Paper 3359-2015

Modelling Operational Risk using Extreme Value Theory and Skew t-copulas via Bayesian Inference using SAS®

Betty Johanna Garzon Rozo, Business School, The University of Edinburgh, UK;
Jonathan Crook, Business School, The University of Edinburgh, UK;
Fernando Moreira, Business School, The University of Edinburgh, UK

ABSTRACT

Operational risk losses are heavy tailed and likely to be asymmetric and extremely dependent among business lines and event types. We propose a new methodology to assess, in a multivariate way, the asymmetry and extreme dependence between severity distributions and to calculate the capital for operational risk. This methodology simultaneously uses several parametric distributions and an alternative mix distribution (the lognormal for the body of losses and the generalized Pareto distribution for the tail) via the extreme value theory using SAS®; the multivariate skew t-copula applied for the first time to operational losses; and the Bayesian inference theory to estimate new n-dimensional skew t-copula models via Markov chain Monte Carlo (MCMC) simulation. This paper analyzes a new operational loss data set, SAS® Operational Risk Global Data (SAS OpRisk Global Data), to model operational risk at international financial institutions. All of the severity models are constructed in SAS® 9.2. We implement PROC SEVERITY and PROC NLMIXED and this paper describes this implementation.

INTRODUCTION

Operational Risk has played a decisive role over the last decade across the bank industry given the significant losses due to operational failures that the financial sector has suffered. Thus, operational risk has become as important as credit risk or market risk. The Basel II accord (2004) allows banks to estimate the regulatory capital that covers their annual operational risk exposure (total Operational Value at Risk - OpVaR) using their own models via the advanced measurement approach (AMA). Under the AMA, the loss distribution approach (LDA) has emerged as one of the most convenient statistical methods to calculate OpVaR.. However, significant problems have arisen for this standard approach. First Operational losses almost never fit a parametric distribution; the main reason of this is their inherently elusive nature: high-frequency low-severity and low-frequency high-severity (McNeil et al. 2005; Racheddi and Fantazzini 2009). Moscadelli (2004) and De Fontnouvelle et al. (2006) show that the tails of the loss distributions functions are in first approximation heavy-tailed Pareto-type. However, because of the high quantile level requirement (99.9%) for the OpR capital charge, precise modelling of extremely high losses is critical. Second, publicly available operational losses datasets possesses several issues. One of the most significant is that the data is collected from a minimum threshold. In this sense the resulting dataset is incomplete and left-truncated. Fitting an unconditional distribution to the observed (incomplete, truncated data) losses would lead to biased estimates of the parameters of both severity and frequency distributions, as shown in Chernobai et al. (2005a).

In brief, several difficulties still remain concerning these advanced modelling techniques. The main points of these recognized problems are as follows: (i) the distribution of such losses has a heavy tail and a major issue is to choose the modelling strategy that accounts for shape and truncation and (ii) the unrealistic and deficient aggregation of event types (ET) or business lines (BL) through a simple conservative sum if the dependence between ET is not identified. To address these problems, this paper attempts to develop a new methodology that integrates the use of extreme value theory (EVT) for modelling the loss severities distribution and skew *t*-copulas functions for undertaking the dependence structure between ET in *n*-dimensions.

This paper presents a complete procedure using SAS® to fill the first gap. All of the models related to the construction of severities are made in SAS® 9.2 and the specific implemented procedures are PROC SEVERITY and PROC NLMIXED.

DATA

This study analyses an updated operational loss data set, SAS® Operational Risk Global Data (SAS OpRisk Global Data), to model operational risk at international financial institutions. The data corresponds to the September 2013 release of SAS OpRisk. The data set contains 29,374 operational losses over the period 1900 to 2013 for firms both across the U.S and outside the U.S. Since the data set is culled from several public sources, the loss announcements are not the firms themselves. Therefore, we needed to identify any possible problems with the quality of the data in order to prepare an adequate dataset. The applied filters to the raw data ensure that losses included in the analysis occurred at financial services firms and the reported loss amounts are reliable. Table 1 summarizes the construction process of the final dataset and the filters applied.

The September 2013 release of SAS OpRisk Global Data produces nominal values for losses that are based on the average CPI for the year with July 2013 as the last month reported by the Bureau of Labor Statistics. Each event in the dataset is associated with a starting year of the loss, last year, and month and year of settlement. We use the starting year to record the existence of an event.

Graphs 1 and 2 illustrate the frequency of loss events in the U.S. and outside the U.S. respectively, aggregated by year during 1980-2013, after filters 1 to 3 were applied. It is clear from both graphs that the frequency of operational risk losses experienced a sharp decline after 2007. This may be explained by the fact that losses could be registered only after several years of occurrence, hence the last years are underpopulated. The events from period 2010-2013 represent only 1.5% and 3.5% of the total amount of losses for each group. Therefore, to diminish the effect of lack of information in the last four years, we limit the sample to the period 1980-2009.

Graph 3 shows the distribution of the number of events across the regions, after filters 1 to 5 were applied. The dataset reports country codes linked to each individual event. The large proportion of events in North America may be explained by the facilities of covering events in this country in comparison to the other regions. The geographic distribution of events supports the decision to exclude firms outside the U.S. in order to get homogeneity of coverage.

Table 2 shows the distribution of the number of loss events across the Business Lines and Event Types and their intersections in the U.S. after applying filters 1 to 6. The shading acts as a heat map and represents the contribution to the total number of losses. The top three Business Lines are Retail Banking, Commercial Banking and Retail Brokerage. Although these three Business Lines account for 72% of the number of events the degree of concentration is not as great as for the Event Types. The top three event types are Clients, Products & Business Practices, External Fraud and Internal Fraud, which account for 94%.

Table 3 shows the distribution of the total gross loss amount across the Business Lines and Event Types and their intersections in the U.S. The shading acts as a heat map and represents the contribution to the total gross loss amount after filters 1 to 5 were applied. The top three Business Lines are Retail Banking, Corporate Finance and Asset Management. These three Business Lines account for 73% of the gross loss amount. Even though the figures are similar to table 2, the gross loss amount is less concentrated among business lines than the number of losses. The top Event Type is Clients, Products & Business Practices that accounts for 87% of the total gross loss value. It is clear that the concentration for Event Type is greater than for the Business Lines in both tables 3 and table 2. Following a comparison with table 2, it is deduced that losses follow the behavior High Frequency Low Impact, especially for External Fraud and Internal Fraud, and Low Frequency High Impact, especially for Asset Management and Corporate Finance. See table 4.

METHODOLOGY

Operational risk losses are characterized by high-frequency low-severity and low-frequency high-severity as was reflected in the introduction. To model these distributions, this paper considers the implementation of the well-known EVT. Therefore, we use the mixed tail distribution that SAS provides: SAS's LOGNGPD[1] function; even though we employ this mix distribution, we do not restrict our models to the LognormalGpd, but consider eight[2] alternative distributions as well. We test simultaneously several parametric distributions and the mix distribution for each business line.

In the LOGNGPD function the parameters x_r and p_n are not estimated with the maximum likelihood method used by PROC SEVERITY, so we need to specify them as *constant* parameters by defining the *dist*_ CONSTANTPARM subroutine. The parameter p_n is fixed to 0.8 by default. However, we use several values (0.9, 0.8, 0.7, 0.6) in order to get a bigger variety of models and to draw a comparison between them.

We estimated the distributions for the vectors of difference in logs of losses (YBL_j). Suppose $X_1, \ldots, X_r$ are *iid,* X_j denotes a vector of operational loss amounts, where $j = 1, \ldots, r$ and r is the number of business lines, in our case $r = 8$, $j = 1, \ldots, 8$; u is a high threshold, in our case $u = \$1M\ (USD)$ one million. Define the excesses losses by $Y_j = X_j - u$ and define Y_j^* as the difference in logs of losses $Y_j^* = LnX_j - Lnu$ with distribution function, df, F^*. Applying the transformation to the data we obtain the vectors $Y_1^*, \ldots, Y_8^*$, but to simplify notation we shall refer to these vectors as YBL_j. However, for drawing comparisons we also model the distributions of the vectors of losses (X_j) and excess losses (Y_j). Thus we present a complete comparison analysis between the distributions for the three different vectors: losses X_j (the data as it is), Y_j (the excess losses) and YBL_j (log of excess losses).

We model severities for each business line using SAS® 9.2. SAS provides a complete procedure PROC SEVERITY[3] that enables us to do the general procedure above mentioned. Seven different statistics of fit were used as selection criteria[4]. The section Code Part 1 illustrates the complete procedure for the vector YBL_6 to fit the multiple predefined distributions, to construct the LOGNGPD distribution and to run goodness of fit across all the distributions. For the other business lines the code was replicated.

