.5E-03                                    | 5.9E-03         |
| Trading and Sales  | 9.2E-03        | 5.7E-03        | 1.0E-01                                  | 2.0E-02                                    | 1.5E-02         |
| Retail Banking     | 4.9E-02        | 6.3E-02        | 3.6E-02                                  | 1.0E-02                                    | 1.6E-02         |
| Commercial Banking | 1.9E-02        | 5.4E-02        | 2.8E-02                                  | 4.8E-03                                    | 9.7E-03         |
| Retail Brokerage   | 4.6E-02        | 6.1E-03        | 1.2E-01                                  | 1.3E-02                                    | 5.3E-02         |
| PS, AS, AM         | 1.3E-02        | 6.9E-03        | 4.6E-02                                  | 7.4E-03                                    | 2.4E-03         |

**Table 23.** Capital Estimates (in \$ mill) for Hypothetical Banks.

| Assumption | Total Assets (\$ bill) |       |       |       |       |       |        |
|------------|------------------------|-------|-------|-------|-------|-------|--------|
|            | 200                    | 300   | 400   | 500   | 600   | 700   | 2000   |
| ER = 5%    | 953                    | 1,479 | 2,159 | 2,783 | 3,503 | 4,464 | 27,714 |
| ER = 6.7%  | 611                    | 925   | 1,312 | 1,663 | 2,061 | 2,588 | 14,250 |
| ER = 10%   | 280                    | 405   | 550   | 674   | 814   | 996   | 4,424  |
| LDCE       | 760                    | 1,140 | 1,520 | 1,900 | 2,280 | 2,660 | 7,600  |

Note: The first three rows of the table gives the capital estimates provided by the severity model G and the frequency model presented in section 5 using 10mill years of simulations. The estimated values are the amount of capital needed to cover losses exceeding \$1mill. Additional capital may be required to cover losses less than \$1mill if the bank has not priced them into their fees and charges. The last row provides the median (total) operational risk capital held by a typical bank in North America as implied by the ratios given in table C1 of Basel's LDCE report (2009).

We computed the operational risk capital requirement for 21 hypothetical banks of various sizes and equity ratios by using the above frequency model and the severity model G of section 4.3. The capital estimates for the 21 banks are given in table 23. The results presented in the second row of table 23 have been computed using the median equity ratio of 6.7%. Hence, the results given in the second row can be considered as the required operational risk capital to cover losses above \$1 mill for a "typical" bank of a given size, as suggested by the fitted model. The 2008 LDCE of the Basel Committee reports that the median ratio between the reported regulatory operational risk capital and the total assets for a "typical" North American bank is 0.38% (Basel, 2009). The last row of table 23 presents benchmark capital estimates based on this ratio. One could use these benchmark estimates to compare the results presented in the second row of the table to test the reasonability of the capital estimates provided by the fitted model. However, it should be noted that the benchmark capital estimate is the required capital to cover the total operational risk of a bank, as implied by the LDCE report, whereas the capital estimates of the first three rows in the table provide the required amount of capital to cover losses exceeding \$1 mill.

One can observe from table 24 that the capital estimates given by the fitted model for a "typical" bank with assets ranging from \$200 bill to \$700 bill is around 3% to 20% less than the industry benchmark. This is expected, given that our estimates are the amount of capital needed to cover

**Table 24.** Capital Estimates (in \$ mill) for Hypothetical Banks after implementing a 1-in-10,000 year probability limit.

| Assumption | Total Assets (\$ bill) |       |       |       |       |       |        |
|------------|------------------------|-------|-------|-------|-------|-------|--------|
|            | 200                    | 300   | 400   | 500   | 600   | 700   | 2000   |
| ER = 5%    | 953                    | 1,479 | 2,140 | 2,678 | 3,380 | 4,024 | 17,462 |
| ER = 6.7%  | 611                    | 925   | 1,312 | 1,641 | 2,023 | 2,477 | 9,479  |
| ER = 10%   | 280                    | 405   | 550   | 674   | 814   | 996   | 3,925  |
| LDCE       | 760                    | 1,140 | 1,520 | 1,900 | 2,280 | 2,660 | 7,600  |

