

# Operational Loss Scaling by Exposure Indicators: Evidence from the ORX Database

Eric Cope and Abderrahim Labbi  
IBM Zurich Research Lab  
in collaboration with  
The ORX Analytics Working Group\*

October 7, 2008

## Abstract

We investigate whether the size of operational risk losses can be correlated with geographical region and firm size. We extend past studies of losses on firm size by applying quantile regression techniques to obtain a more complete view of scaling relationships across the loss distribution. Furthermore, we test whether these distributional scaling relations can be described using simple location-shift or location-scale shift models. In addition, we develop a novel “quantile matching” algorithm to address statistical issues that arise when estimating loss scaling models when the data are subjected to a loss reporting threshold.

## 1 Introduction

The Operational Riskdata eXchange (ORX) is an international consortium of banks whose primary mission is to enable the collection and mutual exchange of operational loss data. With loss records dating back to 2002, ORX currently has the most comprehensive cross-firm loss database for the banking industry, with more than 90,000 loss records collected from 41 banks, which range widely in size, line of business concentration, and region. In this article, we analyze this database to investigate whether banks’ exposures to operational losses can be characterized on the basis of firm or business line size, or the geographical region in which the bank operates. We investigate in particular whether loss data from such a heterogeneous group of banks can be compared using scaling relationships based on these exposure indicators.

*A priori*, we might expect loss severities to vary according to these indicators for a variety of reasons. For one, larger firms may experience larger losses in some areas, such as in those categories where losses result from legal actions the firm, since in many cases judgments of damages are higher for larger firms. The business, legal, and regulatory environments in which

---

\*Significant contributions of the Analytics Working Group are acknowledged, in particular those of John Walter, Giulio Mignola, Alberto Ferreras, John Jordan, Meirong Li, Mark Piche, Rocco Quartu, Peter Schaller, and Roberto Ugoccioni. We also acknowledge the helpful comments of an anonymous reviewer.

the banks operate also differs from region to region, and may also play a role in explaining regional differences in loss severities. In other cases, the size of the firm or the region in which they operate may signal differences in the internal control environment, as banks of different sizes that are operating in different geographies may have different attitudes to risk exposure and be at different maturity levels with regard to risk management. If larger firms tend to have more mature and entrenched cultures of risk management, then it may well be that in certain loss categories, larger firms experience *less* severe losses.

Characterizing loss scaling relationships in terms of exposure indicators is important for a variety of reasons:

- They provide an explanation for differences observed in loss distributions across various member institutions. Given that the Basel II Accord mandates that external data be used by banks under the AMA approach (Basel Committee on Banking Supervision, 2006), it is critical that financial institutions understand how to calibrate external loss data to their internal data in the process of determining regulatory capital requirements.
- They help to predict changes in a bank's exposure to loss if it undergoes substantial organizational change, such as through a merger or expansion of operations into new geographies.
- Scaling models may lead to the development of cross-industry loss distribution benchmarks that would help banks and regulators to compare the performance of different institutions on a consistent basis.

In our study, we developed statistical models to explain differences in loss severity distributions in each business line and event type category in the ORX database, using regional identifiers and proxies for firm or business line size as predictors. We found that in several cases, there were substantial relationships between the scale of loss distributions and the associated indicators. In particular, we find that there are often significant differences in loss sizes between regions, and that loss sizes may be observed to either increase or decrease with business line or firm size.

Past published studies of loss scaling relationships have focused on characterizing the change in mean loss response to changes in the level of the exposure indicator, using least-squares regression techniques. Our study instead applies quantile regression techniques to provide a more complete picture of the manner in which the entire loss distribution changes as the exposure indicators change. In particular, quantile regression allows one to explore the loss response to changes in exposure indicator levels both in the body as well as in the tails of the loss distributions. Quantile regression techniques further allow us to test whether differences in the estimated loss distributions at different factor levels may be described using either a location-

or location-scale shift. The results of such tests can indicate whether simple transformations are valid for scaling loss data of any size from one setting to another. Furthermore, unlike the prior studies cited in this report, we address the fact that loss data is submitted to ORX only if it exceeds a reporting threshold of €20,000, which can lead to distortions in the statistical estimation of scaling models. We have developed and applied a “quantile matching” technique for estimating the models that reduces these distorting effects, to be described in detail later in this report.

The remainder of the article is organized as follows. We discuss the structure and content of the ORX database in Section 2. Section 3 reviews published work on scaling operational loss data and its relevance to consortia such as ORX. We then develop in Section 4 a series of statistical models for scaling loss data, with a comprehensive method for testing and estimating these models through the use of quantile regression techniques. We present results from selected loss categories in Section 5, and conclude by summarizing the major findings of the study in Section 6.

## 2 The ORX Database

The Operational Riskdata eXchange Association (ORX) is the world’s leading operational risk loss data consortium for the financial services industry. Its members use the ORX database for statistical modelling, benchmarking of loss performance, and validation of internal data collection. As a leading industry group, ORX additionally provides a forum for discussion among banks on operational risk issues and the development of industry standards. The association is further committed to advancing fundamental research into operational risk on the basis of its loss database.

As of the end of September, 2007, the ORX loss database contained 86,433 loss events collected from 36 financial institutions. Loss data is submitted by member institutions dating back as far as January 1, 2002, and is subjected to a quality control process that enforces the completeness, accuracy, and consistency of the data, as set down in the ORX Reporting Standards Guidelines (ORX, 2007). Each loss in the database is categorized according to primary and secondary business lines and event types; the primary categories, which shall be our focus in this study, are listed in Table 2. Note that this categorization differs slightly from the Basel II loss taxonomy, although it is straightforward to map the categories of the two classification schemes to each other. The database records the amounts of the gross loss and of any direct and indirect recovery, the dates of occurrence, discovery, and recognition, the country in which the loss was incurred, and an indicator of whether the loss is related to credit or market risk events. In addition to the loss records, each bank submits quarterly data on the gross income

Business Lines	Event Types
1. Corporate Finance	1. Internal Fraud
2. Trading and Sales	2. External Fraud
3. Retail Banking	3. Employment Practices and Workplace Safety
4. Commercial Banking	4. Clients, Products, and Business Practices
5. Clearing	5. Disasters and Public Safety
6. Agency Services	6. Technology and Infrastructure Failures
7. Asset Management	7. Execution, Delivery, and Process Management
8. Retail Brokerage	8. Malicious Damage
9. Private Banking	
10. Corporate Items	

Table 1: Primary business line and event type categories used in classifying ORX data.

of each business line.