CORRECTING FOR REPORTING BIAS IN LOSS SEVERITY

Our data for operational losses are incomplete and left-truncated with a fraction of the data around the lowest quantiles missing, both the number of losses and the values. Fitting unconditional distributions to the observed (incomplete, truncated data) losses would lead to biased estimates of the parameters of both severity and frequency distributions, as shown in Chernobai et al. (2005a). They call this a "naive" approach. As a consequence of this bias, it has been shown that the resulting measure of the risk capital (OpVaR) would be miscalculated (Chernobai et al. 2005a, b). However, some existing empirical evidence suggests that the left-truncation of the data is disregarded when modelling the frequency and severity distribution. The reason given is that the VaR is directly influenced by the upper rather than lower quantiles of the loss distribution. Nevertheless, we found in the literature that to ignore the threshold is wrongly justified (Chernobai et al 2005a, b, c, Giacometti et al. 2007).

---

1 Lognormal distribution function to model the body and a GPD to model the excess distribution above a certain threshold.

2 The distributions tested for each business line were: Weibull, Lognormal, GPD (Generalized Pareto), Exponential, Burr, Gamma, Igauss (inverse Gaussian), Pareto and the mix distribution LogNormalGPD (LOGNGPD).

3 PROC SEVERITY computes the estimates of the model parameters, their standard errors, and their covariance structure by using the maximum likelihood method for each of the distribution model.

4 They are log likelihood, Akaike's information criterion (AIC), corrected Akaike's information criterion (AICC), Schwarz Bayesian information criterion (BIC), Kolmogorov-Smirnov statistic (KS), Anderson-Darling statistic (AD), and Cramér-von Mises statistic (CvM).

As discussed previously, our data consist of a series of operational losses exceeding one million dollars in nominal value. Therefore, we consider the phenomenon of data truncation in order to achieve consistent estimations of the parameters for severity and the frequency distribution. We review the methodology suggested in Chernobai et al. (2005a). They call this method the "conditional" approach, in which the losses are modelled with truncated (conditional) distributions. The severity is estimated using the following conditional density

$$f_\theta^c(x) = f_\theta(x|X \geq H) = \begin{cases} \frac{f_\theta(x)}{1-F_\theta(H)}, & x \geq H \\ 0, & x < H \end{cases} \tag{1}$$

where H is the threshold ($1M)

$\boldsymbol{\theta}$ is the unknown parameter set

$f_\theta(x)$ is the probability density function (pdf)

$F_\theta(x)$ is the cumulative probability function (cdf)

the superscript "c" indicates "conditional"

The general idea is that the proportion of missing data can be estimated on the basis of the respective value of the severity distribution, i.e. the fraction of the missing data is equal to the value $F_\theta(H)$. Once the conditional probability density $\left(f_\theta^c(x)\right)$ is calculated the unknown conditional parameter set $\left(\hat{\theta}^c\right)$ can be calculated in two alternatives ways.

Using Maximum Likelihood Estimation (MLE) procedure, the unknown parameter set is estimated by directly maximizing the constrained log-likelihood function:

$$\hat{\theta}_{MLE}^c = \arg\ \max_\theta \log \prod_{j=1}^{r} \frac{f_\theta(x)}{1 - F_\theta(H)} \tag{2}$$

Using the expectation maximization algorithm. We refer to Dempster et al. (1997), McLachlan and Krishnan (1997) and Meng and van Dyck (1997).

In this paper we use MLE for estimating the unknown conditional parameter; specifically we implement PROC NLMIXED. Rearranging terms leads to the fitted distribution function of the observed sample of the following form

$$\hat{F}^c(x) = f(x) = \begin{cases} \frac{\hat{F}_\theta(x)-\hat{F}_\theta(H)}{1-\hat{F}_\theta(H)}, & x \geq H \\ 0, & x < H \end{cases} \tag{3}$$

So that $\hat{F}_\theta(x) \sim U\left[\hat{F}_\theta(H), 1\right]$ and $\hat{F}^c(x) \sim U[0, 1]$ under the null that the fitted distribution function is true. The true severity distribution function remains unchanged for every data point. However, as we found in the empirical results, equation 1 does not provide a conditional cumulative distribution from 0 to 1 but from a value close to cero. Therefore, an adjustment need to be applied. Next section provides details of this adjustment.

RESULTS AND ANALYSIS

For illustrating purposes we only explain the comparative results for the business line 6. Nevertheless, the resulting models distributions and goodness of fit statistics for X_j and Y_j for $j = 1, .., 8$ (business lines)

present a similar pattern to those reported for business line 6[5] in table 5a. As a result, the inference showed for this particular business line may extend to the other seven.

As table 5a and graph 4 illustrate the results are remarkably different between the vector excess log losses (YBL_6) and the excess losses (Y_6) or between (YBL_6) and the losses (X_6). First, information about the input data set is displayed followed by the "Model Selection Table" and at the bottom “All fit Statistics Table” is shown in Table 5a. It is clear there are large differences in both: the range of values between YBL_6 and Y_6 (maximum 9.39 vs 11,970 respectively) and the standard deviation (1.96 vs 544.39). Thus, we can infer that the scale managed affects the results considerably. The model selection table displays the convergence status, the value of the selection criterion, and the selection status for each of the candidate models. For the vector YBL_6, the Weibull distribution model is selected, because it has the lowest value for the selection criterion, whereas the distribution that presents a better fit for Y_6 and X_6 is the Burr distribution. However, the Burr distribution is the second best model for the vector YBL_6, which indicates the behavior of the distribution of the loss vector is kept regardless the log transformation. The table “All fit statistics” prompts further evaluation of why the model Weibull and the Burr distribution were selected. This table indicates for instance that for the vector YBL_6 the Weibull model is the best according to several statistics (the likelihood-based, AIC, AICC, BIC, AD and CvM). However, the Burr model is the best according to the KS statistics. For the vector Y_6 the closest contender to the Burr distribution is the Gpd distribution, whereas for X_6 there are not contender.

Table 5b presents the model selection table for the vector excess log losses (YBL_j) across all business lines $(j = 1, \dots, 8)$, which are the loss severity distributions of interest. For the vectors YBL_1, YBL_4, YBL_6 and YBL_7 the Weibull distribution model is selected, because it has the lowest value for the selection criterion, whereas the distribution that presents a better fit for YBL_2, YBL_3, YBL_5 and YBL_8 is the LognormalGpd distribution. The following points should be noted regarding the latter distribution. As we mentioned earlier we implement several values for the parameter p_n (i.e. 0.9, 0.8, 0,7 and 0.6) in order to test and get better fits. For instance, the value $p_n = 0.9$ provides the lowest value for the selection criterion in the vector YBL_2, consequently this distribution at this value parameter is selected rather than 0.7 or 0.8. The same situation can be observed for YBL_8. Conversely, vectors YBL_3 and YBL_5 present the best fit at $p_n = 0.7$, which means that 30% of the severity values tend to be extreme as compared to the typical values. These results suggest that business lines 3 and 5 possess the largest tail losses. In general, observe that the Weibull distribution is a short-tailed distribution with a so-called *finite right endpoint*. Weibull distribution is the best fit for business lines 1, 4, 6 and 7. On the contrary, the Lognormalgpd distribution is a long-tailed distribution and modelled appropriately the extreme values of business lines 2, 3, 5 and 8.

The last table Goodness of fit for Log excess losses (table 5c) shows that the model with the LOGNGPD distribution has the best fit according to almost all the statistics of fit for the four business lines 2, 3, 5 and 8 and the model with the Weibull distribution has the best fit according to almost all the statistics of fit for the business lines 1, 4, 6 and 7. The Weibull distribution model is the closest contender to the LOGNGPD model in business lines 3, 5 and 8, for business line 2, the Exponential is the closest. For the Weibull model the Burr distribution fits the data very well for the business line 1, 4, 6 and 7. For the last business line the exponential is also a close model. According to the literature, we expected that the LogNormalGPD distribution to be the most appropriate for modelling all the losses among business lines. However, our results indicate that this is not the case for all the business lines. The most striking result is that severities are not necessarily identically distributed. Thus, the previous results demonstrate that the assumption of identically distributed severities may be erroneous. Finally, Graph 5 provides 8 comparative plots which corroborate the differences among models and visualize the explained results from tables 5b and 5c. The plots are organized in two groups: in the top vectors which follow the Weibull distribution, and in the bottom vectors modeled by the Lognormalgpd distribution.

5 We present the results for the Business Line 6 because this one has the highest number of events (frequency). Then, this business line is a good example for illustrating purpose.

CORRECTING BIAS IN SEVERITIES

In this section we applied the correction for the reported bias. Firstly, we prove that if the theoretical threshold H does not coincide with the minimum value of the vector of losses then the resulting conditional cumulative distribution does not start from 0 but from a value near to zero. Therefore, we replace the expression $F_\theta(H)$ with $F_\theta(x_{min})$ in equations 1, 2 and 3 in order to get $\hat{F}^c(x) \sim U[0,1]$, i.e. a cumulative distribution which goes from 0 to 1. Secondly, we examine the differences between the proportions of missing data in the vector of losses (X_6) (i.e. original data) and the vector of excess log losses (YBL_6). The purpose of this exercise is to visualise the effects of the bias when the logarithms of the values are used rather than the actual values. Finally, the estimation of: (i) the set of parameters estimated by MLE for business line 6, (ii) the conditional proportion of missing data, (iii) the set of conditional new parameters estimated by MLE for $j = 6$ and (iv) the corrected cumulative distributions are reported for the vector of excess log losses YBL_6.

