SAS® GLOBAL FORUM 2020 #SASGF

Paper 5069-2020

Using SAS® OpRisk Global Data to Improve Decision-Making at a Bank

Mentje Gericke, Helgard Raubenheimer, Centre for Business Mathematics and Informatics®, North-West University

ABSTRACT

The management of financial losses is crucial as banks are required to set aside regulatory capital to absorb unexpected losses. Banks also need to calculate economic capital to ensure solvency according to their own risk profile. The main financial risks faced by banks are market, credit, and operational risk. Operational risk, the focus here, includes fraud, improper business practices, and so on. Barings Bank's loss of over USD1 billion due to rogue trading activities is an extreme example of such risk. In order to calculate capital to withstand this risk, the aggregate distribution of expected losses for the next year is determined. The extreme quantiles of this distribution are of specific interest. For instance, a bank should hold capital to survive a one-in-a-thousand-year aggregate operational loss (the 99.9% VaR of the distribution). Companies often have only limited internal data available to accurately model the distribution and therefore use external sources and scenario assessments to supplement their data. Combining the internal data of a given bank with external data is challenging, as such data is collected from differently sized institutions in various regions. This might impact the estimated loss distribution. In this paper, we use SAS® OpRisk Global Data to show how external and internal data can be integrated for use in the capital modeling process. We also suggest measures to challenge experts to adjust scenario assessments based on historical data.

INTRODUCTION

Financial institutions use statistical models to determine their required capital. If they are able to model or predict the amount of the total losses they could potentially suffer in the future, and assign a probability to these losses, they can determine the amount of capital to hold in order to ensure that they can withstand that loss at a certain confidence level. For operational risk capital, this is done by calculating the 99.9% Value-at-Risk (VaR) of the aggregate operational loss distribution.

A popular method for constructing the annual aggregate loss distribution is the loss distribution approach (LDA). Companies and regulators alike rely heavily on the distribution function of expected losses, and it is therefore crucial that it is modeled as accurately as possible. The tail of the distribution is most important, as we are ultimately interested in the extreme quantiles of the distribution in order to calculate capital. When specifically dealing with operational losses, Basel II prescribes that a bank should hold sufficient capital to protect them against a one-in-a-thousand year aggregate loss, i.e. the 99.9% Value-at-Risk of the aggregate operational loss distribution. Ideally, should the bank have a thousand years of historical data, the bank can merely determine the largest loss it had experienced during this period of time to determine the capital requirement. However, in reality, most banks only have about ten years of available loss data. To address this shortcoming, the Basel Committee on Banking Supervision (2011) suggests that loss data from external sources and scenario data can be used by banks in addition to their own internal loss data and controls. For example, external loss data can be compared with internal loss data or it can be used to explore possible weaknesses in the control environment or consider previously unidentified risk exposures. However, the process of incorporating data from external sources requires due consideration because of biases in the external data. Wilson (2007) outline three types of biases inherent in external data, namely the reporting bias, control bias and scale bias. In this paper we address the reporting bias that occurs when institutions use different thresholds to report losses to an external database, and the scale bias that occurs when data is collected from institutions with a different size. Control bias refers to losses that come from institutions with different control mechanisms, and although not specifically addressed in this paper due to limited data on this aspect, it is an area that we have identified for potential further research.

We propose the use of SAS® OpRisk Global Data ("SAS data" or "SAS dataset") to inform the decisions of an individual bank in determining their own operational risk capital. The purpose of our study is to apply a scaling methodology using the SAS data to ensure it is appropriate to the bank when used in their capital modelling process. We show how the external data may potentially be used to challenge business experts to adjust their scenario assessments using the realism of the observed historical data.

METHODOLOGY

AGGREGATE LOSS DISTRIBUTION

The loss distribution approach (LDA) is a popular method used by banks and other financial institutions to determine their operational risk capital. This approach is widely described in the literature (see for example Aue and Kalkbrener (2007), Benito and Lopez-Martin (2018), Lambrigger et al. (2007) and De Jongh et al. (2015)). Under this approach, an organization can estimate the probability distributions of both the severity and the one-year-event frequency using historical data. Having these two distributions, the organization can then compute the probability distribution of the aggregate operational losses (Benito & Lopez-Martin, 2018). The methodology to construct the aggregate loss distribution is briefly described below.

Typically the Poisson distribution is used to model the annual frequency or number of operational loss events over one year. N is a random variable representing the annual number of loss events, i.e. $N \sim Poi(\lambda)$. The random variables $x_1, \ldots, x_N$ denote the loss severities of the loss events and we assume these loss severities are independently and identically distributed. The annual aggregate loss is then given by $A = \sum_{n=1}^{N} x_n$ and the distribution of A is the aggregate loss distribution, which is a compound Poisson distribution that depends on λ and the true severity distribution of $x_1, \ldots, x_N$, denoted by F. In order to determine the aggregate distribution, estimates for λ, the frequency, and F, the severity distribution are needed. We therefore have to decide on a suitable model for F, which can be a class of distributions $f(x, \theta)$. The parameters of λ and θ also needs to be estimated. As previously mentioned, the the 99.9% Value-at-Risk of the aggregate operational loss distribution is of interest for capital estimation, but in most cases it is difficult to do this analytically, and Monte Carlo simulation is often used.