### 3 Literature Review

Several authors have investigated the functional relationship between operational loss sizes and exposure indicators using regression models. Most commonly, these studies fit linear models of log-losses onto bank characteristics that are correlated with firm size, such as log-gross income or log-total assets, possibly also including binary variables indicating the region, business line, or type of loss incurred. These models may be interpreted as a multiplicative model for loss amounts, where a “common” or baseline loss value (corresponding to the exponential of the intercept in the fitted linear regression model) is multiplied by a scaling factor that is dependent on the “idiosyncratic” characteristics of the bank (Na et. al., 2006),

$$L_b = g(r_b^{\text{idio}}) \cdot h(R^{\text{com}}),$$

where  $L_b$  is a loss incurred by a bank or business line  $b$ , the first term is a deterministic function of some vector of bank- or business-line characteristics  $r_b^{\text{idio}}$ , and the second term is a random variable corresponding to the loss distribution of a “standardized” bank. For example, in many of the studies,  $r_b^{\text{idio}}$  is a measure of firm size (call it  $s_b$ ), and  $g$  is a power-law function, so that

$$L_b = s_b^\lambda \cdot h(R^{\text{com}}).$$

Given fitted values for the parameter  $\lambda$ , this model may be used to define a scaling relation among business lines  $B = 1, 2, \dots$  of various sizes using the equalities

$$\frac{L_1}{(s_1)^\lambda} = \frac{L_2}{(s_2)^\lambda} = \dots = h(R^{\text{com}}).$$

Our study employs similar models which determine a common “base distribution” for losses in a given category, which are scaled using multiplicative and exponential factors that depend on the levels of exposure indicators particular to one bank’s setting.

In an early, short study, Shih et. al. (2000) consider a regression model of loss sizes against firm size, measured in terms of revenue. Their model, which was fitted using the OpVar database (at the time of the study, this database was maintained by PricewaterhouseCoopers, and it has a lower reporting threshold of US\$1 million), may be written as:

$$\log(Y_i) = b_0 + b_1 \log A_i + e_i,$$

where  $Y_i$  is the loss size,  $A_i$  is the bank revenue,  $e_i$  is an error term, and  $b_0$  and  $b_1$  are parameters to fit. We note that this model corresponds to a power-law relationship between loss sizes and firm size, as measured by revenue. The OLS fit of the model indicated that a positive linear relationship between losses and firm size was present. The authors then note that the errors in the loss model appear to increase in variability with firm size, and correct for this heteroscedasticity by using a weighted least squares fitting procedure, which also returned a positive linear relationship. In this study, we explicitly model heteroscedasticity by characterizing how various quantiles of the loss distributions change with respect to the levels of the exposure indicators.

This methodology was also used in a study by Chappelle et. al. (2005) to combine internal and external data to compute an overall loss distribution for a given business line / event type category. They use the external data only to model losses above a high level, while using internal data only to model the distribution of smaller losses. They report similar results as those of Shih et. al. (2000).

Na et. al. (2006) perform a similar regression modeling exercise as was done by Shih et. al. (2000), but based on bank-internal and -external data supplied by ABN-AMRO. It is not clear from their article, however, what the source of the external data was. Gross income is the predictor used to model the size of business lines. A key difference in the experiments performed in this study is that not only loss severities, but also aggregate weekly losses and loss frequencies are modeled. In addition, the regressions are performed on mean losses by business line, as well as the standard deviation by business line. The authors found that the model did not fit the severity distributions well, as the values of  $R^2$  were quite low, and the estimated coefficients

varied widely according to whether internal, external, or combined internal/external data were used in the model fitting. However, both the aggregate loss data and the frequency data showed strong evidence of power-law relationships with gross income.

Dahen and Dionne (2007) extend the model of Shih et. al. (2000) to include not only total assets as a predictor of loss sizes, but also indicators for the region in which the loss occurred, the business line, and the event type. The study took loss data reported by financial institutions from 1994 and 2006 in Fitch’s OpVar database, which has a lower reporting threshold of US\$1 million. The regression model employed was

$$\log(Y_i) = b_0 + b_1 \log(A_i) + \sum_{j=1}^3 b_{1+j} R_{ij} + \sum_{j=1}^7 b_{4+j} B_{ij} + \sum_{j=1}^6 b_{11+j} E_{ij} + e_i$$

where  $Y_i$  is the loss amount,  $b_0$  is the common risk component,  $A_i$  is the bank’s total assets,  $R_{ij}$  is a region indicator (one indicator for each of the regions US, Canada, and Europe),  $B_{ij}$  and  $E_{ij}$  are business line and event type indicators, respectively, and  $e_i$  is an error term, assumed to be normally distributed.

This model was fit using ordinary least squares. The authors report an adjusted  $R^2$  value of 0.11, which, though still quite low, is larger than the value reported by Shih et. al. (2000). Analysis of variance showed that the firm size variable contributes very little to the overall  $R^2$  (0.006), although its coefficient value is statistically different from zero at a 90% significance level. The location, business line, and event type variables also each include statistically significant indicators, and each contribute nearly equally to the overall  $R^2$  value.

Dahen and Dionne next consider a reduced regression model, which only included the independent variables which appeared significant at the 90% level in the full model (log of total assets, the regional indicators for the US and Canada, and indicators for the business line Commercial Banking and event type Clients, Products, and Business Practices):

$$\log(Y_i) = b_0 + b_1 \log(A_i) + b_2 R_{i,US} + b_3 R_{i,Canada} + b_4 B_{i,CB} + b_5 E_{i,CPBP} + e_i.$$

All the terms in the above model except for  $b_0$  are considered part of the “idiosyncratic” term. Using the estimated coefficients  $\hat{b}_j$  from the reduced model, the authors then proposed a means of scaling a loss occurring in one bank B to another bank A based on the ratio of the corresponding estimated idiosyncratic terms for each bank:

$$\text{Loss}_A = \frac{g_A}{g_B} \text{Loss}_B,$$

where

$$g_A = \exp \left[ \hat{b}_1 \log(\text{Assets}_A) + \hat{b}_2 R_{A,\text{US}} + \hat{b}_3 R_{A,\text{Canada}} + \hat{b}_4 B_{A,\text{CB}} + \hat{b}_5 E_{A,\text{CPBP}} \right].$$

Here, the region, business line, and event type indicators can be set to scale a loss from Bank B differently if it is considered to occur in various regions within Bank A, as well as in different business lines, event types, etc. Dahlen and Dionne tested this scaling model for predicting the distribution of losses observed at Merrill Lynch and showed that both the predicted average loss and standard deviation were statistically equal at a 95% confidence level.