Using an empirical threshold. To begin with we analyze the vector of losses $j = 6$ (X_6 scale in \$M). We fix the threshold *H* equal to \$1M (USD) and to take the parameters obtained from the distribution that fits this loss vector best, the Burr distribution (see results in table 5a). The parameters of the distribution of the vector of losses X_6 and the value of the missing data $F(H)$ are presented in table 6. Then we implemented the PROC NLMIXED in SAS for estimating the unknown conditional parameter set $(\hat{\theta}^c)$. This procedure fits nonlinear mixed models—that is, models in which both fixed and random effects enter nonlinearly. PROC NLMIXED enables us to specify any continuous distribution, having either a standard form (Burr, Weibull) or the conditional distribution function (eq. 1) which we code using SAS programming statements.

Four comparative plots are prepared in Graph 6. These plots enable us to visually verify how the models (CDF vs Conditional CDF at H=1) differ from each other. The plot in the top left side displays the full cumulative distribution function (CDF) for the vector X_6. In the top right side we can see an expanded image of the same distribution where it is clear what data is missing (i.e. *F(H)* = 0.035919). The plot in the bottom left presents a zoom of the conditional cumulative distributions $\hat{F}^c(X_6)$ and the plot in the bottom right shows a bigger zoom of $\hat{F}^c(X_6)$, where we can see clearly that the Conditional CDF at H=1 does not start at 0, as it should. We re-estimated the whole procedure adjusting to the real threshold. Table 8 and graph 7 evidences that the new reached value by $\hat{F}^c(x)$ is zero. Table 8 demonstrates the parameter values of the fitted distributions to the vector of losses X_6 and the estimated fraction of the missing data F(H), under the 'naive' and the conditional approaches.

Table 9 reports the MLE estimates, the value of the maximised log-likelihood function, and the estimate of the proportion of missing data for the distribution of the vectors of excess log losses YBL_6. As we can see the proportion of missing data for YBL_6 (0.0011) is significantly lower than for X_6 (0.036). The MLE procedure converges and both parameters (shape and scale) slightly decreased. Graph 8 illustrates in the top left the cumulative distribution for the vector YBL_6 under the naive approach. The plot in the top right is a closest image of the CDF, which visualizes that the starting point of the CDF is not zero. The plots in the bottom are the conditional CDF. In the left the full conditional CDF is presented and in the right a closer image is shown. We can visually corroborate that the $\hat{F}^c(YBL_6)$ starts from zero as is evidenced by the plot in the bottom right. Finally, same procedure was applied to the vectors excess log losses $YBL_j \; for \; j = 1, \ldots, 8$. We fit the conditional density to each business line.

CONCLUSION

The described procedure for modelling severities enables us to achieve multiple combinations of the severity distribution and to find which fits most closely. Hence, we achieve an accurate estimation of the whole severity losses distribution. We show that the model with the natural logarithm of operational excess losses is more feasible for two main reasons (i) presentation of data on a logarithmic scale can be helpful when the data covers a large range of values, our dataset contains values of losses from \$1Million to \$21,000M (USD); (ii) the use of the logarithms of the values rather than the actual values reduces a wide

range to a more manageable size. We present a comparison analysis between the distributions for the three different vectors: losses X_j (the data as it is), Y_j (the excesses losses) and YBL_j (log of excesses losses). The attained results provide convincing evidence to suggest that the selection of modelling excesses log losses is appropriated. We refer to table 5a.

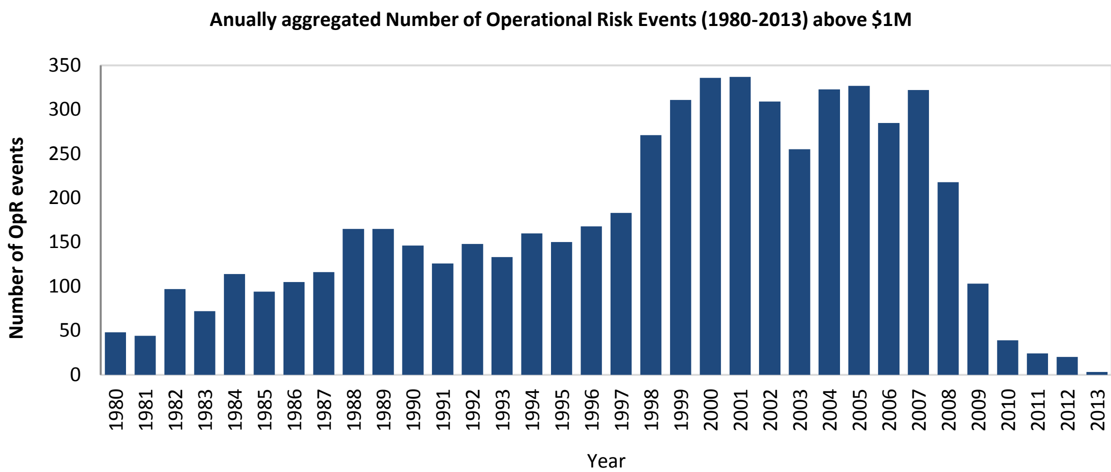
In the correction for reporting bias we prove that using the empirical threshold instead the theoretical threshold allows us to get a conditional cumulative probability function starting from zero. Thus, we achieve an accurate set of MLE parameters. Further, under the conditional approach the scale parameters (if relevant) are decreased, the shape 1 parameters increased and the shape 2 parameters (if relevant) decreased under the correct model.

LIST OF TABLES AND GRAPHS

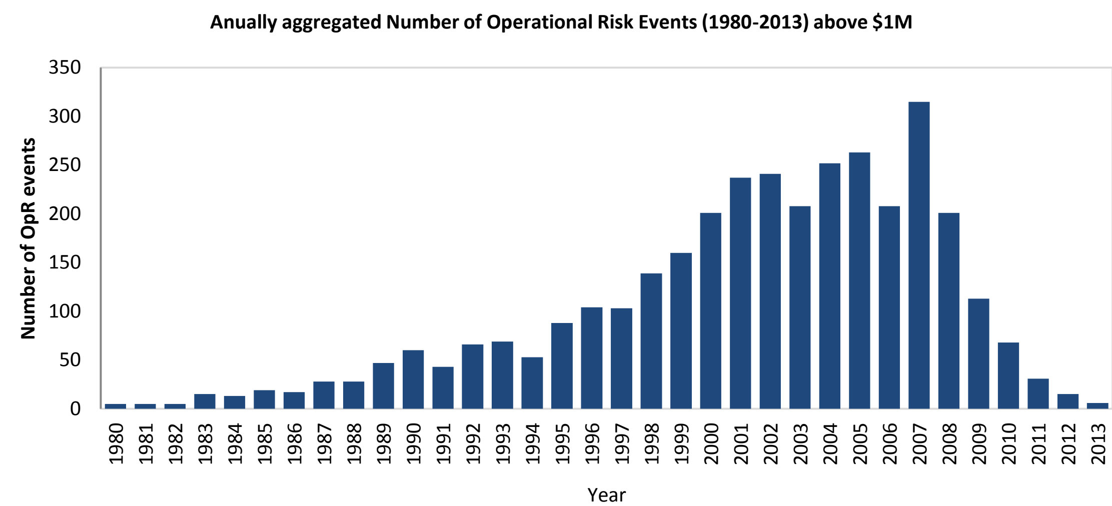
Selection Procedure	
Description of the filter	Reason
1. OpR events occurring in the Financial Industry (NAICS[6] Industry Code 52)	We are interested just in the Financial sector
2. OpR events occurring after 1980	There are relatively few earlier observations. Data from 1900-1979 represent only 2.2% of the total dataset.
3. Limited to losses exceeding \$1 M (USD)	The dataset was collected from public places and may not include all of a bank's internally recorded Operational Losses, particularly those of a smaller magnitude. It also looks like large losses are more suitable for understanding and assessing the OpRisk exposure (De Fontnouvelle 2006).
4. Extract loss events in early 2013 and limit the sample to events originating between the beginning of 1980 and the end of 2009	Since losses may take months or even years to materialize, it is likely that many events are currently taking place but have not yet been discovered. So the last several years of the database may be underpopulated. See the trend of Graph 1 and 2
5. Excluding Insurance Business Line	The study is based on Basel's matrix. Therefore we just analyzed the 8 business lines and 7 event types define in the matrix.
6. Exclusion of foreign firms outside U.S	Foreign firms may not be as well covered as U.S. firms. Electronic archiving for foreign publications may be more recent than for U.S. publications, and the vendors may not possess the language skills necessary for checking all foreign publications. Also media coverage of small firms may diverge from that of large firms. Taking into account only events in the U.S. ensures some homogeneity in the sample. 60% of events occurred in the U.S. after filters 1 to 5 were applied. See graph 3

Table 1. Filters applied to the dataset

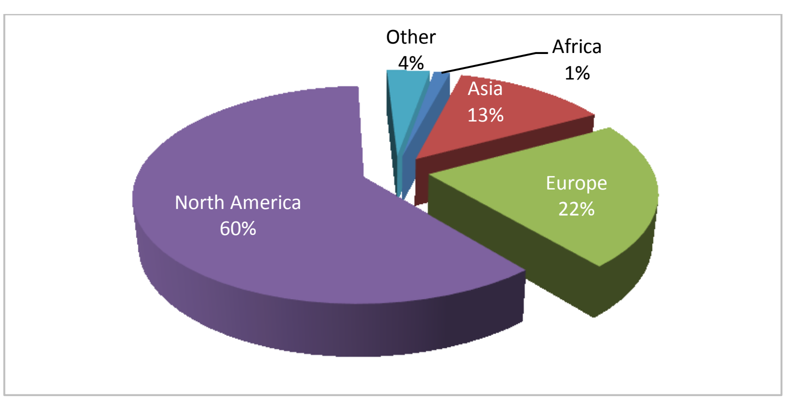
6 The NAICS is the standard used by Federal statistical agencies in classifying business establishments for the purpose of collecting, analysing, and publishing statistical data related to the U.S business economy.