Other numerical methods can also be used. For our own risk capital estimations calculated in this paper, we make use of the single-loss approximation (SLA) method suggested by Böcker & Klüppelberg (2005), that we summarize as follows: if F is the true underlying severity distribution function of the individual losses, and λ is the true annual frequency, then the $100(1-\gamma)\%$ VaR of the compound loss distribution may be approximated by $F^{-1}(1 - \gamma/\lambda)$.

The focus of our paper is on the estimation of the severity distribution $f(x, \theta)$ using the internal data of an individual bank, but also using data from an external database. We therefore do not expand on the benefits or limitations of the aggregation approaches mentioned above.

SAS® OPRISK GLOBAL DATA

It is practice in operational risk management to use different data sources for modelling future losses. Banks have typically been collecting their own loss data for a period of time. In addition, certain external loss databases exist, including publicly available data, insurance data and consortium data. The Basel Accord (2011) also suggests the use of scenario assessments to improve severity distribution estimation.

For our investigation, we use SAS® OpRisk Global Data. The SAS® OpRisk Global Data is a comprehensive and accurate repository of information on publicly reported operational losses in excess of USD100 000, containing more than 32 000 events across all industries worldwide. For each publicly available operational loss, the SAS dataset provides the loss amount together with some other information about the company where the loss occurred. This include, among other, a description of the loss event, as well as the region, the size of assets and other information associated with each loss.

In our study we have only included losses incurred in the financial industry as it is aimed at decision-making at a bank. We have also only included losses above USD 1 million and we therefore had 10,935 available data points. We discuss the distribution of the data in more detail under the heading "Explanatory variables".

ALLOWING FOR REPORTING BIAS

The SAS® OpRisk Global Data contains information obtained from several online information providers and other publications. A team of seasoned SAS operational risk research analysts maintain the database in accordance with strict data quality standards and review it periodically in order to update it and to ensure accuracy and completeness. Ganegoda & Evans (2012) argues that most external databases, but especially those maintained by vendors collecting publicly reported losses, suffer from reporting bias. Wilson (2007) explains that larger losses (and also those associated with larger firms) are more likely to be reported in the media due to factors such as the size and nature of the loss. This is due to the fact that not all operational losses are reported on public platforms and this is especially true for smaller losses. As a result, public databases may contain a disproportionately high number of large losses, and one should make allowance for this bias in fitting a statistical model, or else the tail of the distribution will be overestimated.

Ganegoda & Evans (2012) draws on a method first introduced by De Fontnouvelle *et al.* (2006) to assign a weight to each loss in the external database. This means that smaller losses will carry a greater weight and large losses will carry a smaller weight. We briefly explain their methodology below.

They firstly assume that a (log) loss y_i is only reported in the public domain if it exceeds a certain truncation or observation point, t_i. This truncation point t_i is a stochastic variable and should not be confused with the threshold at which losses are captured in the database, being USD100 000 in the case of the SAS database. To explain this, if a loss is greater than the unobserved truncation point, but lower than the USD100 000 threshold, the research analyst responsible for compiling the database will observe the loss, but not include it in the database. On the other hand, if the unobserved truncation point is higher than the log USD100 000 threshold, the analyst will not know about the loss, and it will for this reason not be included in the database. Therefore, only losses greater than both t_i and USD100 000 will be included in the database.

Because they assume that the loss amount, y_i and truncation point, t_i are independent, the distribution of losses in the database is given by:

$$f(y_i|y_i > t_i) = \frac{f(y_i)G(y_i)}{\int_{\mathbb{R}} f(y)G(y)\,dy}.$$

They recommend using a Logistic distribution for $G(.)$, which is given by

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

where τ 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 the probability of being reported increases with the size of the loss.

$z_i = y_i - u$ is defined as the excess log loss over a high enough threshold u, and it is shown that z_i can be approximated using an exponential distribution. They obtain the following likelihood equation

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

where $h(z_i;b) = \frac{1}{b}\exp(-\frac{z_i}{b})$ and $\tau^* = (\tau - u)$. The parameters b, τ^* and a are estimated by maximising the likelihood function and finally the normalized weights to be assigned to each loss is calculated as

$$w_i' = \frac{nw_i}{\sum_{i=1}^{n} w_i},$$

where:

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

In order to confirm the existence of reporting bias in the SAS dataset, we carried out likelihood ratio tests of the restriction that the reporting probabilities are constant across all losses for each threshold level (i.e. that 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, confirming the existence of reporting bias.

We estimated the parameters b, τ^* and a for different choices of the threshold u and found that the parameter values for b and a stabilized after the USD6 million threshold. We therefore used the associated estimates $\hat{a} = 1.08941$ and $\hat{t} = (3.59446 + \log(6))$ to calculate corresponding weights for all the losses reported in the SAS database.

ALLOWING FOR SCALING BIAS

The scaling methodology we apply to the external SAS data in order to model the severity distribution of operational losses will correct for the scale bias alluded to earlier in this paper.

We use the method introduced by Ganegoda and Evans (2012) using regression analysis based on the Generalized Additive Models for Location Scale and Shape (GAMLSS) framework to model the scaling properties of operational losses. They explain that the GAMLSS framework has the ability to model all the distributional parameters and therefore offers flexibility in estimating the scaling properties of a model.

In their paper, Ganegoda and Evans (2012) argue that a good scaling model should also be able to make allowance for the variation of model parameters for different business lines and event types. The discussion below provides the technical background to their approach.