An alternative method by which to approach the scaling question is based on the assumption that the loss distributions arise from a given parametric family of distributions, and that the parameter values corresponding to a given bank's loss distribution are functionally related to the value of the bank's exposure indicators. For example, many parametric families of distributions include scale parameters that govern the size of the quantiles of the distributions; fitted scale parameter values can be regressed on firm size variables. Shape and location parameters can likewise also be fitted using the same technique. This method was applied in Wei (2007) to a dataset of 376 losses of over US\$10 million taken from the Fitch OpVar database. Wei models the severity distribution using several parametric classes of distributions, including the Generalized Beta of the second kind (GB2), Burr Type XII, Generalized Pareto (GPD), and Lognormal families of distributions, both without using covariate information, as well as modeling the scale parameters of these distributions as  $\alpha A_b^\beta$ , where  $A_b$  are the total assets of the bank  $b$ , and  $\alpha$  and  $\beta$  are parameters to fit. Wei uses a likelihood ratio test to determine if the model with covariate is significantly better than the model without covariate, and concludes that for the data he considers, the GB2 distribution with the covariate provides the best overall fit.

## Contrasting Approach of the Current Study

Our study does not make any parametric assumptions about the loss functions, and instead builds regression models that have similar interpretations as the previous studies cited above. However, our analysis explicitly addresses several limitations of these studies that make them unsuitable for scaling loss data across institutions. First, OLS methods estimate the response only at the mean loss level given the level of certain exposure indicators. The response in the mean levels may not match the response at various quantiles of the loss distributions, as very large losses for banks may scale differently for banks of different sizes than small losses. In addition, in the log-linear models considered above, the mean of the log-losses does not correspond to any intuitively recognizable statistic in the original losses, as the exponential of the mean of the log losses is not the mean of the losses. As we shall discuss below, *quantile*

*regression* techniques provide a means of estimating changes in losses as a function of exposure indicators at all levels of the loss distributions, which provide meaningful relations among raw loss data even when they are fit to log-losses.

Second, if loss severities are being modeled through a regression relationship, the  $R^2$  value is of limited importance in assessing the goodness of the fit, as we are not seeking to “explain away” the variation of losses observed at different levels of the exposure indicators, but rather, we simply want to find a value by which to shift loss values at one level to make them comparable in location and scale to loss values at another level. A good regression relationship should therefore be measured in terms of the degree of closeness or similarity among the scaled loss distributions, and not by how much of the variation in losses is accounted for by the regression line. There may still be considerable variability in the residuals of the regression model; however, as long as the distribution of these residuals follows the scaling model at all levels of the exposure indicators, the regression model is providing an acceptable scaling relation. We shall address this issue by applying statistical tests of the equality of loss distributions at various exposure indicators levels, which adjust for the location and/or scale differences in these distributions across the levels.

Third, the loss databases used in each study all present data that are subjected to a reporting threshold, meaning that only losses in the tails of the severity distributions are reported. Given the possible differences in location and/or scale of loss distributions at various exposure indicator levels, the reporting threshold corresponds to an unknown and different quantile level of the loss distribution of each loss category. This creates difficulties in properly aligning the distributions of different loss categories when estimating a simple scaling relation. We introduce a “quantile matching” algorithm that attempts to reduce the distortions introduced during the truncation of loss data by the reporting threshold requirement.

## 4 Scaling Analysis Methodology

### 4.1 Quantile Regression and Tests of Location- and Location-Scale Shift

Our study applied quantile regression (QR) techniques (Koenker, 2005) to characterize the relationship between loss sizes and the levels of exposure indicators. Our primary motivations for using quantile regression were:

- We could determine whether or not “large” losses varied differently from “small” losses with the level of the exposure indicator. For example, if the linear QR fits at the median and at the 0.95 quantile levels are not approximately parallel to each other, then we may conclude scaling relations used for comparing large losses should be different from the

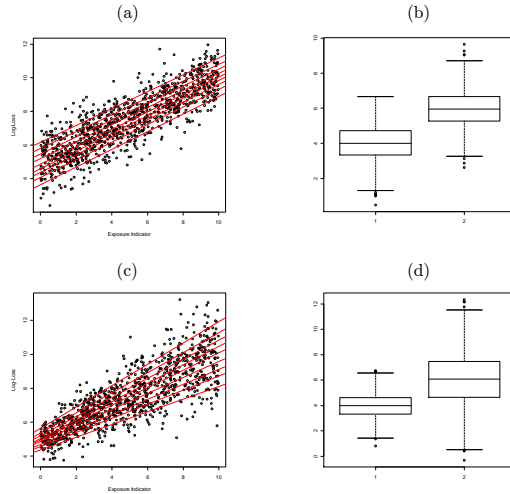


Figure 1: Examples of location- and location-scale shift models using simulated data. Figure (a) shows that the fitted quantile regression lines in a location-shift model are nearly parallel; the boxplots shown in (b) corresponding to loss distributions at two exposure indicator levels are simply shifted versions of each other. Figures (c) and (d) show the corresponding graphs for a location-scale model. In (c), the quantile regression lines fan out from each other such that the distances among them remain in a fixed proportion. The same relation holds between the horizontal lines in the boxplots in (d).

scaling used for small or “average” losses.

- Because quantiles of distributions are preserved under monotone transformations of the data (unlike means), QR fits provide meaningful results even when applied to log-transformed data. We generally chose to work with log-transformed loss data as the raw data often exhibited heavy-tailed behavior, and location or location-scale shifts of log-loss distributions can be interpreted simply as scale- or scale-power shifts of the original, raw data.
- We could test formally whether location- or location-scale shift hypotheses for the data were statistically acceptable. If so, then the differences among loss distributions can be characterized using just one or two parameters, and loss data of any size can be scaled appropriately through linear transformations. Koenker (2005) defines a formal statistical test of these hypotheses based on the theory of quantile regression, which he terms the “Khmaladze test,” that we have applied throughout the analysis.