Graph 1. Operational Risk Event Frequency U.S



Graph 2. Operational Risk Event Frequency Outside U.S



Graph 3. Distribution of Number of Loss Events by Region (1980-2009)

Number of Events	B D & S F	C, P & B P	D to P A	E P & W S	E, D & P M	E F	I F	Total	% of Total
Agency Services	0	55	0	0	7	7	5	74	2%
Asset Management	0	260	1	6	6	11	52	336	8%
Commercial Banking	3	197	7	19	16	432	128	802	19%
Corporate Finance	0	261	0	20	0	12	19	312	7%
Payment and Settlement	2	46	1	1	12	12	13	87	2%
Retail Banking	3	664	16	42	23	505	444	1,697	40%
Retail Brokerage	0	398	1	43	22	20	116	600	14%
Trading & Sales	1	306	0	13	14	5	46	385	9%
Total	9	2,187	26	144	100	1,004	823	4,293	
% of Total	0%	51%	1%	3%	2%	23%	19%		

1% – < 5% | 5% – 10% | > 10% of total

BD & SF, Business Disruption and System Failures, C,P &BP, Clients, Products & Business Practices, D to PA, Damage to Physical Assets, EP &WS, Employment Practices & Workplace Safety, ED & PM, Execution, Delivery & Process Management, EF, External Fraud, IF, Internal Fraud

Table 2. Distribution of Frequency of Operational Risk Events by Business Line by Event Type in U.S. (1980-2009)

$M (USD)	B D & S F	C, P & B P	D to P A	E P & W S	E, D & P M	E F	I F	Total	% of Total
Agency Services		6,580			281	2,673	309	9,843	3%
Asset Management		57,252	95	68	128	454	1,356	59,354	17%
Commercial Banking	312	10,017	1,770	350	259	7,242	6,224	26,174	7%
Corporate Finance	0	64,740		420		980	2,703	68,843	19%
Payment and Settlement	48	1,686	2	1	171	248	295	2,453	1%
Retail Banking	251	116,121	213	1,604	275	3,309	7,142	128,916	36%
Retail Brokerage		15,524	4	2,043	167	118	2,423	20,280	6%
Trading & Sales	9	35,352		414	529	126	2,317	38,748	11%
Total	620	307,274	2,085	4,901	1,811	15,150	22,771	354,610	
% of Total	0%	87%	1%	1%	1%	4%	6%		

1% – < 5% | 5% – 10% | > 10% of total

Table 3. Distribution of Gross Losses by Business Line by Event Type in U.S. (1980-2009)

	Table 2a	Table 3a		Table 2a	Table 3a
Business Line	Low Frequency	High Impact	Event Type	High Frequency	Low Impact
Asset Management	8%	17%	External Fraud	23%	4%
Corporate Finance	7%	19%	Internal Fraud	19%	6%

Table 4. Comparison between Distribution of Frequency vs Gross of Loss by Business Line by Event Type in U.S

Descriptive Statistics

	YBL6	Y6	X6
Statistics	Ln (X6) - Ln($1 M)	X6 - $1M	X6
Observations	1697	1697	1697
Minimum	0.01	0.01	1.01
Maximum	9.39	11970.77	11971.77
Mean	1.96	74.97	75.97
Stand Deviation	1.6	544.39	544.39

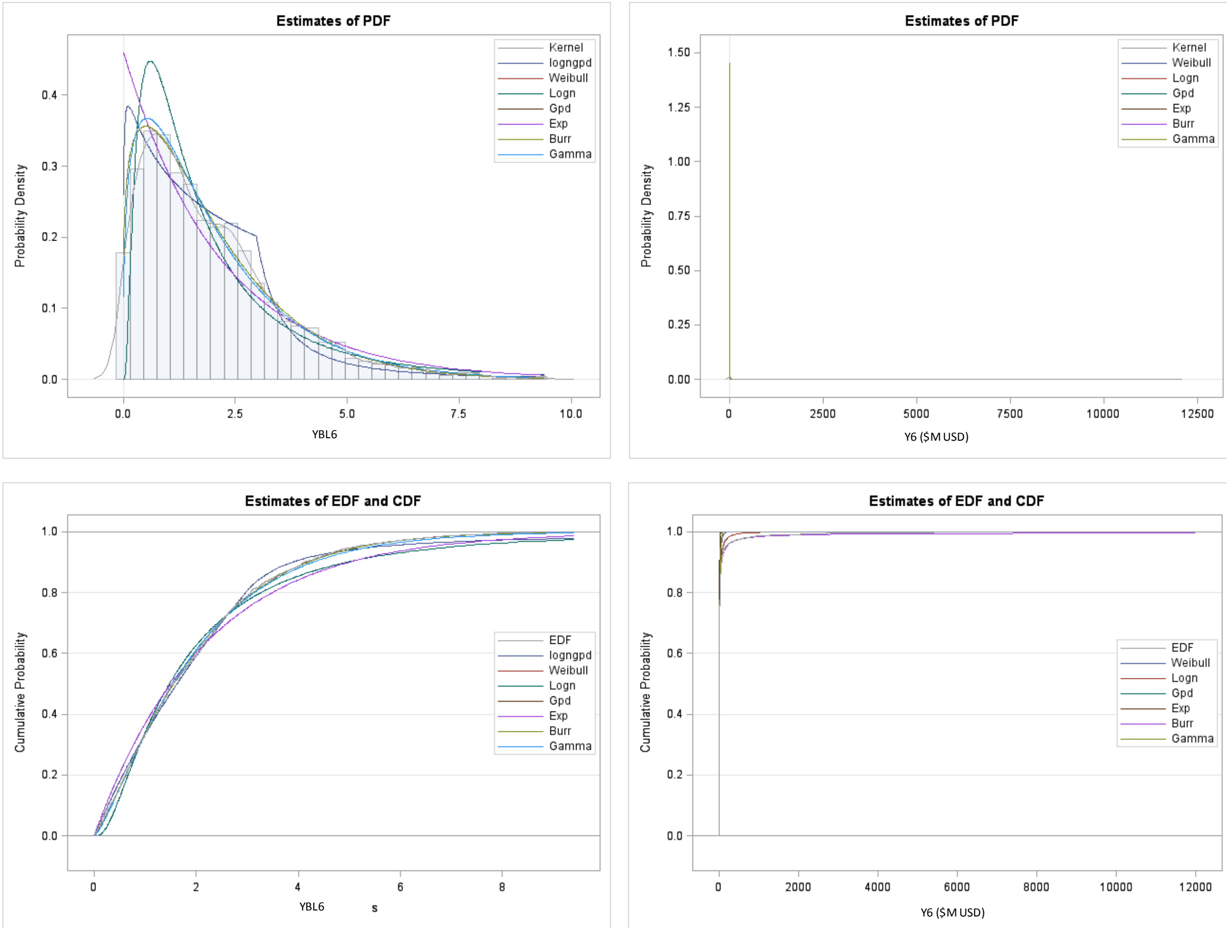
Model Selection Table

	YBL6			Y6			X6		
Distribution	Converged	Custom Object	Selected	Converged	Custom Object	Selected	Converg ed	Custom Object	Selected
Logngpd	Yes	0.2952	No	Yes	4.74555	No	Yes	6.1798	No
Weibull	Yes	0.0672	Yes	Yes	1.7436	No	Yes	6.4364	No
Logn	Yes	1.2173	No	Yes	0.21613	No	Yes	3.0204	No
Gpd	Yes	1.8715	No	Yes	0.06472	No	Yes	3.968	No
Exp	Yes	1.8715	No	Yes	13.5027	No	Yes	8.1232	No
Burr	Maybe	0.0681	No	Yes	0.06471	Yes	Yes	0.2247	Yes
Gamma	Yes	0.1219	No	Yes	2.87092	No	Yes	7.0122	No