We consider log losses denoted by $\boldsymbol{y} = (y_1, \dots, y_n)^T$, a random sample of independent observations. We assume that these log losses follow some parametric distribution $f(y_i; \vartheta_i)$ with parameter vector ϑ_i. For the sake of simplicity and in line with Ganegoda and Evans' (2012) notation, we assume that $\vartheta_i = \left(u_i, \sigma_i\right)^T$ is a vector of only two distributional parameters.

A set of link functions are defined that specifies the relationship between the linear predictor and the distributional parameters of each distribution component distribution as:

$$\begin{aligned} g_1(u_i) &= \eta_{i1} = \exp(\beta_{11} + \beta_{12}X_{i12} + \cdots + \beta_{1p}X_{i1p}), \\ g_2(\sigma_i) &= \eta_{i2} = \exp(\beta_{21} + \beta_{22}X_{i22} + \cdots + \beta_{2p}X_{i2p}), \end{aligned} \tag{1}$$

for $i = 1, \dots, n$, where X_{ijp} is the value of the pth explanatory variable relating to the observation y_i in the jth distributional parameter, and β_{jp} is the parameter corresponding to X_{ijp}. The set of equations are simplified with the help of matrix notation as follows:

$$\begin{aligned} g_1(u_i) &= \boldsymbol{X_1\beta_1}, \\ g_2(\sigma_i) &= \boldsymbol{X_2\beta_2}, \end{aligned}$$

where, $\boldsymbol{X_j}$ are the matrix of the jth distributional parameter, and $\boldsymbol{\beta_j}$ are the corresponding parameter vectors. The maximum likelihood estimates of $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are then obtained by solving:

$$\max_{\boldsymbol{\beta_1},\boldsymbol{\beta_2}} \sum_{i=1}^{n} w_i' \log f(y_i; \boldsymbol{\beta_1}, \boldsymbol{\beta_2}).$$

In order to solve the above equation, we used the PROC NLP function in SAS Enterprise Guide.

EXPLANATORY VARIABLES

The literature suggests that the extent of operational risk losses can be impacted by a number of factors associated with the firm where the loss occurs. These are included as the explanatory variables in the scaling model explained above, and are discussed in more detail under this section.

Numerous studies have suggested that there may be some relationship between the size of a firm and the operational loss amount (for examples, refer to Shih *et al.* (2000), Dahen and Dionne (2010) and Cope and Labbi (2008)). The SAS data include a number of variables that could potentially be indicative of the size of the firm, including revenue, net income, asset value, shareholder equity and the number of employees for the fiscal year in which it experienced the loss. It is reasonable to assume that there is a positive correlation between these variables, and therefore we have only selected a single variable to represent the size of the firm, namely the assets of the firm, as an explanatory variable.

Ganegoda and Evans (2012) also suggest that the equity ratio, being the proportion of equity used to finance the company's assets, can give an indication of the risk taking tendency of management. It provides a measure of leverage used and given that both the assets and shareholder equity are provided in the SAS data, this ratio could easily be computed. It was used as the second explanatory variable in our scaling model.

The third explanatory variable included in our model was the geographic region in which the firm operates, being Africa, Asia, Europe, North America, Other Americas or Other. Wilson (2007) explains that all operational losses arise as a result of a specific set of circumstances due to a lack of, or failure in controls. The reason for including region as an explanatory variable is based on the assumption that the circumstances should be similar in different geographic regions and should therefore impact on the size of operational losses. Cope and Labbi (2008) showed differences in the loss distributions for banks of various sizes and operating in different geographies. Table 1 provides summary statistics about the losses in the different regions.

Table 1: Summary Statistics per Geographical Region

			Log-losses (USD Million)		
Region	Number of losses	% of losses	50th percentile	90th percentile	99th percentile
Africa	120	1,1%	1,49	5,08	6,14
Asia	1 429	13,1%	2,10	5,03	7,18
Europe	2 752	25,2%	2,43	5,69	7,91
North America	6 210	56,8%	1,98	4,76	7,21
Other	306	2,8%	1,83	4,60	6,59
Other Americas	118	1,1%	2,89	5,59	7,45
	10 935	100%			

Note: Given the relatively small number of losses reported in *Africa*, *Other* and *Other Americas,* we have grouped these losses together.

The Basel Committee on Banking Supervision (2005) specifies that a bank's activities should be categorized into a number of business lines, and a comprehensive set of non-overlapping operational event types should be defined and applied across the various business lines.

Some business lines are considered more risky than others and may potentially suffer higher losses, and hence the severity distribution will be impacted by the business line, being our fourth explanatory variable.

Our final explanatory variable was event type, as it is found that different types of loss events are associated with higher losses. A list of the categories of business lines and event types used in our analysis is provided in Table 2 and Table 3.

Table 2: Summary Statistics per Business Line

			Log-losses (USD Million)		
Business line	Number of losses	% of losses	50th percentile	90th percentile	99th percentile
Agency Services	171	1,6%	3,14	5,61	7,77
Asset Management	502	4,6%	2,61	5,01	7,51
Commercial Banking	1 990	18,2%	2,21	4,88	7,20
Corporate Finance	574	5,2%	2,90	5,66	8,18
Insurance	1 881	17,2%	2,19	4,80	7,01
Payment and Settlement	219	2,0%	2,40	5,83	7,67
Retail Banking	3 561	32,6%	1,59	4,58	7,56
Retail Brokerage	788	7,2%	1,48	3,79	6,37
Trading & Sales	1 249	11,4%	3,18	6,08	8,40
	10 935	100%			

Note: Given the relatively small number of losses reported under *Agency Services*, *Asset Management* and *Payment and Settlement,* we have grouped the losses in these categories together. We will refer to this category later in the paper as *AS, AM and PS*.