The interpretation of location- and location-scale shift models may be characterized according to geometrical relationships among the quantile regression lines. When the location-shift hypothesis is true, we should expect to see that all quantile regression lines are parallel to each other. In the case of a location-scale shift, the quantile regression lines fan out in such a way that the

vertical distances between the regression lines at each exposure indicator level are in constant proportion to each other. See Figure 1. Formally, a location-scale shift model can be expressed as (Koenker, 2005)

$$Y_i = X_i\alpha + (X_i\gamma)\varepsilon_i, \quad (1)$$

where  $Y_i$  is a (log-)loss value,  $X_i$  is a (row) vector of independent variables,  $\alpha$  and  $\gamma$  are (column) vectors of regression parameters to be fitted, and  $\varepsilon_i$  is a random variable having a fixed distribution  $F_0$ . The terms  $X_i\alpha$  and  $X_i\gamma$  represent values that add to and multiply the random variable  $\varepsilon_i$ , respectively, so that the distribution of the loss  $Y_i$  for any given exposure indicator level  $X_i$  is just a shifted and scaled version of the base distribution  $F_0$ . The parameters  $\alpha$  and  $\gamma$  therefore represent location-shift and scale-shift parameters for this base distribution, respectively. We may alternatively express the regression relation (1) according to the quantiles of the distribution of the response variable, for a given  $X_i$ , as

$$Q(\tau | X_i) = X_i\beta(\tau),$$

where  $\tau \in (0, 1)$  represents the probability level of the quantile, and  $\beta(\tau) = \alpha + \gamma F_0^{-1}(\tau)$ .

A location-shift model is a special case of (1), where the scale-shift term  $X_i\gamma$  is identically set to 1, i.e.,

$$Y_i = X_i\alpha + \varepsilon_i. \quad (2)$$

In this case, changes in the values of the exposure indicators  $X$  contribute only to an additive increase or decrease in the location of the loss distribution. We note that typically an intercept term is included in our model, so that the first value of the vector  $X_i$  will be 1. The model (2) has exactly the same form as an ordinary linear regression model, except perhaps for the minor difference that the “error” terms  $\varepsilon_i$  are not necessarily assumed to have zero mean. A more substantial difference between the quantile regression approach that we take and traditional regression methods lies in the manner in which the models are fit. Rather than first focus on estimating the coefficients  $\alpha$  under the assumption that the  $\varepsilon_i$ ’s follow a common distribution, we instead first test whether the  $\varepsilon_i$ ’s follow a common distribution under the assumption that a linear model of the form specified describes the location shift. Only when we are satisfied that this hypothesis holds do we then estimate the values of the location-shift coefficients  $\alpha$ . In a sense, we are checking the goodness-of-fit of the model before we estimate the model parameters, rather than the other way around, as is usually done in regression analysis.

The Khmaladze test (Koenker, 2005) is a statistical test of the hypothesis that a given quantile regression model follows a location- or location-scale shift model. The test has the property that the distribution of its test statistic is asymptotically independent of the loss dis-

tributions being compared; like the Kolmogorov-Smirnov test of equality between distributions, the Khmaladze test may be performed by referring its test statistic to a single table of critical values, regardless of the distribution of the losses. This is of great advantage when the base loss distribution is not known in advance of the model fitting. In the analysis, we first applied the Khmaladze test for location-shift, and if the test passed, we estimated the location-shift parameters  $\alpha$ . Otherwise, we would then apply the Khmaladze test for location-scale shift, and if that test passed, we estimated  $\alpha$  and  $\gamma$ . If neither test passed, we simply reported the differences in the fitted regression models at different quantile levels. Location-shift parameters  $\alpha$  were estimated using ordinary least-squares fitting. Location-scale shift parameters  $\alpha$  and  $\gamma$  were estimated as follows: We first fitted a series of quantile regression lines at several probability levels  $\tau_i$ . Let  $\hat{\beta}(\tau_i)$  denote the vector of estimated coefficients at each level, and let  $\hat{F}_0^{-1}(\tau_i) = X_0 \hat{\beta}(\tau_i)$  denote the estimated quantile at level  $\tau_i$  of the loss distribution at a fixed exposure indicator level  $X_0$ . Using the relation  $\beta(\tau) = \alpha + \gamma F_0^{-1}(\tau)$  given above, we estimated the terms in  $\alpha$  and  $\gamma$  using an ordinary least-squares regression model of each component of  $\hat{\beta}(\tau_i)$  against  $\hat{F}_0^{-1}(\tau_i)$ . Each such regression yielded an estimate of the components of the shift parameter vectors  $\alpha$  and  $\gamma$ .

## 4.2 The Loss Reporting Threshold Problem and the Quantile Matching Algorithm

Scaling relationships properly concern the distributions of all losses experienced by banks, and not only those above the reporting threshold of €20,000. For example, if we hypothesize that the quantiles of one bank’s losses are a scale multiple of another bank’s losses, we mean that there exists a factor  $a$  such that for all  $x > 0$ ,

$$F_A(x) = F_B(ax),$$

where  $F_A$  and  $F_B$  are the complete loss distribution functions of the two banks. To understand what would happen if we instead assume that the scaling relation holds only for losses observed above €20,000 euro, note that for  $a > 1$  we would assume that the lowest quantile of the conditional distribution of bank B (conditional on the loss exceeding 20,000), would be  $20000 \cdot a$ ; however the true lowest quantile for bank B’s losses is 20,000. This mismatch will be present to some degree for all probability levels. See Figure 4.2 for an illustration of the possible distortions among quantiles that can arise due to the loss reporting threshold. We shall use the term “unconditional” loss distributions to refer to the distribution of all losses incurred by a bank, as well as the “conditional” loss distribution to refer to the distribution of losses exceeding the reporting threshold.