All Fit Statistics Table

	YBL6								Y6								X6							
Distribution	Custom	-2 Log Likelihoo	AIC	AICC	BIC	KS	AD	CvM	Custom	-2 Log Likelihood	AIC	AICC	BIC	KS	AD	CvM	Custom	-2 Log Likelihood	AIC	AICC	BIC	KS	AD	CvM
logngpd	0.2952	5701	5711	5711	5738	1.11	4	0.3	4.7456	13366	13376	13376	13403	5.43	40.48	4.75	6.1798	14764	14774	14774	14801	5.75	54.49	6.1892
Weibull	0.06715 *	5592 *	5596 *	5596 *	5607 *	0.71	1.87 *	0.07 *	1.7436	13982	13986	13986	13997	2.77508	63.804	1.754	6.4364	20380	20384	20384	20394	6.126	522.1	6.4544
Logn	1.21727	6093	6097	6097	6107	2.12	26.1	1.21	0.2161	12325	12329	12329	12340	1.07251	5.18966	0.219	3.0204	13197	13201	13201	13212	4.572	44.13	3.0322
Gpd	1.87152	5698	5702	5702	5712	2.22	15.1	1.88	0.0647	12249 *	12253 *	12253 *	12263 *	0.68451 *	1.56884 *	0.065	3.968	13076	13080	13080	13091	5.829	40.04	3.98
Exp	1.87152	5698	5700	5700	5705	2.22	15.1	1.88	13.503	46258	46260	46260	46265	6.73122	2243	13.51	8.1232	37961	37963	37963	37969	5.52	1653	8.1418
Burr	0.06807	5592	5598	5598	5614	0.71 *	1.89	0.07	0.0647 *	12249	12255	12255	12271	0.68756	1.57052	0.065 *	0.2247 *	12430 *	12436 *	12436 *	12453 *	1.484 *	4.13 *	0.224 *
Gamma	0.12193	5608	5612	5612	5623	0.82	2.81	0.12	2.8709	21702	21706	21706	21716	3.79944	501.008	2.884	7.0122	29521	29525	29525	29536	6.01	990.5	7.031

Table 5a. Comparison in model selection between the Loss distributions for: excess log losses (YBL6), excess losses (Y6) and losses (X6)



Graph 4. Comparison between Loss distributions: excess log (YBL6) vs excess (Y6)

Model Selection Table

	YBL1		YBL2		YBL3		YBL4		YBL5		YBL6		YBL7		YBL8	
Distribution	Custom Obj	Selected	Custom Obj	Selected	Custom Obj	Selected	Custom Obj	Selected	Custom Obj	Selected	Custom Obj	Selected	Custom Obj	Selected	Custom Obj	Selected
logngpd 0.9	0.15010	No	0.04736	Yes	0.18532	No	0.07728	No	0.05168	Maybe	0.29520	No	0.43920	No	0.10069	Yes
logngpd 0.7			0.08234	Maybe	0.03692	Yes	0.05609	No	0.03708	Yes					0.16284	No
Weibull	0.05438	Yes	0.16568	No	0.07823	Maybe	0.03638	Yes	0.05613	Maybe	0.06715	Yes	0.05650	Yes	0.37494	No
Logn	0.15376	No	0.61818	No	0.87360	No	0.28155	No	0.18372	No	1.21727	No	0.38020	No	1.13482	No
Gpd	0.79746	No	1.64947	No	3.00964	No	2.36002	No	0.32561	No	1.87152	No	1.30314	No	1.43571	No
Exp	0.79746	No	1.64947	No	3.00964	No	2.36002	No	0.32561	No	1.87152	No	1.30314	No	1.43571	No
Burr	0.05446	No	0.16647	No	0.07943	No	0.03682	Maybe	0.05630	No	0.06807	No	0.05660	No	0.37607	No
Gamma	0.08754	No	0.26814	No	0.19406	No	0.08408	No	0.08139	No	0.12193	No	0.07209	No	0.51017	No

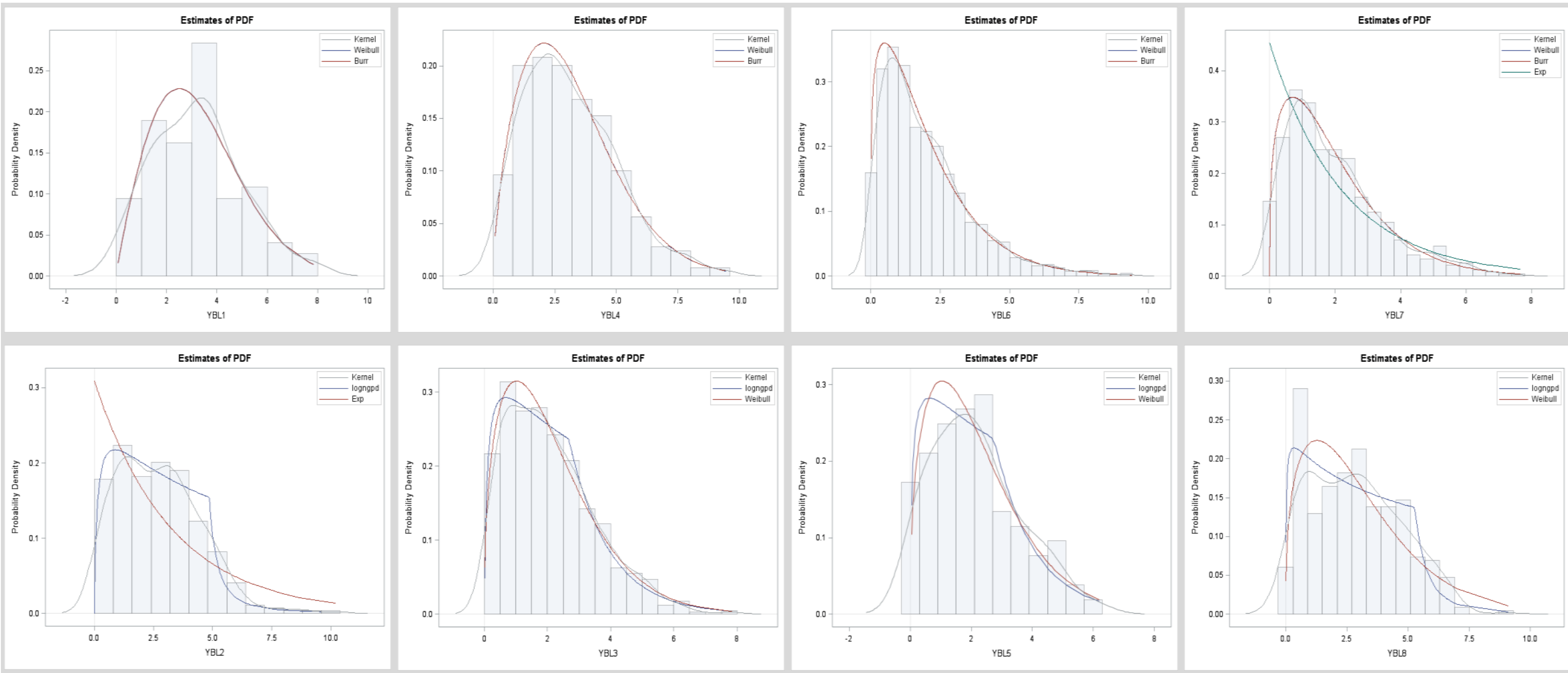
Table 5b[7]. Selection Model for Log excess losses and goodness of fit

7 YBL1 = Agency services, YBL2 = Asset Management, YBL3 = Commercial Banking, YBL4 = Corporate Finance, YBL5 = Payment and Settlement, YBL6 = Retail Banking, YBL7 = Retail Brokerage, YBL8 = Trading and Sales