Table 3: Summary Statistics per Event Type

			Log-losses (USD Million)		
Event type	Number of losses	% of losses	50th percentile	90th percentile	99th percentile
Business Disruption and System Failures	60	0,5%	3,03	5,24	6,24
Clients, Products & Business Practices	5 660	51,8%	2,63	5,55	7,91
Damage to Physical Assets	76	0,7%	1,95	5,04	7,28
Employment Practices and Workplace Safety	349	3,2%	1,84	4,23	6,55
Execution, Delivery & Process Management	489	4,5%	1,47	4,15	6,88
External Fraud	2 229	20,4%	1,46	3,91	6,36
Internal Fraud	2 072	18,9%	1,61	4,57	7,03
	10 935	100%			

Note: Given the relatively small number of losses reported under *Business Disruption and System Failures*, *Damage to Physical Assets* and *Employment Practices and Workplace Safety,* we have grouped the losses in these categories together. We will refer to this category later in the paper as *SF, D and EP*.

MODEL APPLICATION

The first step in our model selection process was to find a base model that closely follows our data, but without taking into account any of the explanatory variables set out above. In other words, we first selected an appropriate probability distribution assumption to be used in our subsequent model fitting. For this purpose we used the SEVERITY procedure in SAS. We consider six different parametric models, of which we have included the density functions in Appendix B.

In order to select the best base model, we considered three goodness of fit tests. These are twice the negative log-likelihood (-2LogLikelihood), the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both the AIC and BIC are based on the -2LogLikelihood, and smaller values of all these criteria indicate a better model. Both the AIC and BIC penalize models with more parameters, but the BIC even more so, and we therefore used the BIC as our main determining factor in selecting our best-fit model. The BIC is defined as:

$$BIC = -2LogLikelihood + kln(n),$$

where k is the number of estimated parameters in the model and n is the number of observations used in the model.

Our results showed that the Burr distribution had the lowest BIC value, but the fact that it has three parameters introduced potential complications for our scaling model. For this reason we decided to use the Gamma distribution that ranked second among our potential models, and because it was also the severity distribution used by Ganegoda and Evans (2012). The Gamma model only has two parameters, being the location and scale parameters that was estimated to be $\hat{\alpha} = 1.084527$ and $\hat{\theta} = 0.88478$.

Based on the fact that the Gamma distribution was identified as the most appropriate distribution function for our data, we assumed that the Gamma model was also appropriate as our base to continue the modelling process.

The Gamma distribution was fitted to the data again, this time allowing for the explanatory variables introduced in the previous section. We carried out a step-wise selection of these variables in order to determine the parameters θ and α of the Gamma distribution using the link functions introduced in Equation 1. The first step involved a forward selection of variables only for θ, followed by a forward selection of variables for α given the model we had obtained for θ. Thereafter we followed a backward elimination of variables for θ, given the selected model for θ and α and a backward elimination of variables for α.

Based on the step-wise selection method described above, we found that Log-assets and seven other business line and event type explanatory variables were significant to the scale parameter θ. None of the region variables were found to be significant for θ.For the shape parameter α, Log-assets, the region variable Asia, Retail Brokerage and Execution, Delivery & Process Management were found to be significant explanatory variables. The parameter estimates of the final model given by the step-wise selection method are shown in Table 4.

Table 4: Estimated Parameter Values for our Final Model

Explanatory variable	$\boldsymbol{\theta}$		$\boldsymbol{\alpha}$	
	Estimate	Std. Error	Estimate	Std. Error
Intercept	-0,446390	0,084198	0,411743	0,080968
Log-assets	0,029426	0,006454	-0,02159	0,007091
Equity ratio	-	-	-	-
Africa, Other Americas, Other	-	-	-	-
Asia	-	-	-0,0892	0,037367
Europe	-	-	-	-
Corporate finance	0,189131	0,049063	-	-
AS, AM & PS	0,129315	0,040584	-	-
Commercial Banking	0,258638	0,031836	-	-
Insurance	-	-	-	-
Retail Banking	-0,08911	0,024295	-	-
Retail Brokerage	-	-	-0,20123	0,044886
Clients, Products & Business Practices	0,100373	0,042772	-	-
Execution, Delivery & Process Management	-	-	-0,27172	0,070796
External Fraud	-0,31829	0,046103	-	-
Internal Fraud	-0,18629	0,046095	-	-

Note: The model was fitted using *North America*, *Trading and Sales* and *SF, D and EP* as the baseline categories for geographical region, business line and event type respectively.

MODEL DIAGNOSTICS AND RESULTS

Ganegoda & Evans (2012) uses normalized quantile residuals, $\hat{r}_i$, to verify the adequacy of the fitted GAMLSS models. For a response variable Y with a continuous cumulative distribution function $F(y_i; \hat{\theta}_i)$, the normalized quantile residuals are defined as $\hat{r}_i = \Phi^{-1}[F(y_i; \hat{\theta}_i)]$, where Φ^{-1} is the inverse cumulative distribution function of the standard Normal distribution. According to Rigby & Stasinopoulos (2005), the error $\hat{r}_i$ should be standard Normally distributed if the model is adequate. We show the QQ plot of estimated residuals against the theoretical quantiles of the standard Normal distribution in Figure 1 graphically indicating the normality of the estimated residuals.