Note: The first three rows of the table gives the capital estimates provided by the severity model G and the frequency model presented in section 5 using 10mill years of simulations. The estimated values are the amount of capital needed to cover losses exceeding \$1mill. Additional capital may be required to cover losses less than \$1mill if the bank has not priced them into their fees and charges. The last row provides the median (total) operational risk capital held by a typical bank in North America as implied by the ratios given in table C1 of Basel's LDCE report (2009).

losses above \$1 mill, whereas the benchmark capital estimate is the amount of capital needed to cover the total operational risk. Furthermore, our simplified assumption that the frequencies of each BL/ET are independent, would reduce our capital estimates by around 8.3% as explained earlier. Given these facts, we conclude that the capital estimates provided by our model for the first six banks (i.e. banks with assets of \$200 bill to \$700 bill) are reasonable as well as consistent with industry standards. However, the estimated capital for a bank with \$2000 bill assets by our model is \$ 14.25 bill, which is 1.87 times larger than the benchmark estimate. Although the estimate is significantly higher than the benchmark, it is in no means an unrealistic estimate. The ratios given in table C1 of LDCE report (Basel, 2009) suggest that the 75<sup>th</sup> percentile of the regulatory capital reported by a bank with \$2000 bill assets was in the range of \$11.6 bill to \$16 bill.

The reason behind our capital estimates for large banks being significantly high are the infinite mean models we obtained earlier for certain BL/ET categories in those banks. As discussed in subsection 4.6.1 a solution for high capital estimates yield by infinite mean models is to use a probability limit. We tested several probability limits and found a limit of 1-in-10,000 years would yield capital estimates which are consistent with the industry benchmark. The capital estimates after implementing the probability limit is given in table 24. Note that the probability limit is only implemented when simulating losses from infinite mean models. Hence, capital estimates do not change for small banks in which severity models for all BL/ET categories had tail indices greater than one.

## 5.2 Backtesting for Two US Banks

In practice regulators require banks to backtest their operational risk models to test whether the estimated capital from the model is sufficient to cover past loss experience. To see how our model performs against such a test we carried out backtesting for Banks A and B of subsection 4.5. The backtesting was carried out using the frequency model described in subsection 5.1 and the severity model given in table 12 and table 13.

The plots of the logarithm of estimated capital and the total log-losses for bank A and B are given in figure 11 and figure 12. The Y-axis of the plots has not been labelled to protect anonymity of the banks. Both plots indicate that the estimated capital is sufficient to cover losses above \$1 mill.

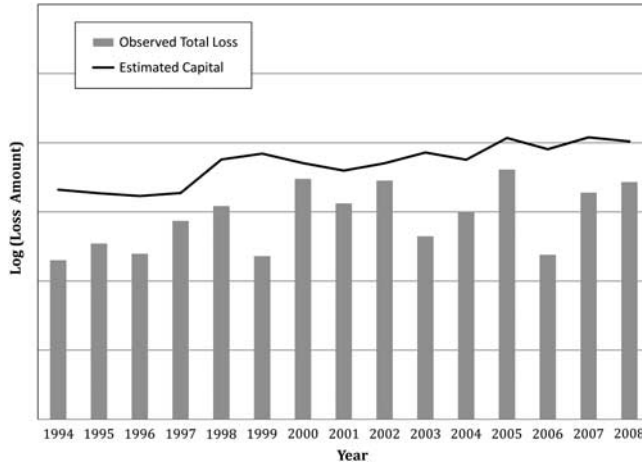


Figure 11. Backtesting Results for Bank A.