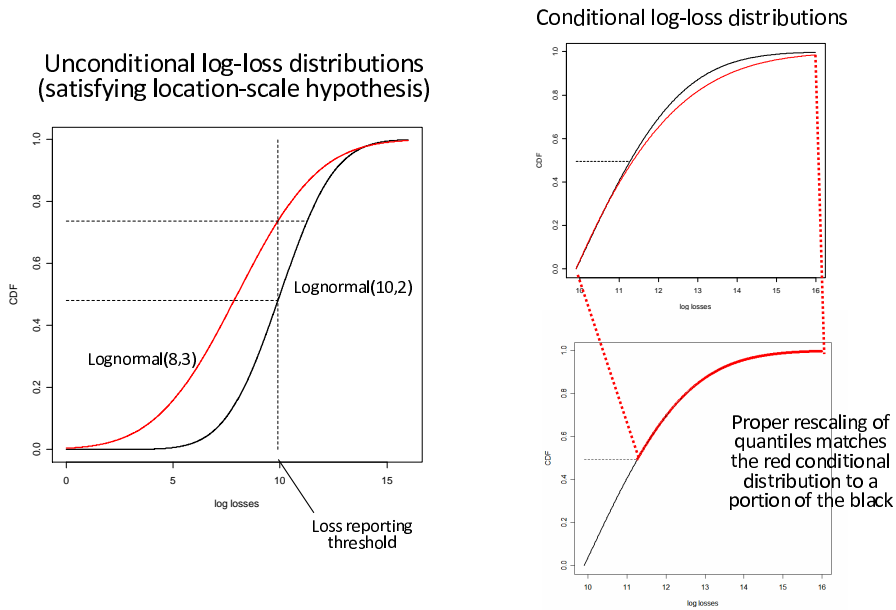


Figure 2: Illustration of the distorting effects of the loss reporting threshold. The figure at left displays two distributions of log-losses *not* subject to the reporting threshold that differ by a location-scale shift. Losses observed above the reporting threshold appear to follow the distributions in the upper right graph. These loss distributions appear quite similar and there may be no apparent need to apply any scaling relation; however, the proper scaling relation is shown at the bottom right, where the losses following the red distribution are only matched with the upper portion of the black distribution. Failure to recognize larger differences in the underlying loss distributions due to the masking effects of the loss reporting threshold can therefore lead to gross inaccuracies in the estimates of the true scaling relations.

In order to match two loss distributions that differ in scale (but are otherwise equal) so that the quantiles of their unconditional loss distributions match each other, we need to be able to identify the percentile levels of each corresponding to a given loss amount. For example, if we knew that the reporting threshold of €20,000 corresponded to the 30th percentile of one unconditional loss distribution and the 40th percentile of another distribution, then we could easily determine a scaling that would match losses above the 40th percentile of the first distribution with the losses of the second. Those losses that occur between the 30th and 40th percentile of the first distribution would remain unmatched, as we do not observe any points in this range from the second distribution, due to data truncation. The problem of course is that we have no means of knowing what the percentile level of the reporting threshold is for the unconditional loss distributions, as we do not know how many losses were truncated below the €20,000 level.

One approach to this problem would be to assume that the losses arise from a common parametric family of distributions, derive a likelihood function using a conditional version of the density function for this family, and then use maximum likelihood methods to estimate the parameters. The fitted distributions can then be used to estimate the percentile level of the loss reporting threshold. While this is a very straightforward approach, it comes at the cost of making strong assumptions on the distributions of the unreported losses under the €20,000 threshold. For example, even if the tail behavior of loss events above the €20,000 threshold is closely matched by the tail of a lognormal distribution, this does not imply that the distribution of losses under threshold matches this distribution. By making this assumption, we are effectively guessing as to the percentile level of the threshold of each distribution. In addition, this method assumes that a parametric family of distributions will generally fit well for all loss category distributions for which scaling relations are desired.

Rather than commit to strong assumptions regarding the distribution of unobserved losses below the reporting threshold, we introduced a “quantile matching” technique for matching like quantiles among several loss distributions. Although this method does assume that a location- or location-scale shift model is valid for the entire unconditional loss distribution – even below the reporting threshold – it avoids making any assumption about the percentile level of the reporting threshold and the overall shape of the unconditional loss distributions. To explain how this method works, we first introduce some notation. Let  $F$  and  $G$  denote two unconditional loss distributions for which we hypothesize that there exist some factors  $a$  and  $b$  such that  $F(y) = G(a+by)$  for all  $y$  (i.e.,  $G$  is a location-scale shift of  $F$ ). We further define  $\bar{F}(y) = 1 - F(y)$  and  $\bar{G}(y) = 1 - G(y)$  to denote the complement (survival) distributions, respectively. Next, let  $t$  denote the loss reporting threshold of €20,000, and let  $\tilde{F}(y) = \frac{\bar{F}(y)}{\bar{F}(t)}$  denote the (complement) conditional loss distribution of losses that exceed the threshold limit  $t$ , and similarly define  $\tilde{G}(y)$ .

Assuming that  $a + bt > t$ , we can derive the following relationship between  $\tilde{F}$  and  $\tilde{G}$ :

$$\tilde{F}(y) = \frac{\overline{F}(y)}{\overline{F}(t)} = \frac{\overline{G}(a + by)}{\overline{G}(a + bt)} = \frac{\tilde{G}(a + by)}{\tilde{G}(a + bt)}. \quad (3)$$

Comparing the first and last terms of this expression, we see that  $\tilde{F}$  and  $\tilde{G}$  nearly follow a location-scale model, except that the percentiles of  $\tilde{F}$  must be multiplied by the constant  $\tilde{G}(a + bt)$  in order to be brought in line with those of  $\tilde{G}$ . Equation (3) implies the following relationship between the quantiles  $\tilde{F}^{-1}(\tau)$  and  $\tilde{G}^{-1}(\tau)$  of these two distributions:

$$\tilde{G}^{-1}(\tau\tilde{G}(a + bt)) = a + b\tilde{F}^{-1}(\tau) \quad (4)$$

for all  $\tau \in (0, 1)$ . If instead of dealing with two distinct distributions, we are working in a general quantile regression setting where the location-scale model (1) is considered, we may rewrite (4) as

$$\tilde{F}_X^{-1}(\tau\tilde{F}_X(X\alpha + (X\gamma)t)) = X\alpha + (X\gamma)\tilde{F}_0^{-1}(\tau) \quad (5)$$