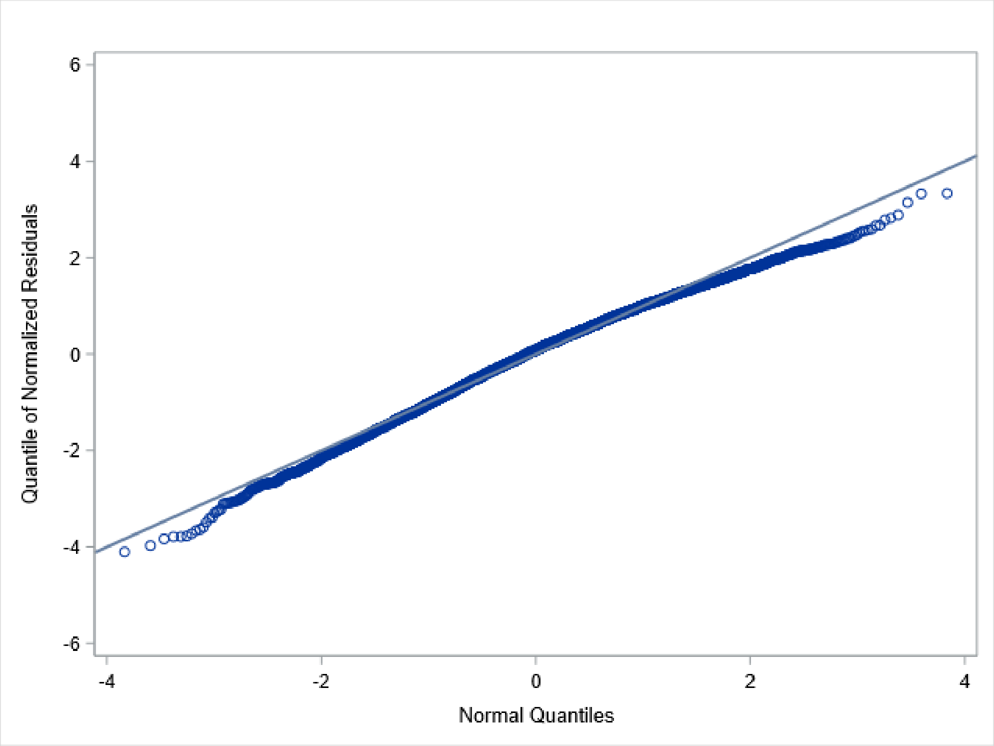

Figure 1: Normalized Quantile Residual Plot

As a further validation of our model, we have simulated 1 000 000 losses and tested the goodness-of-fit by comparing the quantiles of the simulated losses with the observed losses using a QQ plot. The QQ plot is shown in Figure 2.

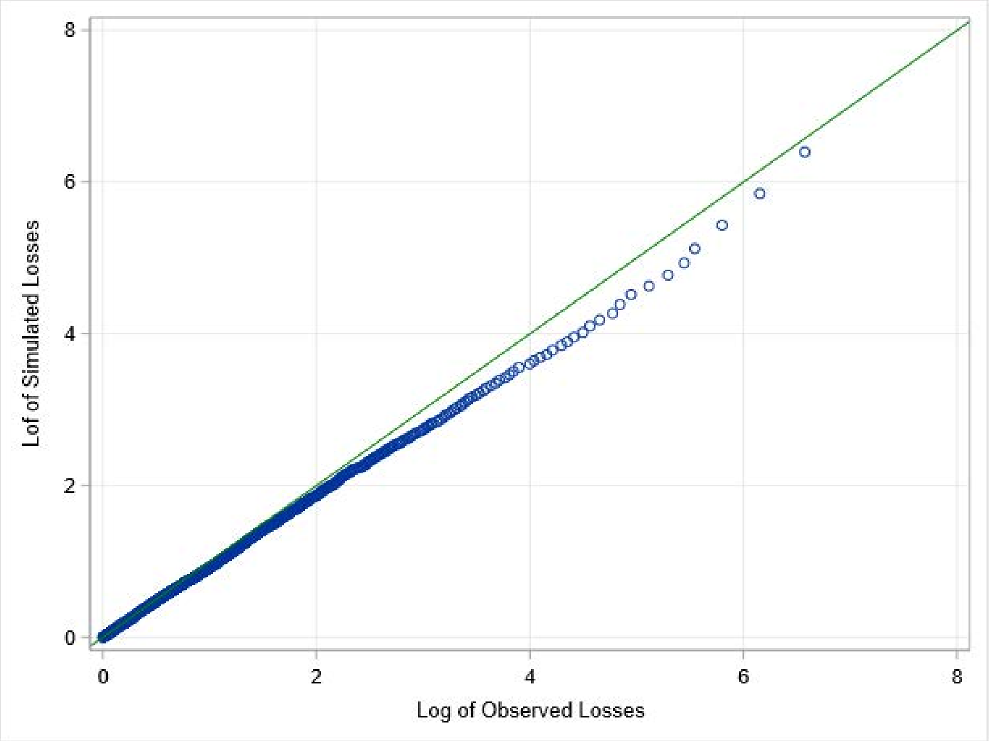

Figure 2: QQ Plot of Simulated Losses vs. Observed Losses

SCENARIO ANALYSIS

We previously noted that the purpose of our model is to assist banks with their scenario analysis process, and this may be specifically helpful for banks with limited internal data. In order to show how our model can be used to determine quantiles of our aggregate distribution, we first consider an approach for banks to obtain scenarios that could be used in risk capital models.