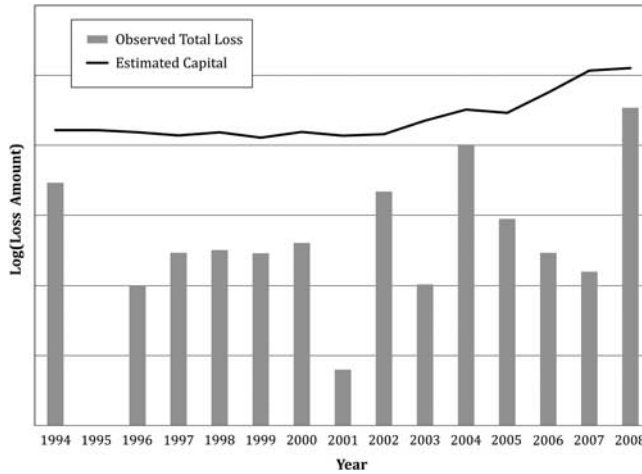


Figure 12. Backtesting Results for Bank B.

Furthermore, we find that under the null hypothesis of model being ‘correct’, non-rejection region for number of exceptions ( $X$ ) based on the Kupiec’s POF test (Kupiec, 1995) is  $0 \leq X \leq 2$  for the 95% confidence level and  $0 \leq X \leq 1$  for the 99% confidence level. The results of the backtest are well within the non-rejection regions for the both confidence levels. Hence, we conclude that the backtesting results further validate our severity model.

## 6 Conclusion and Discussion

One of the major challenges of quantifying operational risk for banks is addressing the issue of lack of data. Most of the large operational losses are infrequent and therefore it is impossible to model these losses by solely relying on bank level data. As a solution to this Basel II requires banks to use

industry wide external data to complement their internal data. However, combining internal data with external data raises further issues. In this paper we have demonstrated how external data collected from various banks of different sizes and different unknown loss reporting thresholds can be combined together to quantify the operational risk of a bank of a given size.

Two of the main challenges in combining data from different banks is the presence of scale bias and reporting bias. We used a method based on extreme value theory developed by De Fontnouvelle *et al.* (2006) to compute inverse probability weights to correct for reporting bias. Then using those weights a regression analysis was carried out to determine the relationship between the loss severity and the size of banks, in order to correct for the scale bias. Most of the previous work had studied only the relationship between the mean of the log-losses and the size of a bank. We argue that a model which only accounts for the scaling in the mean of the log-loss distribution would not be able to capture the tail behaviour of operational losses properly, which is the main interest for capital calculation. A good scaling model should be able to account for the variation in location, scale and shape of the distribution across different business lines and event types. Hence, all the parameters of the distribution should be tested for scaling properties. To achieve this we employed the recently introduced GAMLSS framework to carry out the regression analysis. The model selection process revealed that the operational losses are best explained by a log-gamma distribution with both distributional parameters being related to the size of the bank.

We validated the fitted model by simulating losses for the US banking industry, and two individual banks in the US. Tests demonstrated that the simulated losses follow the observed losses reasonably well. Furthermore, a capital calculation exercise for several hypothetical banks by using simplified assumptions about the frequency distribution revealed that the capital estimates provided by the model is consistent with the industry standards. In addition, backtesting of the model using two US banks also confirmed the validity of the model.

The results of this study suggest that the size of a bank is an important indicator of its operational loss severities than previous research has suggested. In particular, the model indicates that there is a considerable variation in the tail index of the loss distribution across banks of different sizes. We find that for certain BL/ET combinations, 99<sup>th</sup> percentile loss can increase by up to 25% to 35% when the size of a bank doubles.

An important observation we make is that the tail index for certain BL/ET categories can become less than 1 for very large banks. That is, the loss distribution of certain BL/ET categories is extremely heavy-tailed for very large banks. This can have significant implications for capital calculation as well as for risk management as discussed in the text. In particular, the results show that the loss distribution of CPBP events is extremely heavy-tailed and a significant proportion of the total loss due to CPBP events is caused by a few individual catastrophic losses. This phenomenon is referred to as the one-loss-causes-ruin problem and creates further challenges to estimating the tail of the loss distribution. Although the use of external data as outlined in this paper can improve the estimation of the tail of the loss distribution, given that catastrophes are high degree unique events, the importance of using scenario analysis cannot be stressed enough in this particular context. However, it should be noted also that scenario analysis also has its limitations since dreaming up worst-case scenarios are often very difficult for a number of reasons. Therefore, we stress the importance of having a system to monitor near-miss losses for early detection of possible crises and having a well established crisis management system in addition to the risk management system to assist a bank to steer through a catastrophic event.