where  $F_X$  denotes the loss distribution at exposure indicator level  $X$ , and  $F_0$  denotes a “base” distribution at some given exposure indicator level  $X_0$ . We may further write the left-hand side of (5) simply as  $\tilde{F}_{X,\alpha,\gamma}^{-1}(\tau)$ , where  $\tilde{F}_{X,\alpha,\gamma}$  represents the complement conditional distribution of losses at the exposure indicator level  $X$  exceeding the level  $X\alpha + (X\gamma)t$ . The identity results from the observation that, for  $y > X\alpha + (X\gamma)t$ ,

$$\tilde{F}_{X,\alpha,\gamma}(y) = \frac{\tilde{F}_X(y)}{\tilde{F}_X(X\alpha + (X\gamma)t)},$$

which implies that  $\tilde{F}_{X,\alpha,\gamma}^{-1}(\tau) = \tilde{F}_X^{-1}(\tau\tilde{F}_X(X\alpha + (X\gamma)t))$ . Thus,

$$\tilde{F}_{X,\alpha,\gamma}^{-1}(\tau) = X\alpha + (X\gamma)\tilde{F}_0^{-1}(\tau),$$

and this relation has the form of a location-scale model as in (1), where  $\tilde{F}_{X,\alpha,\gamma}$  represents the loss distribution at the exposure indicator level  $X$ , conditional on the loss exceeding  $X\alpha + (X\gamma)t$ . These observations suggest the following iterative algorithm to compute appropriate values of  $\alpha$  and  $\gamma$  in a location-scale shift model:

### Quantile Matching Algorithm

1. Initialize a data set  $S = \{(X_i, y_i) : i = 1, \dots, n\}$  to include all the available data points.
2. Fit a quantile regression model using the points in  $S$  to estimate location and scale pa-

parameters  $\alpha$  and  $\gamma$ .

3. Define the set  $S'$  according to

$$S' = \{(X_i, y_i) : y_i \geq X_i\alpha + (X_i\gamma)z\},$$

where

$$z = \max_i \left\{ \frac{t - X_i\alpha}{X_i\gamma} \right\}. \quad (6)$$

4. If  $S = S'$ , then stop and return the current values of  $\alpha$  and  $\gamma$ . Otherwise, set  $S \leftarrow S'$  and return to step 2.

The mechanics of the algorithm may be understood as follows. In step 2, the parameters  $\alpha$  and  $\gamma$  are estimated, which can be used to determine a threshold value  $X\alpha + (X\gamma)t$  corresponding to any exposure indicator level  $X$ . The distribution of data values  $y_i$  exceeding the levels  $X_i\alpha + (X_i\gamma)t$  follows the distribution  $\tilde{F}_{X_i, \alpha, \gamma}$ ; we therefore remove all data points such that  $y_i < X_i\alpha + (X_i\gamma)t$  from the data set  $S$  in Step 3 and estimate the parameters  $\alpha$  and  $\gamma$  again using this new data set. Therefore, each loop in the iteration first estimates  $\alpha$  and  $\gamma$ , then sets a new loss data reporting threshold based on these estimates, truncates the data according to the new threshold, and finally re-estimates  $\alpha$  and  $\gamma$ . The algorithm stops looping when the values of  $\alpha$  and  $\gamma$  no longer change, or if the algorithm begins to cycle through a set of values for these parameters. The algorithm can be adapted to estimate location-shift models by estimating only the parameters  $\alpha$  in step 2 and ignoring the terms  $X_i\gamma$  in both equations of step 3. In this case we may replace Step 3 with the simpler, equivalent version:

$$S' = \left\{ (X_i, y_i) : y_i \geq t + X_i\alpha - \min_i \{X_i\alpha\} \right\}.$$

We have yet to explain the motivation for the choice of the parameter  $z$  in (6), however, which is written in place of  $t$  in Step 3. When estimating a location- or location-scale shift model as in (1) or (2), the modeler has the freedom to choose the location and scale of the base distribution  $F_0$ , which will in turn affect the values of the parameters  $\alpha$  and  $\gamma$ . For example, in a location-shift model, we could instead replace the base distribution  $F_0$  with the distribution  $F'_0$ , where  $F'_0(y) = F_0(a + y)$  for some  $a$  and for all  $y$ . To compensate for this change, the value of  $\alpha$  in this model would have to change to  $\alpha'$ , where  $\alpha' = \alpha + a \cdot e_1$ ,  $e_1$  being a vector whose first component equals one but whose other components all equal zero. Due to the presence of these extra “degrees of freedom” in our estimation procedure, we are faced with a range of values of  $\alpha$  and  $\gamma$  from which to choose in step 2 of the algorithm. We want to choose these parameters such that the resulting threshold values are as unrestrictive (i.e., small) as possible, under the

constraint that their value never drops below  $t$  at any  $X_i$ . This can be shown to be equivalent to finding the minimal value of  $z$  such that  $X_i\alpha + (X_i\gamma)z \geq t$  for all  $i$ , whose solution is given by (6).

The quantile matching algorithm is useful for fitting shift models, but recall that we are interested in first determining whether shift models are appropriate. We can do this by performing the Khmaladze test for the shift model on the data values in  $S$  returned by the quantile matching algorithm. However, since quantile matching selects the data points in  $S$  on the assumption that a shift model is valid, testing the validity of a shift model after this process removes the data points will likely exhibit bias. To minimize this bias, we fitted  $\alpha$  and  $\gamma$  in step 2 of the algorithm using only very low quantiles of the distribution. For example, if we are fitting a location-shift model, we estimate the location shift parameters  $\alpha$  at the 0.1 probability level. In this case we effectively assume that the location-shift model only holds among the lowest 10% of the observed data points; the upper 90% need not satisfy this hypothesis in order for the quantile matching procedure to be effective.

The algorithm typically terminates after only a few steps, converging absolutely or cycling over a small set of similar solutions. When a shift model was not appropriate for the data, the algorithm often deleted too few or too many data points from the sample. Although the Khmaladze test usually fails in such cases, we have found that graphically displaying the remaining data points in the set  $S$  using boxplots or scatterplots is always advisable to control for instances where excessive amounts of data are deleted, or where the shift model assumptions are obviously violated.

## 5 Analytic Procedure and Selected Results

In the analysis undertaken in this study, we only used those loss data that were not listed as related to credit or market risk events, and where the net loss amount after direct recovery exceeded the reporting threshold of €20,000. The total number of loss events fulfilling these requirements was 78,282. Each loss was further associated with a geographic region, as well as with income statistics for both the relevant business line and the bank as a whole.