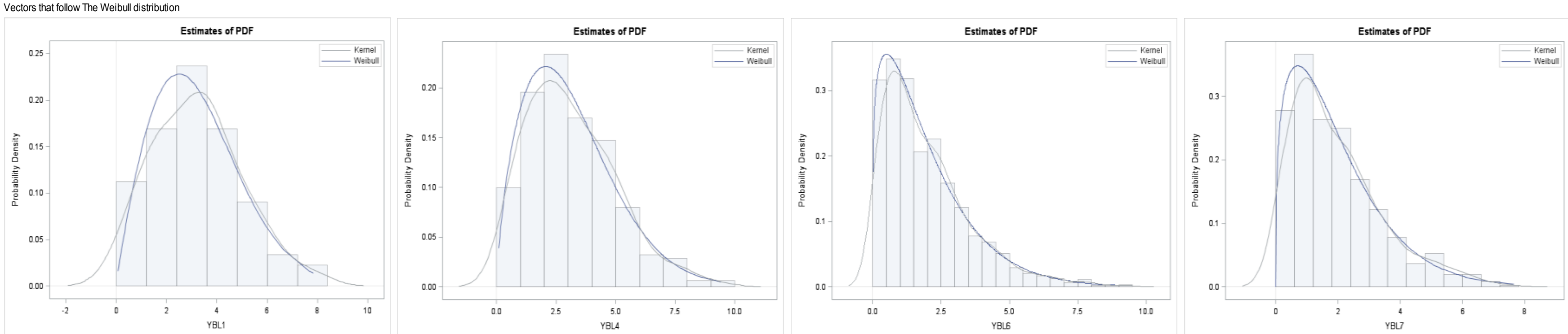
Business Line	All Fit Statistics Table								
	Distribution	Custom	-2 Log Lk	AIC	AICC	BIC	KS	AD	CvM
	logngpd 0.9	0.15010	325	335	336	347	0.75684	1.85620	0.14303
	logngpd 0.7								
	Weibull	0.05438 *	288 *	292 *	292 *	297 *	0.61198 *	0.33684 *	0.05362 *
	Logn	0.15376	331	335	336	340	0.75861	2.12128	0.14532
YBL1	Gpd	0.79746	324	328	328	332	1.45013	4.83323	0.81166
	Exp	0.79746	324	326	326	328	1.45013	4.83323	0.81166
	Burr	0.05446	288	294	294	301	0.61223	0.33762	0.05368
	Gamma	0.08754	296	300	300	305	0.67678	0.76479	0.08362
	logngpd 0.9	0.04736 *	1577	1587	1587	1606	0.53254 *	2.69840 *	0.04768 *
	logngpd 0.7	0.08234 *	1621	1631	1631	1650	0.70612 *	3.26276 *	0.07900 *
	Weibull	0.16568	.	.	.	.	0.91588	3.87053	0.16076
	Logn	0.61818	.	.	.	.	1.52647	12.82981	0.60653
YBL2	Gpd	1.64947	1352	1356	1356	1363	2.43908	13.15085	1.65642
	Exp	1.64947	1352 *	1354 *	1354 *	1357 *	2.43908	13.15085	1.65642
	Burr	0.16647	.	.	.	.	0.91383	3.88201	0.16143
	Gamma	0.26814	1360	1364	1364	1371	1.06876	5.26544	0.26160
	logngpd 0.9	0.18532	2764	2774	2774	2798	1.30902	2.73021	0.18777
	logngpd 0.7	0.03692 *	2631	2641	2641	2665	0.47018 *	0.87401 *	0.03740 *
	Weibull	0.07823	2629 *	2633 *	2633 *	2642 *	0.70407	1.13464	0.07638
	Logn	0.87360	2846	2850	2850	2860	1.70868	12.64882	0.86384
YBL3	Gpd	3.00964	2806	2810	2810	2819	2.71169	22.15918	3.01723
	Exp	3.00964	2806	2808	2808	2813	2.71169	22.15918	3.01723
	Burr	0.07943	2629	2635	2635	2649	0.70664	1.15170	0.07748
	Gamma	0.19406	2651	2655	2655	2664	0.83625	2.48649	0.19030
	logngpd 0.9	0.07728	1231	1241	1241	1260	0.69848	0.83104	0.08014
	logngpd 0.7	0.05609	1226	1236	1237	1255	0.60291 *	0.63869	0.05801 *
	Weibull	0.03638 *	1221 *	1225 *	1225 *	1233 *	0.63228	0.48562 *	0.03614
YBL4	Logn	0.28155	1301	1305	1305	1312	0.96584	4.27679	0.27357
	Gpd	2.36002	1345	1349	1349	1357	2.42019	15.88608	2.37150
	Exp	2.36002	1345	1347	1347	1351	2.42019	15.88608	2.37150
	Burr	0.03682	1221	1227	1227	1239	0.63118 *	0.49051	0.03645
	Gamma	0.08408	1232	1236	1236	1244	0.71189	1.08164	0.08115

Business Line	All Fit Statistics Table								
	Distribution	Custom	-2 Log Lk	AIC	AICC	BIC	KS	AD	CvM
	logngpd 0.9	0.05168 *	304	314	315	327	0.55604 *	0.57842 *	0.05434
	logngpd 0.7	0.03708 *	292 *	302	303	315	0.50666 *	0.47526 *	0.03702 *
	Weibull	0.05613	294	298 *	298 *	303 *	0.63640	0.72377	0.05248
	Logn	0.18372	345	349	349	354	1.00916	4.13016	0.17081
YBL5	Gpd	0.32561	309	313	313	317	1.09947	2.23698	0.33161
	Exp	0.32561	309	311	311	313	1.09947	2.23698	0.33161
	Burr	0.05630	294	300	300	307	0.63714	0.72650	0.05259
	Gamma	0.08139	299	303	304	308	0.71889	1.11848	0.07564
	Logngpd 0.9	0.29520	5701	5711	5711	5738	1.10627	4.00170	0.29695
	logngpd 0.7								
	Weibull	0.06715 *	5592 *	5596 *	5596 *	5607 *	0.71479	1.87061 *	0.06629 *
	Logn	1.21727	6093	6097	6097	6107	2.11862	26.06905	1.20879
YBL6	Gpd	1.87152	5698	5702	5702	5712	2.21773	15.12896	1.87526
	Exp	1.87152	5698	5700	5700	5705	2.21773	15.12896	1.87526
	Burr	0.06807	5592	5598	5598	5614	0.71272 *	1.88589	0.06710
	Gamma	0.12193	5608	5612	5612	5623	0.81807	2.81447	0.12000
	Logngpd 0.9	0.43920	2444	2454	2454	2476	1.45760	10.66206	0.44701
	logngpd 0.7								
	Weibull	0.05650 *	.	.	.	.	0.64025 *	7.78898 *	0.05634
	Logn	0.38020	.	.	.	.	1.36954	17.35249	0.37254
YBL7	Gpd	1.30314	2004 *	2008	2008	2016	2.05064	16.32207	1.30867
	Exp	1.30314	2004	2006 *	2006 *	2010 *	2.05064	16.32207	1.30867
	Burr	0.05660	.	.	.	.	0.64457	7.79932	0.05633 *
	Gamma	0.07209	2025	2029	2029	2038	0.73586	8.34982	0.07040
	logngpd 0.9	0.10069 *	1483 *	1493 *	1493 *	1513 *	0.74088 *	1.63337 *	0.10181 *
	logngpd 0.7	0.16284 *	1493 *	1503 *	1503 *	1522	1.02947 *	2.41536 *	0.15967 *
	Weibull	0.37494	1504	1508	1508	1516 *	1.30131	3.95281	0.36816
	Logn	1.13482	1642	1646	1646	1654	2.14302	11.76488	1.12089
YBL8	Gpd	1.43571	1582	1586	1586	1594	2.54424	11.38871	1.44085
	Exp	1.43571	1582	1584	1584	1588	2.54424	11.38871	1.44085
	Burr	0.37607	1504	1510	1510	1522	1.30253	3.96515	0.36922
	Gamma	0.51017	1521	1525	1525	1533	1.50466	5.09513	0.50254

Table 5c. Goodness of fit for Log excess losses



Vectors that follow The Weibull distribution



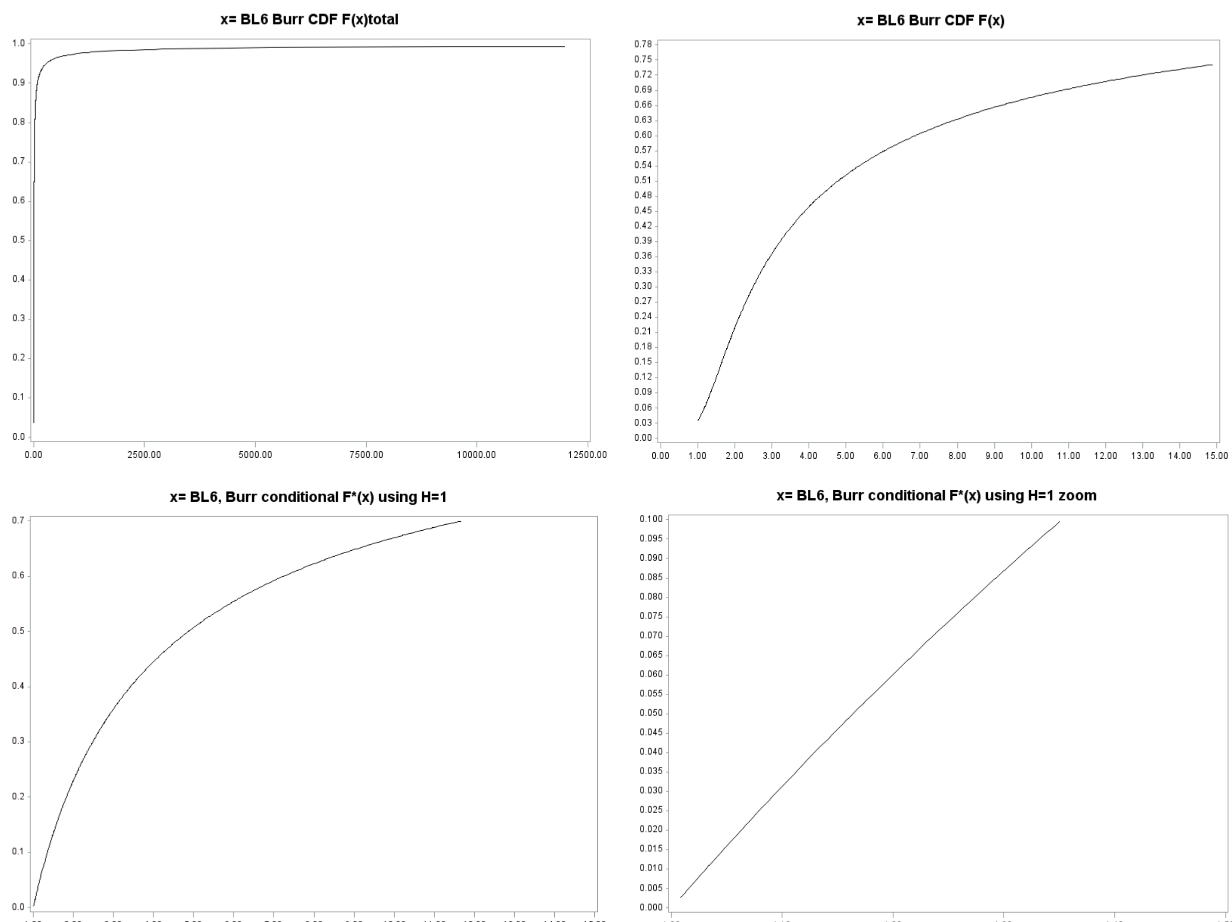
Graph 5. Comparison between Loss excess log distributions: the Weibull vs the Lognormalgpd distribution