De Jongh *et al.* (2015) describe one such approach to scenario analysis, and refer to this as the 1-in-c years method. They explain that the scenario makers are asked the following question: "What loss level q_c is expected to be exceeded only once every c years". They suggest popular choices for c to be 7, 20 and 100 years and motivate their first choice of 7 as the number of years of historical data available to a bank.

In order to determine the quantiles of our aggregate distribution that correspond to our 1-in-c year losses as suggested above, we draw on the work done by De Jongh *et al.* (2015) to combine historical data and scenarios. They explain that if the annual loss frequency is Poisson distributed with parameter λ, and the underlying severity distribution is F, and if scenario makers know the exact values of λ and F, then the scenario assessments provided for q_c, being the loss only exceeded once in c years, should be:

$$q_c = F^{-1}(1 - \tfrac{1}{c\lambda}). \qquad (2)$$

They construct a spliced distribution function, using backward-looking historical information for the "expected" (or "body") part of the distribution and forward-looking scenario information for the "unexpected" (or "tail") part. They select a number b with corresponding quantile q_b and denote $F_e(x)$ as the conditional distribution function of a random loss $X{\sim}F$, given that $X \leq q_b$, and $F_u(x)$ as the conditional distribution function, given that $X > q_b$. The distribution function for $F_u(x)$ is then given by:

$$F_u(q_c) = \frac{[F(q_c)-F(q_b)]}{[1-F(q_b)]} \; for \; q_c > q_b. \qquad (3)$$

Because we only model losses greater than USD1 million, we effectively only model the unexpected part of the severity distribution as explained above and we therefore need to make allowance for the expected part of our distribution if we intend to use the model to determine capital estimates in the tail of the distribution. If we do not make allowance for losses below USD1 million, we will under-estimate our required risk capital. To explain this further using the notation set out above, our q_b is equal to USD1 million, although this is a pre-determined amount and not specifically related to the loss amount only exceeded every b years. We therefore do not know the probability that losses would exceed USD1 million, i.e $P(X > 1) = 1 - F(1)$, and for comparative purposes we assume that $F(1)$ is between 0,95 or 0,98. Although not exact, these assumptions are based on data from the Loss Data Collection Exercise done by Basel in 2008. What this means is that we will adjust our quantiles using Equation 2 to make allowance for the fact that we are conditionally modelling above USD1 million. Table 5 shows the adjusted probabilities for different values of $F(1)$, i.e. the cumulative probability that losses would be less than USD1 million. The probabilities are calculated using Equations 1 and 2 and assuming an annual frequency of 6.58627. The reason for selecting this value for our annual frequency, will be explained in the following section.

Table 5: Adjusted probabilities for different values of $\boldsymbol{F}(\mathbf{1})$

Scenario point	Cumulative prob. on $\boldsymbol{F}(.)$	Cumulative prob. on $\boldsymbol{F_u}(.)$ for values of $\boldsymbol{F}(\mathbf{1})$			
		0.95	0.96	0.97	0.98
1-in-10 year	0,984746	0,696338	0,620423	0,493897	0,240845
1-in-20 year	0,992373	0,848169	0,810211	0,746948	0,620423
1-in-100 year	0,998475	0,969634	0,962042	0,949390	0,924085
99.9% VaR	0,999848	0,996963	0,996204	0,994939	0,992408

RESULTS FOR AN INDIVIDUAL BANK

In this section we show how the model we have built on SAS data, can be utilized by an individual bank. We assume that we have the internal loss data for the bank, and for this purpose we have extracted the loss data for the Bank of America Corporation from the SAS data. In the remainder of this paper we refer to Bank of America Corporation as our individual bank. Table 6 provides a summary of the number of operational losses above USD1 million for our indivdiual bank, and reported in the SAS database. It should be noted that it is expected that the bank itself would have a significant higher amount of data, given that the internal data would not suffer from reporting bias. In addition, it is expected that the internal data would include information on losses below USD1 million. This information could be used to model the body of our severity distribution, although in using our model we do adjust our quantiles to allow for this fact.

Table 6: Bank of America Corporation Loss Data Points per Business Line and Event Type

	Business line					
Event type	Clients, Products & Business Practices	Employment Practices and Workplace Safety	Execution, Delivery & Process Management	External Fraud	Internal Fraud	Total
Agency Services	2	-	-	-	-	2
Asset Management	10	-	1	-	-	11
Commercial Banking	1	-	-	3	1	5
Corporate Finance	10	1	-	-	-	11
Insurance	1	-	-	-	-	1
Payment and Settlement	-	-	1	-	-	1
Retail Banking	26	2	1	21	7	57
Retail Brokerage	28	9	6	-	7	50
Trading & Sales	34	1	5	1	3	44

We run our model again, but this time excluding the loss data of our individual bank. The results of our new model is shown in Table 7.

