HEC Montréal

Scaling model for the severity of external operational loss data with endogenous Markov regime-switching

Par

John Ioannis Simitzis

Sciences de la gestion

(Économie Financière Appliquée)

Mémoire présenté en vue de l'obtention du grade

de maîtrise ès sciences (M.Sc.)

Septembre 2012

Abstract

Large operational losses are generally infrequent but potentially disastrous events that must be accounted for in the financial risk management framework of bank holding companies. Internal databases lack the observations necessary to implement any meaningful models for calculating necessary capital. We propose a scaling model that accounts for the heterogeneity of this risk all the while accounting for potential selection bias in the external data and possible changes in operational risk dynamics that may be attributed to the recent financial crisis. We find evidence of high and low regime periods in the data which, when properly accounted for using a regime-based method, ameliorate the effectiveness of the scaling mechanism we develop. We find proof that the scaled external losses stem from the same distribution as that of the internal bank holding company, and can thus be used in acquiring more internal data to elaborate a model for operational risk.

Keywords:

Operational risk, bank holding companies, scaling model, selection model, external data, Markov endogenous regime-switching.

Sommaire

Les grandes pertes opérationnelles sont généralement des évènements rarissimes, mais qui peuvent s'avérer désastreux d'où l'importance de bien les considérer dans l'élaboration des systèmes de gestion des risques financiers d'institutions bancaires. Les bases internes de ces institutions n'ont pas le nombre d'observations nécessaires pour développer un modèle fiable afin de calculer le capital à mettre de côté. On propose un modèle de normalisation qui prend en considération l'hétérogénéité de ce risque tout en traitant le biais de sélection potentiel lié aux pertes externes et la possibilité d'un changement dans la dynamique du processus de risque opérationnel qui pourrait être attribuable ou non à la crise financière récente. Nous constatons la présence de périodes de régime haut ou bas dans les données qui, lorsque bien incorporés dans un modèle à régimes, aident à améliorer l'efficacité de notre mécanisme de normalisation. On vérifie que les pertes externes normalisées proviennent de la même distribution que les pertes internes d'une institution bancaire et pourront être utilisées dans l'élaboration d'un modèle pour traiter le risque opérationnel.

Mots-clés:

Risque opérationnel, institutions bancaires, modèle de normalisation, modèle de sélection, données externes, changements endogènes de régimes Markoviens.

Table of Contents

List of tables

List of figures

List of abbreviations

LDA:	Loss Distribution Approach
NBER:	National Bureau of Economic Research
M$:	Millions of dollars
B$:	Billions of dollars
USD:	United States dollars
CDF:	Cumulative Distribution Function
EDF:	Empirical Distribution Function
RBr:	Retail Brokerage (Business Line)
TS:	Trading and Sales (Business Line)
PS:	Payment and settlement (Business Line)
CF:	Corporate finance (Business Line)
AS:	Agency services (Business Line)
CB:	Commercial banking (Business Line)
RB:	Retail banking (Business Line)
AM:	Asset management (Business Line)
DPA:	Damage to physical assets (Risk Type)
CPBF:	Clients, products, and business practises (Risk Type)
EPWS:	Employment practises and workplace safety (Risk Type)
EF:	External fraud (Risk Type)
IF:	Internal Fraud (Risk Type)
EDPM:	Execution, delivery, and process management (Risk Type)
BDSF:	Business disruption and system failures (Risk Type)
Reg:	Regime (Interaction variable)
OLS:	Ordinary Least Squares regression
GLS:	Generalized Least Squares regression
MLE:	Maximum Likelihood Estimator(s)
BHC:	Bank Holding Company

Acknowledgements

I wish to thank my director, M. Georges Dionne, for his constant availability and support over the course of my work and for having shared with me his extensive expertise and knowledge in terms of financial risk management and econometrics. I was inspired by his teachings in my risk management class, which resulted in my interest to challenge myself further and continue learning. I will never forget the "interview" I was submitted to in order to work on this project. I was asked if I knew how to apply an econometric model in particular, which I did. It ended up being the only model I knew that was useful in accomplishing this, forcing me to learn incessantly as I progressed. This was exactly what I had hoped for.

I would also like to thank my family and friends who supported me throughout the whole process, but would like to highlight the individual contributions of my father, who has been pushing me from before I can remember to accomplish what he knew I was capable of.

I would like to thank my fiancé Jasmine who supported me on a daily basis and was always there for me, and surely always will.

Last but not least, I thank the jurors who have accepted to evaluate my work.

1. Introduction

Regardless the sector in which it evolves, three particular types of risk emerge at the core of any financial institution: market, credit and operational. These institutions face the challenge of developing internal models to properly and objectively quantify, supervise and manage these risks.