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Theta	1.34031	0.31264	4.29	<.0001
Alpha	0.13595	0.14014	0.97	0.3321
Gamma	4.12532	3.76227	1.1	0.2730
F(H)	0.035919			

Table 6. Parameter Estimates for X_6 which follows Burr Distribution and F(H)

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
theta	1.3119	0.2961	1697	4.43	<.0001	0.05	0.7312	1.8927	0.0213
alpha	0.931	0.3529	1697	2.64	0.0084	0.05	0.2389	1.6231	-0.09654
gamma	0.787	0.2361	1697	3.33	0.0009	0.05	0.324	1.25	-0.12498
$\hat{F}^c(H)$	0.0026266								

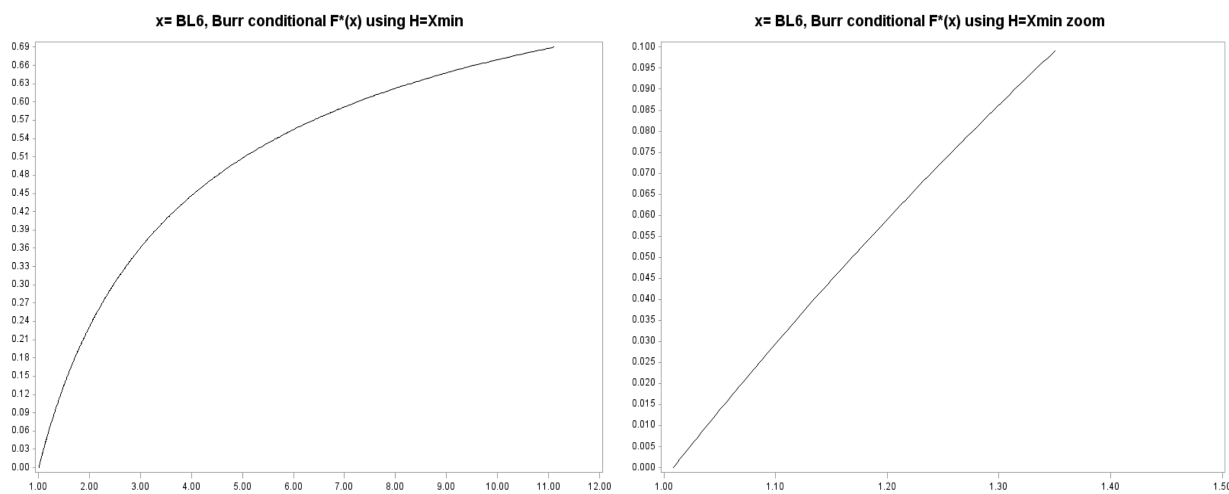
Table 7.Conditional MLE Parameter for X_6 which follows Burr Distribution and $\widehat{F}^c(H)$



Graph 6. CDF and conditional CDF for X_6

Method	MLE Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Low er	Upper	Gradient
Naive approach	Theta	1.34031	0.31264	4.29	<.0001					
	Alpha	0.13595	0.14014	0.97	0.3321					
	Gamma	4.12532	3.76227	1.1	0.2730					
	$\hat{F}(H)$	0.035919								
H = *1*	theta	1.3119	0.2961	1697	4.43	<.0001	0.05	0.7312	1.8927	0.0213
	alpha	0.931	0.3529	1697	2.64	0.0084	0.05	0.2389	1.6231	-0.09654
	gamma	0.787	0.2361	1697	3.33	0.0009	0.05	0.324	1.25	-0.12498
	$\hat{F}^c(H)$	0.0026266								
H = *Xmin*	theta	1.3007	0.3354	1697	3.88	0.0001	0.05	0.6428	1.9586	-0.12226
	alpha	1.0265	0.4125	1697	2.49	0.0129	0.05	0.2174	1.8357	0.139696
	gamma	0.7247	0.2284	1697	3.17	0.0015	0.05	0.2767	1.1727	0.133526
	$\hat{F}^c(H)$	0								

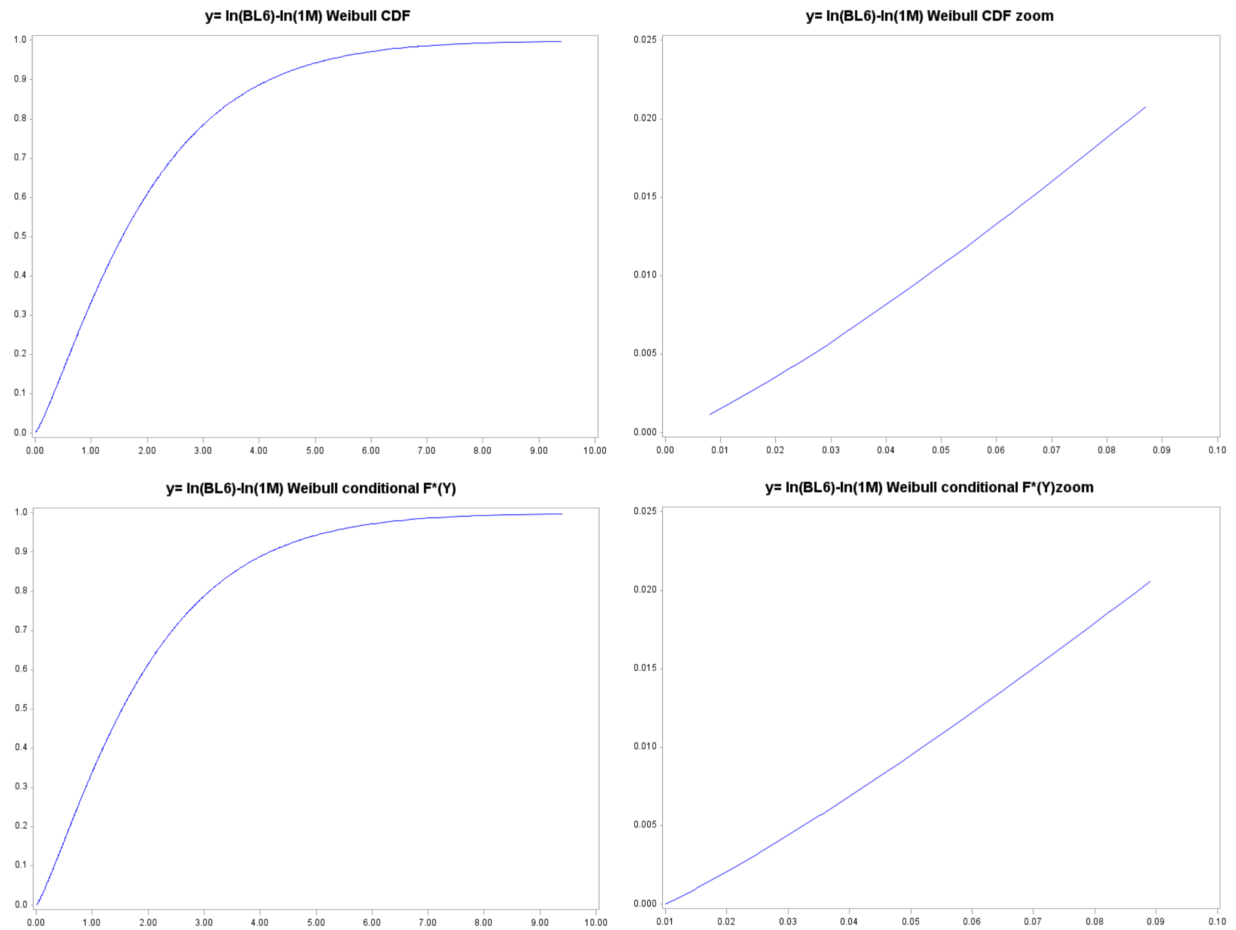
Table 8. Parameter for X_6 under the naive approach, H=1 $\hat{F}^c(H)$ and H=Xmin