Table 7: Re-estimated Parameter Values for our Model, Excluding the Individual Bank's Data

Explanatory variable	$\boldsymbol{\theta}$		$\boldsymbol{\alpha}$	
	Estimate	Std. Error	Estimate	Std. Error
Intercept	-0,41268	0,085442	0,402985	0,08192
Log-assets	0,027787	0,006581	-0,02033	0,007212
Equity ratio	-	-	-	-
Africa, Other Americas, Other	-	-	-	-
Asia			-0,08989	0,037436
Europe	-	-	-	-
AS, AM & PS	0,137837	0,040684	-	-
Commercial Banking	0,250548	0,031946	-	-
Corporate Finance	0,193883	0,04964		
Insurance	-	-	-	-
Retail Banking	-0,09341	0,024465	-	-
Retail Brokerage	-	-	-0,21234	0,046812
Clients, Products & Business Practices	0,082871	0,043878	-	-
Execution, Delivery & Process Management	-	-	-0,28332	0,072153
External Fraud	-0,33385	0,04706	-	-
Internal Fraud	-0,20375	0,047088	-	-

Using the results of our new model, we simulate 1 000 000 losses for each of two business lines, namely retail banking and retail brokerage. Ideally we would want to simulate losses only for one event type within a business line (for example, external fraud in retail banking), but given the limited number of data points per individual bank, we have grouped all losses over event type within a single business line, i.e. assuming that event types are independent.

In order to obtain estimates for a 1-in-10 year, 1-in-20 year and 1-in-100 year loss per business line, we need to make an assumption for λ, the annual loss frequency of losses. For this, we again refer to Ganegoda & Evans (2012), and they approximated that a bank with USD1 billion assets would experience 0.00823 losses per year, based on data from a Loss Data Collection Exercise done by Basel in 2008. They further show that the total number of losses per year can be weighted to obtain a frequency for each business line and event type within the bank. We used a similar approach and assumed that our individual bank has assets of USD2 trillion (based on the SAS data), in order to estimate frequencies for our two business lines. The estimated annual frequencies for retail banking was therefore $6{,}586$ and 1.485 for retail brokerage.

In Table 8, for our two business lines, we show the model estimates for our scenario points for a 1-in-10 year, 1-in-20 year and 1-in-100 year loss, corresponding to the quantiles for the adjusted probabilities of our fitted distribution as shown in Table 5. Note that we only show the scenario point estimates for the assumption that $F(1) = 0.98$, i.e. the probability that losses are above USD1 million, is 0.02. We also show the 1-in-1000 year estimate, given that this would be the amount corresponding to the 99.9% Value-at-Risk and therefore the regulatory capital required for the business line. In addition to the point estimates, we show the distribution free 90% confidence intervals for these quantiles.

Given that we have loss data specific to our individual bank, also referred to as "internal loss data", we could use this data in isolation to fit a model specific to our individual bank. The concern with this approach is that the data, and especially when working within a specific business line and event type, is fairly limited as shown in Table 6.

We are only working with publicly available data for the individual bank, but even if one had access to all the bank's collected data, it tends to be limited and even more so for higher losses. This point also illustrates the need for banks to augment their own internal data with data from external sources. For the same two business lines under consideration, we fit a Gamma-distribution only to the internal data points. We compare the same quantiles estimated from these models to the estimates from our GAMLSS model described under the Methodology section. Table 8 provides a summary of the results obtained from the two models for the two business lines.

Table 8: Estimated Scenario Points per Business Line for Different Models

	Retail banking		Retail brokerage	
	Individual **bank's data**	Model	Individual **bank's data**	Model
1-in-10 year	0.139162	0.265663 (0.265;0.266)	-	- (-)
1-in-20 year	0.684096	0.877871 (0.876;0.879)	-	- (-)
1-in-100 year	2.179431	2.296219 (2.292;2.300)	0.901160	0.845764 (0,844; 0,847)
1-in-1000 year	4.469461	4.404938 (4.392;4.419)	2.518588	2.917611 (2,911;2,924)

Table 8 shows that for the retail banking business line, the estimated scenario points are similar for both models, where the first model is based on internal data and the second model on external data, but tailored for the unique explanatory variables specific to the individual bank. For the retail brokerage business line, where the internal data is even more scarce, the difference between the estimates of the two models is greater.

The estimated scenario points for 1-in-10 years and 1-in-20 years are zero for the retail brokerage business line. This is due to the fact that the estimated annual frequency of losses in this business line is only 1.485. As a result, our individual bank is not expected to observe losses higher than USD 1 million in this business line in 10 or even 20 years' time.

DISCUSSION

In operational risk management, banks use different data sources for modelling future losses. Given that most banks only have limited internal data, they often subscribe to external data consortiums or make use of other external data sources to augment their own data. In addition, many organizations use expert scenario assessments to inform the magnitude of extreme losses, that is, the tail of the loss distribution.

The Basel Accords suggest ways that banks can use scenario assessments to improve the estimation of the loss distribution. The Basel Committee on Banking Supervision (2011b) emphasizes that the scenario process is qualitative by nature and that outputs from such a process would contain significant uncertainties. Therefore, the purpose of our study is to show how external data, and specifically SAS® OpRisk Global Data, can be used to inform or challenge these more subjective scenario assessments.

We showed how the SAS data can be used to estimate the severity distribution of losses. Given the explanatory variables for a specific bank, the distribution $f(y;\hat{\boldsymbol{\beta}}_i)$ may be used to determine quantiles of the aggregate loss distribution, and these in turn can be compared to the scenario assessments of the experts. We assumed that experts or scenario makers are asked to answer the following question: 'What aggregate loss level is expected to be exceeded once in c years?'. Once we have selected an appropriate distribution function, the quantiles can be determined that relate to the scenario assessments provided by the experts. For example, the 1-in-100 year loss predicted by our expert should be in line with the 99% quantile of our aggregate loss distribution. Therefore, if the loss scenario points provided by the experts deviate too far from the quantiles of the loss distribution that was estimated by the data, one can revert back to the expert and request them to justify the difference. Using internal and external data, and specifically for units of measure where adequate historical data is available, one should be able to model future expected losses fairly well. However, the more significant benefit of our scaling model is for banks where very limited or no internal within a business line is available. In such a case the bank may use the model based on external data and use it's own characteristics to infer values expected future losses.