The scaling model developed in this study is of particular importance to banks, as it can be used to complement their internal data and improve the estimates of the tail of the distribution. For example, it is possible to estimate the tail index of the operational loss distribution of a particular bank as  $\hat{\alpha} = \omega\hat{\alpha}_{Int} + (1-\omega)\hat{\alpha}_{Ind}$ , where  $\hat{\alpha}_{Int}$  is the estimate of the tail index given by the bank's internal data,  $\hat{\alpha}_{Ind}$  is the estimate of the tail index given by the proposed model using industry data, and  $\omega$  is a credibility weight decided by the experts. The bank could assign a higher weight to the internal estimate  $\hat{\alpha}_{Int}$  if it has satisfactory amount of high-quality data, however when the quality of internal data is poor or not relevant due to a recent change in the operational environment such as a merger or an acquisition, the bank will need to assign a higher weight to the external estimate  $\hat{\alpha}_{Ind}$ . In practice, *ad hoc* procedures are often used to decide the credibility weights. However, it is also possible to use a full credibility theory approach to combine the internal and external estimates as proposed by Bühlmann *et al.* (2007) and Lambrigger *et al.* (2007). Another use of the proposed model is that it can serve as a good starting point for building internal models and as a reference for the banks as well as the regulators when examining the appropriate capital levels. We consider the model is an improvement over the existing models found in the literature for two main reasons: 1) it takes into account the scaling properties of all the parameters in the loss distribution rather than just the location parameter, and 2) the scaling properties of the tail index are explicitly modelled, which is more significant to the capital calculation rather than the mean of log-losses. However, a major limitation of our model is that simulated losses are not sensitive to the internal controls of a bank. Therefore it would be interesting to extend the model to include other explanatory variables that reflect the quality of risk management and governance of individual banks. For example, variables such as the annual risk management budget and the cost of employee training might be able to serve as proxies for the quality of risk management and staff competence. The number of past fiduciary breaches and regulatory fines may serve as proxies for the quality of corporate governance. We believe an extension of our model which takes into account the internal controls would provide better capital estimates which reflect the individual circumstances of the banks. However, in order to facilitate such study the industry will need to be generous enough to share some of their confidential data with academia.

## References

- Altamuro, J. & Beatty, A. (2010). How does internal control regulation affect financial reporting? *Journal of Accounting and Economics*, 49(1-2), 58–74.
- Arnold, B.C., Balakrishnan, N. & Nagaraja, H.N. (1992). *A first course in order statistics*. Wiley-Interscience, New York.
- Artzner, P., Delbaen, F., Eber, J.M. & Heath, D. (1999). Coherent measures of risk. *Mathematical finance*, 9(3), 203–228.
- Basel. (1988). International Convergence of Capital Measurement and Capital Standards. Basel Committee on Banking Supervision.
- Basel. (1996). Amendment to the Capital Accord to Incorporate Market Risks. Basel Committee on Banking Supervision.
- Basel. (2006). International Convergence of Capital Measurement and Capital Standards: A Revised Framework – Comprehensive Version. Basel Committee on Banking Supervision.
- Basel Committee on Banking Supervision (2003). The 2002 Loss Data Collection Exercise for Operational Risk. Bank for International Settlements.
- Basel Committee on Banking Supervision (2009). Results from the 2008 Loss Data Collection Exercise for Operational Risk Bank for International Settlements.
- Baud, N., Frachot, A. & Roncalli, T. (2002). *Internal data, external data and consortium data for operational risk measurement: How to pool data properly?*, 1–18. Groupe de Recherche Operationnelle, Credit Lyonnais, France.