Graph 7. CDF and conditional CDF for X_6 at H=Xmin

Method	Parameter	Estimate	Standard Error	t Value	Pr > |t|	Alpha	Low er	Upper	Gradient
Naïve approach	Theta	2.10604	0.11282	18.67	<.0001				
	Tau	1.21248	0.12208	9.93	<.0001				
	$\hat{F}(H)$	0.0011648							
H = *YBL6min*	Theta	2.0806	0.04435	46.91	<.0001	0.05	1.9936	2.1676	0.000013
	Tau	1.2018	0.02346	51.22	<.0001	0.05	1.1558	1.2479	-0.00005
	$\hat{F}^c(H)$	0							

Table 9. Parameter for YBL_6 under the naive approach and $\hat{F}^c(H)$ for H=YBL6min



Graph 8. CDF and conditional CDF for YBL_6 at H = YBL6min

CODE

PART 1. MODELLING SEVERITY DISTRIBUTIONS

Programme 1.1. Fitting multiple predefined distributions

```
/*--- Set the search path for functions defined with PROC FCMP ---*/

  options cmplib=(Sashelp.Svrtdist);
  proc severity data=USdata crit=aicc;
      loss YBL6;
      dist _predefined_;
  run;

  options cmplib=(Sashelp.Svrtdist);
  proc severity data=USdata crit=aicc;
      loss X6;
      dist _predefined_;
  run;
```

Programme 1.2. Constructing a mix distributions Lognormal + GPD = LOGNGPD

```
/***------- Define a mix distribution LOGNGPD -------***/
/*------- Define Lognormal Body-GPD Tail Mixed Distribution -------*/

  proc fcmp library=sashelp.svrtdist outlib=work.sevexmpl.models;
        function LOGNGPD_DESCRIPTION() $256;
            length desc $256;
            desc1 = "Lognormal Body-GPD Tail Distribution.";
            desc2 = " Mu, Sigma, and Xi are free parameters.";
            desc3 = " Xr and Pn are constant parameters.";
            desc = desc1 || desc2 || desc3;
            return(desc);
        endsub;

        function LOGNGPD_SCALETRANSFORM() $3;
            length xform $3;
            xform = "LOG";
            return (xform);
        endsub;

        subroutine LOGNGPD_CONSTANTPARM(Xr,Pn);
        endsub;

        function LOGNGPD_PDF(x, Mu,Sigma,Xi,Xr,Pn);
            cutoff = exp(Mu) * Xr;
            p = CDF('LOGN',cutoff, Mu, Sigma);
            if (x < cutoff + constant('MACEPS')) then do;
                return ((Pn/p)*PDF('LOGN', x, Mu, Sigma));
            end;
            else do;
                gpd_scale = p*((1-Pn)/Pn)/PDF('LOGN', cutoff, Mu, Sigma);
                h = (1+Xi*(x-cutoff)/gpd_scale)**(-1-(1/Xi))/gpd_scale;
                return ((1-Pn)*h);
            end;
        endsub;
```

```
function LOGNGPD_CDF(x, Mu,Sigma,Xi,Xr,Pn);
    cutoff = exp(Mu) * Xr;
    p = CDF('LOGN',cutoff, Mu, Sigma);
    if (x < cutoff + constant('MACEPS')) then do;
        return ((Pn/p)*CDF('LOGN', x, Mu, Sigma));
    end;
    else do;
        gpd_scale = p*((1-Pn)/Pn)/PDF('LOGN', cutoff, Mu, Sigma);
        H = 1 - (1 + Xi*((x-cutoff)/gpd_scale))**(-1/Xi);
        return (Pn + (1-Pn)*H);
    end;
endsub;

subroutine LOGNGPD_PARMINIT(dim,x[*],nx[*],F[*],Ftype,
                            Mu,Sigma,Xi,Xr,Pn);
    outargs Mu,Sigma,Xi,Xr,Pn;
    array m[2] / nosymbols;
    array xe[1] / nosymbols;
    array nxe[1] / nosymbols;

    eps = constant('MACEPS');

    Pn = 0.7; /* Set mixing probability */
    _status_ = .;
    call streaminit(56789);
    Xb = svrtutil_hillcutoff(dim, x, 100, 25, _status_);
    if (missing(_status_) or _status_ = 1) then
        Xb = svrtutil_percentile(Pn, dim, x, F, Ftype);

    /* prepare arrays for excess values */
    i = 1;
    do while (i <= dim and x[i] < Xb+eps);
        i = i + 1;
    end;
    dime = dim-i+1;
    call dynamic_array(xe, dime);
    call dynamic_array(nxe, dime);
    j = 1;
    do while(i <= dim);
        xe[j] = x[i] - Xb;
        nxe[j] = nx[i];
        i = i + 1;
        j = j + 1;
    end;

    /* Initialize lognormal parameters */
    call logn_parminit(dim, x, nx, F, Ftype, Mu, Sigma);
    if (not(missing(Mu))) then
        Xr = Xb/exp(Mu);
    else
        Xr = .;

    /* Initialize GPD's shape parameter using excess values */
    call gpd_parminit(dime, xe, nxe, F, Ftype, theta_gpd, Xi);
endsub;

subroutine LOGNGPD_LOWERBOUNDS(Mu,Sigma,Xi,Xr,Pn);
```

```
            outargs Mu,Sigma,Xi,Xr,Pn;

            Mu    = .; /* Mu has no lower bound */
            Sigma = 0; /* Sigma > 0 */
            Xi    = 0; /* Xi > 0 */
        endsub;
   quit;
```

Programme 1.3. Fitting goodness of test

```
/***----- run a goodness of test for BL6 seven distributions-----***/

   options cmplib=(work.sevexmpl);
   proc severity data=USdata obj=cvmobj print=all plots=pp;
      loss YBL6;
      dist logngpd weibull logn gpd exp burr gamma;

   /* Cramer-von Mises estimator (minimizes the distance *
      between parametric and nonparametric estimates)    */
      cvmobj = _cdf_(YBL6);
      cvmobj = (cvmobj -_edf_(YBL6))**2;
   run;
```

REFERENCES

Basel Committee on Banking Supervision Basel II (2004). International Convergence of Capital Measurement and Capital Standards: a revised framework. Bank for International Settlements. Basel, June. URL www.bis.org

Chernobai, A.S., C. Menn, S. Truck and S. Rachev (2005a). A note on the estimation of the frequency and severity distribution of operational losses. Mathematical Scientist 30(2), 87-97.

Chernobai, A.S., C. Menn, S. Truck and S. Rachev (2005b). Estimation of operational value-at-risk in the presence of minimum collection thresholds. Thecnical Report, University of California, Santa Barbara, CA.

Chernobai, A.S., S.T. Rachev, F.J. Fabozzi (2005c). Composite Goodness-of-Fit Tests for Left-Truncated Loss Samples. Thecnical Report, University of California, Santa Barbara, CA.

De Fontnouvelle P., V. Dejesus-Rueff, J.S. Jordan, and E.S. Rosengren (2006). Capital and Risk: New Evidence on Implications of Large Operational Losses. *Journal of Money, Credit, and Banking,* Vol. 38, No. 7, 1819-1846

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological) 39(1), 1–38.

Giacometti, R., S.T. Rachev, A.S. Chernobai, M. Bertocchi, G. Consigli (2007). Heavy-tailed distributional model for operational risk. *The Journal of Operational Risk* 2(1), 55-90

McLachlan, G., and Krishnan, T. (1997). The EM Algorithm and Extensions (Wiley Series in Probability and Statistics). Wiley, New York.

McNeil, A.J., R. Frey, and P. Embrechts (2005). Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press, Princeton, NJ

Meng, X. L., and van Dyk, D. (1997). The EM algorithm – an old folk-song sung to a fast new tune. Journal of the Royal Statistical Society, Series B (Methodological) 59(3), 511–567.

Moscadelli, M. (2004). The modelling of operational risk: experience with the analysis of the data collected by the Basel Committee. Banca D'Italia, Termini di discussion No. 517

Rachedi O., and D. Fantazzini (2009). Multivariate models for operational risk: A copula approach using Extreme

Value Theory and Poisson shock models. To appear in chapter 7: G.N. Gregoriou (ed.), Operational Risk: Towards Basel III, Best Practices and Issues in Modeling, Management and Regulation. New York: John Wiley & Sons

ACKNOWLEDGMENTS

The author would like to add her thanks to Geoffrey Taylor, SAS Academic Programme Manager, for his valuable assistance in getting access to the data set, and other staff in SAS Institute Inc. for their kind support.

CONTACT INFORMATION

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

Betty Johanna Garzon Rozo
The University of Edinburgh, Business School
29 Buccleuch Place, office 3.02, Edinburgh, UK
Email: bjgr_patric@hotmail.com