We performed regression analysis for each primary business line and primary event type, with the exception of EL8: Malicious Damage, for a total of 17 analyses (10 business lines, 7 event types). In each case, we used log-losses as the response variable, and we considered linear models of log-losses based on the following variables:

- **Quarterly Business Line Gross Income** (QTR.BL.GI) was considered in the analyses of business lines, with the exception of BL10, Corporate Items, which relates to losses that can only be properly attributed to the corporate level and not to any particular business

line. This statistic was associated with each loss recognized by the reporting bank in that business line during that quarter.

- **Quarterly Total Gross Income** (QTR.TOT.GI) was computed as the sum of all business line gross incomes reported by the bank in each quarter, and associated with all losses reported by that bank in that quarter. The distribution of these income values across institutions is given (in billions of Euro) in the following table:

min	25%	50%	mean	75%	max
0.002	0.922	2.00	2.83	3.38	15.3

- **Average Business Line Gross Income** (AVG.BL.GI) represents the average over all quarters of income reported by the bank in that business line, and is associated with all losses reported by that bank in that business line.
- **Average Total Gross Income** (AVG.TOT.GI) was computed as the average over all quarters of the sum of all business line gross incomes, and associated with all losses reported by that bank. These income values are distributed as:

min	25%	50%	mean	75%	max
0.046	0.805	1.85	2.51	3.10	10.8

- **Region** The region reported by the bank in which the loss was incurred. Regions included Western Europe, North America, Latin America, Asia/Pacific, Eastern Europe, and Africa. Western Europe and North America were by far the most represented regions, accounting for 31,200 and 40,086 loss records, respectively. In the study, we focus exclusively on the differences between these two regions.

An initial graphical exploration of the data helped to determine the functional form of the regression model and to eliminate any data with outlying levels of the exposure indicator. As most income statistics were highly correlated, we restricted the regression models to incorporate at most one income measure. The selection of a single income measure does not therefore imply that the other income measures were insignificant predictors, only that they did not produce as good a fit as the selected income variables. Once a suitable regression model was found for each individual exposure indicator, the models were combined in a multiple quantile regression analysis to arrive at a predictive model for loss distributions given the value of the exposure indicators in each category. A stepwise selection procedure was used to eliminate any variables that did not appear to be significant in the multiple regression model.

We next provide complete results from fitting models to losses in two business lines: Trading and Sales and Retail Banking. These two loss categories were chosen to illustrate the range

of regression relationships that we obtained: different shift models, with different relationships between loss severities and region and firm size measures. As mentioned above, the goodness of fit of these models is determined via the Khmaladze test, since acceptance of the null hypothesis is an indication that the shift model is appropriate for the data. Because the models are all fit for log-losses, each additive term in the fitted models corresponds to a multiplicative term for the corresponding model for the raw loss data, as the addition of some value  $x$  to a given log-loss amount corresponds to a multiplication of the raw loss value by  $10^x$ . In the reported model statements, all income variables should be understood to be in units of billions of euro, and that we indicate the regional variable as  $1\{\text{NA}\}$ , as it is a binary indicator variable whose value is one if the loss is from North America, and zero if from Western Europe. In addition, the estimated standard errors of each regression coefficient is listed in parentheses at the right of each regression model. Note however that these are standard errors associated only with the least-squares fittings of  $\alpha$  and  $\gamma$ ; they do not take into account the variability that is introduced through the application of the quantile matching procedure.

We emphasize that all the results presented in this paper are based on data available up to September, 2007; the results are subject to change in the future as more data become available to the consortium.

## BL 2: Trading and Sales

Losses in Trading and Sales were observed to follow a location-shift model with average total gross income as a variable. Because income was correlated to region (North American banks generally reported higher incomes than Western European banks), the regional indicator was found to be insignificant when added to this model ( $p$ -value 0.16), so we did not include it in the final model. We further discuss the role of regional differences below. The regression model was fit using 6,931 data values in this category, 3,428 of which remained after quantile matching.

$$\hat{Q}(\tau|X) = \hat{F}_0^{-1}(\tau) + 0.0946 \cdot \text{AVG.TOT.GI} \quad (0.0030)$$

Standard errors of the coefficients are listed in parentheses at right. In this model, we see that larger firms experience significantly larger losses in this category, with nearly 25% higher losses for every €1 billion in average total revenues.

The quantiles of the estimated base distribution  $\hat{F}_0$  were:

Min	25%	50%	75%	Max
4.278	4.447	4.695	5.071	7.516

Average business line gross income also proved to be a highly significant predictor, producing a location model fit whose Khamaladze test statistic was generally higher than in the case of average total bank gross income. Average gross income within the Trading and Sales business line was highly correlated in our sample ( $r = 0.79$ ) with average total bank gross income. Average total bank gross income was also assumed to be less variable than the corresponding business line income, especially when considering quarterly business line gross income. In addition, we found that both the form of the regression model and the size of the regression coefficients remained stable when considering only banks where Trading and Sales made up a substantial portion of the revenue, such that average gross income for Trading and Sales was at least 20% of the average total bank gross income. (No bank's Trading and Sales income exceeded 45% of total income.) For these reasons, we opted to use average total bank gross income in the model as a measure of firm size.

As mentioned above, regional differences did not appear to provide strong explanatory power over and above the income differences as shown in the model; this was in part due to the fact that the correlation between the regional indicator and average total gross income was somewhat high at 0.32. However, a separate fit of log-losses onto the regional indicator showed that a location-scale model was appropriate, such that North American losses were generally greater than European losses, with the difference being slightly larger at the high end of the loss distribution. To compare, the 25th percentile of the North American log-loss distribution (after quantile matching was applied) exceeded that of Western Europe by 0.097; at the 75th percentile, however, the difference in log-losses was 0.127. These differences are statistically significant and do indicate the presence of regional variation, although the measurement of this variation is partly confounded by income differences among the banks.