- Bühlmann, H., Shevchenko, P.V. & Wüthrich, M.V. (2007). A “toy” model for operational risk quantification using credibility theory. *Journal of Operational Risk*, 2(1), 3–19.
- Cope, E. & Labbi, A. (2008). Operational loss scaling by exposure indicators: Evidence from the ORX database. *The Journal of Operational Risk*, 3(4), 25–45.
- Dahen, H. & Dionne, G. (2010). Scaling models for the severity and frequency of external operational loss data. *Journal of Banking & Finance*, 34(7), 1484–1496.
- De Fontnouvelle, P., DeJesus-Rueff, V., Jordan, J.S. & Rosengren, E.S. (2006). Capital and risk: New evidence on implications of large operational losses. *Journal of Money, Credit, and Banking*, 38(7), 1819–1846.
- De Fontnouvelle, P., Rosengren, E. & Jordan, J. (2007). Chapter 10: Implications of Alternative Operational Risk Modeling Techniques. In M. Carey & R.M. Stulz (Eds.), *The Risks of Financial Institutions* (pp. 475–512). University of Chicago Press, Chicago.
- Dunn, P.K. & Smyth, G.K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3), 236–244.
- Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling extremal events for insurance and finance*. Springer, Berlin: London.
- Evans, J. (2001). Pitfalls in the Probability of Ruin Type Risk Management. *Database of Actuarial Research Enquiry/Financial and Statistical Methods/Risk Pricing and Risk Evaluation Models/Probability of Ruin*, Paper 30, pp. 501–510. Casualty Actuarial Society.
- Gompers, P., Ishii, J. & Metrick, A. (2003). Corporate governance and equity prices. *The Quarterly Journal of Economics*, 118(1), 107–155.
- Kupiec, P.H. (1995). Techniques for Verifying the Accuracy of Risk Management Models. *Journal of Derivatives*, 3(2), 73–84.
- Lambrigger, D.D., Shevchenko, P.V. & Wüthrich, M.V. (2007). The Quantification of Operational Risk using Internal Data, Relevant External Data and Expert Opinions. *The Journal of Operational Risk*, 2(3), 3–27.
- McNeil, A., Frey, R. & Embrechts, P. (2005). *Quantitative risk management: Concepts, techniques, and tools*. Princeton University Press, New Jersey.
- Medova, E.A. & Berg-Yuen, P.E.K. (2009). Banking capital and operational risks: comparative analysis of regulatory approaches for a bank. *Journal of Financial Transformation*, 26, 85–96.
- Moscadelli, M. (2004). The modelling of operational risk: experience with the analysis of the data collected by the Basel Committee. Banca d'Italia.
- Na, H., Van Den Berg, J., Miranda, L. & Leipoldt, M. (2006). An econometric model to scale operational losses. *The Journal of Operational Risk*, 1(2), 11–31.
- Nešlehová, J., Embrechts, P. & Chavez-Demoulin, V. (2006). Infinite mean models and the LDA for operational risk. *Journal of Operational Risk*, 1(1), 3–25.
- R Development Core Team. (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rigby, R.A. & Stasinopoulos, D.M. (2001). The GAMLSS project: a flexible approach to statistical modelling. In B. Klein & L. Korsholm (Eds.), *New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling*. Odense, Denmark.
- Rigby, R.A. & Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3), 507–554.
- Rigby, R.A. & Stasinopoulos, D.M. (2010). A flexible regression approach using GAMLSS in R. <http://gamlss.org/images/stories/papers/book-2010-Athens.pdf> [accessed 30-Jul-2012].
- Rootzén, H. & Klüppelberg, C. (1999). A Single Number Can't Hedge against Economic Catastrophes. *Ambio*, 28(6), 550–555.

- Shih, J., Khan, A.S. & Medapa, P. (2000). Is the Size of an Operational Loss Related to Firm Size? *Operational Risk Magazine*, 2(1), 1–2.
- Stasinopoulos, D.M. & Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7), 1–46.
- Ultsch, A. (2002). Proof of Pareto's 80/20 law and Precise Limits for ABC-Analysis. Technical Report 2002/c, DataBionics Reseach Group, University of Marburg.
- Wei, R. (2007). Quantification of operational losses using firm-specific information and external database. *Journal of Operational Risk*, 1(4), 3–34.
- Wilson, S. (2007). A Review of Correction Techniques for Inherent Biases in External Operational Risk Loss Data. *Working Papers*, Australian Prudential Regulation Authority.