CONCLUDING REMARKS

In this paper we have showed how SAS® OpRisk Global Data can be used by a bank, when they do not have their own internal loss data, to build statistical capital models. We have also provided ways in which a bank can use a model only based on external data to inform or challenge the scenario assessments provided by experts. Scenario assessments are often used as a significant component of operational risk management, but given the subjective nature of these assessments, it is important to have an objective measure to check whether the expert's opinion is not biased or completely unrealistic. Although experts may not change their views based on the results of statistical models, they may be required to justify why their assessments deviate from the data. Our suggested model take into account the reporting bias included in any external database, but also shows that operational losses are dependent on certain factors specific to a bank, for example size and region, but also the business line and event type associated with operational losses.

REFERENCES

Aue, F., & Kalkbrener, M. (2007). LDA at work: Deutsche Bank's approach to quantifying operational risk. *Journal of Operational Risk , 1*(4), 49-93.

Basel Committee on Banking Supervision. (2005). *International convergence of capital measurement and capital standards. A revised framework.* Bank for International Settlements., Basel, Switzerland.

Basel Committee on Banking Supervision. (2011a). *Principles for the sound management of operational risk. Report 195.* Bank for International Settlements.

Basel Committee on Banking Supervision. (2011b). *Operational risk: Supervisory guidelines for the advanced measurement approaches. Report 196.* Bank for International Settlements.

Benito, S., & Lopez-Martin, C. (2018). A review of the state of the art in quantifying operational risk. *Journal of Operational Risk, 13*(4), 89-129.

Böcker , K., & Klüppelberg, C. (2005). Operational VaR: a Closed-Form Approximation. *Risk*, 90-93.

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., & Rosengren, E. (2006). Capital and risk: New evidence on implications of large operational losses. *Journal of Money, Credit, and Banking*, 1819–1846.

De Jongh, P.J., De Wet, T., Raubenheimer, H., & Venter, J. (2015). Combining scenario and historical data in the loss distribution approach: a new procedure that incorporates measures of agreement between scenarios and historical data. *Journal of Operational Risk, 10*(1), 45-76.

Ganegoda, A., & Evans, J. (2012). A scaling model for severity of operational losses using generalized additive models for location scale and shape (GAMLSS). *Annals of Actuarial Science, 7*(1), 61-100.

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.

Rigby, R., & Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society*, 507-554.

Shih, J., Samad-Khan, A., & Medapa, P. (2000). Is the size of an operational loss related to firm size? *Operational risk magazine.*, 1-2.

APPENDIX A: PROBABILITY DENSITY AND DISTRIBUTION FUNCTIONS

Table A1 provides a summary of the parametric models that were fitted to our data in order to determine the base model.

Table A1: Probability Density and Distribution Functions

Distribution	Par 1	Par 2	Par 3	Probability density function	Probability distribution function
Burr	$\theta > 0$	$\alpha > 0$	$\gamma > 0$	$f(x) = \alpha\gamma z^{\gamma}$	$F(x) = 1 - \left(\frac{1}{1 + z^{\gamma}}\right)^{\alpha}$
Gamma	$\theta > 0$	$\alpha > 0$		$f(x) = \frac{z^{\alpha}\exp(-z)}{x\Gamma(\alpha)}$	$F(x) = \frac{\gamma(\alpha, z)}{\Gamma(\alpha)}$
Generalized Pareto	$\theta > 0$	$\xi > 0$		$f(x) = \frac{1}{\theta}(1 + \xi z)^{-1-\frac{1}{\xi}}$	$F(x) = 1 - (1 + \xi z)^{-\frac{1}{\xi}}$
Inverse Gaussian (Wald)	$\theta > 0$	$\alpha > 0$		$f(x) = \frac{1}{\theta}\sqrt{\frac{\alpha}{2\pi z^{3}}}\exp(-\frac{\alpha(z-1)^{2}}{2z})$	$F(x)$ $= \Phi\left((z-1)\sqrt{\frac{\alpha}{z}}\right)$ $+ \Phi(\left(-(z\right.$ $\left.+ 1)\sqrt{\frac{\alpha}{z}}\right)\exp(2\alpha)$
Lognormal	$-\infty \leq \mu \leq \infty$	$\sigma > 0$		$f(x)$ $= \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\ln x - u}{\sigma}\right)^{2}\right)$	$F(x) = \Phi\left(\frac{\ln x - u}{\sigma}\right)$
Pareto	$\theta > 0$	$\alpha > 0$		$f(x) = \frac{\alpha\theta^{\alpha}}{(x + \theta)^{\alpha+1}}$	$F(x) = 1 + \left(\frac{\theta}{x + \theta}\right)^{\alpha}$

Notes:

- $z = \frac{x}{\theta}$
- θ denotes the scale parameter for all the distributions.
- $\gamma(a, b) = \int_0^b t^{\alpha-1}\exp(-t)\,dt$, the lower incomplete gamma function.
- $\Phi(y) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{y}{\sqrt{2}}\right)\right)$, the standard normal cumulative density function.
- The function $a(x, \phi)$ does not have an analytical expression and is evaluated using series expansion methods.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mentje Gericke
+27 182992567
mentje.gericke@nwu.ac.za