We finally note that the Trading and Sales business line encompasses a variety of activities within a bank, which may be associated with different loss distributions. In the ORX databases, losses are distinguished in this business line category if they related to equities, global markets, corporate investments, or treasury/funding. Most of the reported losses were concentrated in the first two categories (2,908 equities losses and 4,304 global markets losses). We found that losses were typically higher in the global markets area than in equities, with the differences being slightly greater at the high end of the distribution. At the 25th percentile (after quantile matching), the difference in log-loss values was 0.035; at the 75th percentile this differences increases to 0.074. We intend to continue to explore variability among secondary business lines further in future work.

### BL 3: Retail Banking

Retail Banking losses were observed to follow a location-scale shift model with quarterly business line gross income and regional factors as variables. We restricted the data to those banks whose quarterly retail banking income was in the range €1–3B. The regression model was fit using 18,965 data values, 18,136 of which remained after quantile matching.

$$\begin{aligned}\hat{Q}(\tau|X) &= -0.869 + 1.216\hat{F}_0^{-1}(\tau) && (0.018, 0.004) \\ &+ (0.487 - 0.120\hat{F}_0^{-1}(\tau))\text{QTR.BL.GI} && (0.009, 0.002) \\ &+ (0.0519 - 0.0161\hat{F}_0^{-1}(\tau))1\{\text{NA}\} && (0.0115, 0.0024)\end{aligned}$$

In order to understand the range of variation in the slopes of the various quantile regression lines fit under this location-scale shift model, one must consider the range of the quantiles of the estimated base distribution  $\hat{F}_0$ . These were estimated to be:

Min	25%	50%	75%	Max
4.314	4.422	4.585	4.850	7.305

Based on these quantile levels, the following table lists the regression coefficients fitted at selected quantiles of the loss distribution:

	Min	25%	50%	75%	Max
Intercept	4.378	4.509	4.707	5.030	8.015
QTR.BL.GI	-0.033	-0.046	-0.066	-0.098	-0.394
1{NA}	-0.017	-0.019	-0.022	-0.026	-0.065

It is interesting from this graph that loss severities decrease with income at all quantile levels; also, losses are less severe in North America than in Western Europe. From the above table, we that there is considerable variation in the effects of firm size and region, depending on the what percentile of the loss distribution one observes. For example, at the 25th percentile level, each increase of €1 billion corresponds to an 11% decrease in the corresponding quantile of the loss distribution, while at the 75th percentile, the same increase in income leads to a 25% decrease. The decrease in these quantile levels as one moves from Western Europe to North America ranges between about 4% and 6%, respectively.

## 6 Summary of Results

We summarize the overall results of the model fits for each exposure indicator in Table 2, which indicate if loss distributions are increasing, decreasing, or do not change with the level of firm

Loss Category	Loss severity increasing (Incr) or decreasing (Decr) with firm size?	Larger losses in Western Europe (WE) or North America (NA)?	Shift Model Type: Location-shift (L) or Location-scale shift (LS)?
Corp Fin	Incr	NA	L
Trad & Sales	Incr	NA	L
Ret Banking	Decr	WE	LS
Comm Banking	Incr	WE	LS
Clearing	Incr	WE	LS
Agency Serv	Incr	NA	L
Asset Mgmt	Decr	NA	L
Ret Brokerage	Incr	NA	L
Priv Banking	Decr	NA	L
Corp Items	Incr	WE	L
Int Fraud	Decr	WE	L
Ext Fraud	Decr	WE*	LS
Empl Practices	Incr	WE	L
Clients Prods	Incr	NA*	LS
Disasters	Incr	-	LS
Tech & Infr	-	NA	LS
Exec Delivery	Incr	NA	L

Table 2: Summary of results, indicating the direction of the signs of the regression coefficients. A hyphen indicates that no significant relationship was found. A star indicates that the indicated region showed higher losses only at the upper end of the loss distribution; among the lower values of the distribution, the other region had comparatively higher losses. This effect can occur with location-scale shift models. Finally, we note that the increasing trend of losses with firm size for Execution, Delivery and Process Management was only observed among banks with average quarterly total gross incomes of less than €3 billion; above this income level, no significant trends were observed.

size indicators, and whether the location, location-scale, or neither hypothesis was accepted.

We conclude from this table that simple scaling models are appropriate for characterizing the relationship between loss distributions and exposure indicators. While in some cases the loss distributions appear to scale equally with the exposure indicator level at every probability level, the number of location-shift models selected indicates that frequently large losses scale differently from small losses. Because of the regulatory focus on large losses, it is essential that scaling relations for extreme events be appropriately characterized. In addition, we see that larger firms and business lines usually, but not always, incur larger losses. Exactly why we may observe decreasing relationships between loss sizes and firm sizes in such categories as Retail Banking, Retail Brokerage, Asset Management, and Internal and External Fraud is the subject of continued investigation.

Our study highlights the utility of using quantile regression techniques to characterize differences in loss distributions for banks of various sizes and operating in different geographies. We arrive at a more complete picture of scaling relations than previous studies were able to show using only least-squares regression techniques. The methods allow us to compare and scale losses from different institutions in a more sensitive manner, and in particular characterize more precisely the response of very high quantiles of the loss distributions.

## References

- Basel Committee on Banking Supervision. “International Convergence of Capital Measurement and Capital Standards.” 2006.
- Chapelle, Ariane, Yves Crama, Georges Hübner, and Jean-Philippe Peters. “Measuring and managing operational risk in the financial sector: an integrated framework.” (2005). Available at SSRN: <http://ssrn.com/abstract=675186>.
- Dahen, Hela, and Georges Dionne. “Scaling models for the severity and frequency of external operational loss data.” Canada Research Chair in Risk Management Working Paper 07-01, 2007.
- Koenker, Roger. *Quantile Regression*. Cambridge University Press, 2005.
- Na, Heru Sataputera, Jan van den Berg, Lourenco Couto Miranda, and Marc Leipoldt. “An econometric model to scale operational losses.” *J. Oper. Risk* 1 (2006): pp. 11–31.
- Operational Riskdata eXchange Association. “ORX Reporting Standards: An ORX Member’s Guide to Operational Risk Event/Loss Reporting.” Available at <http://www.orx.org/>.
- Shih, J., A. Samed-Khan, and P. Medapa. “Is the size of operational loss related to firm size?” *Operational Risk*, 2000.
- Wei, Ran. “Quantification of operational losses using firm-specific information and external database.” *J. Oper. Risk* 1 (2007): pp. 3–34.