Due to extensive data and research, most financial institutions have satisfactory market and credit risk management departments. Over the past few years, the world has witnessed a plethora of mergers & acquisitions in the banking sector which have given form to massive entities of growing intricacy and risk that must be managed. Moreover, the extensive automation of services and e-commerce in general put additional strain on their integrated systems. For such reasons and due to recent events that have shone light on potential gargantuan mishaps, operational risk has grown exponentially in importance. This type of risk affects all departments, which we will here forth call "business lines" that the bank may have, but in possibly very different ways; that we will qualify as "risk types". Rogue trading such as witnessed at Société Générale in 2008 which cost the French bank a whopping 4.9B€ is very different in nature to, say the failed transaction processing event at Wells Fargo Bank that cost it 150M$. One common denominator though, is the rarity and relative severity of these individual loss events. Their unequivocal heterogeneity is a testament to the inherent complexity of a potential model to predict them.

In June 1999, regulatory authorities became involved with a mission to develop a framework incorporating operational risk. The ultimate goal such authorities aim to achieve through regulation is the realignment of capital requirements with the actual exposure to risk, which encompasses an integrated approach. This stipulates the inclusion of correlation effects between the bank's different risks (as aforementioned). Still far from that reality, operational risk models are still in their

infancy, let alone at the stage of understanding how it relates to other risks. This master's thesis will seek to continue on this path by proposing and developing an efficient model to predict the level of operational losses, conditional on there being a loss, allowing banks to estimate their value at risk (VaR) related to their exposure to operational risk and set appropriate capital aside for it. It is important to note that dedicating capital to cope with potential losses does not supersede the need to adapt their supervision and management techniques to account for operational risk.

In 2001, the Basel Committee officially defined operational risk as the risk of loss resulting from inadequate or failed internal processes, people, and systems or from external events. It was inefficacious to collect data before a general definition such as this one was agreed upon and finally established within the banking community. An extensive database is required when trying to implement models, a privilege not yet conceivable. Internal data collection could still take a number of years more in order to observe enough high-impact losses that are of interest because they are unlikely occurrences. Resorting to the use of external data is therefore more than a simple viable option; it is an essential supplement that will allow for a more comprehensive outlook on tail events. One cannot simply combine the two and use the resulting distribution to model the loss patterns. According to the Basel Committee on Banking and Supervision, the correct combination of internal and external loss data is thus an important step to be considered.

Most operational risk research falls into the loss distribution approach (LDA) category. Ultimately, what is needed is a final aggregate loss distribution from which the capital requirement can be calculated; defined by the Basel Committee as being a one-year holding period with a confidence level of 99.9%. LDAs require extensive data and stipulate that the desired loss distribution that will be used to calculate the value-at-risk stems from a mathematical convolution between the two types of statistical distributions: the severity of the losses –defined as the amplitude of a loss conditional on there being one- and their frequency. VaR estimation precision will be

increased by modeling for the best parameters in order to exploit the multivariate nature of the distribution that is chosen for each of the two components. Conceptually, it is important to understand the role such variables play in differentiating the shape of the right-hand-side tail between two hypothetical bank entities having the same total operational loss amount; one displaying more high-frequency low-severity losses as opposed to the other that more low-frequency but high-severity losses. For the same amount of aggregate loss, these two hypothetical banks share a very different exposure to operational risk. The sporadic nature and form of the data, showing a wide range of losses make it preferable to estimate each distribution separately to allow for more parameters and greater accuracy (Cummins et al., 1990; Frees and Valdez, 2008).

On that note, the model we are developing comes from a relatively unexplored wing of operational risk research: scaling models. A scaling mechanism is essentially a normalization formula that can be used to combine an external database to an internal one by projecting an external loss from one institution onto another (the internal one). In order to integrate the external event into a more complete internal database, the level of control variables explaining the severity (or frequency) of operational losses determine how important the loss would be if the same loss observed in the external database were to happen in the institution we are studying.

Building on a methodology previously elaborated by Dahen and Dionne (2010) to make better use of external data in order to predict the severity of potential operational losses, our ultimate goal is to incorporate new data related to the recent financial crisis. We construct and test a scaling mechanism for the severity of operational losses using yearly and quarterly data on bank holding companies from the United States and an external database of operational losses from public sources spanning years 1994 until 2010. We cope with potential selection bias of our loss database due to its 1M$ minimum loss threshold by using a 2-step Heckman selection model. The first step is a model for the probability that a bank holding

company from our database suffers a minimum loss of 1M$. Conditional on there being a loss (severity), the second step consists of running a panel OLS regression that will provide us with the necessary coefficients that explain level of operational losses and will be used for our scaling mechanism. In order to account for the crisis and its effects, we will also look into a regime switching model to capture a possible change in the dynamics of operational risk during the crisis. A great advantage of this method is that these regimes need not be based on macroeconomic factors, but can be endogenous; defined by movement or changes in the dynamics of operational risk itself.

We find proof that the potential selection bias related to our use of external data is not really cause for concern. Furthermore, regimes aid in explaining the operational loss process by capturing a change in dynamics post 2003, allowing us to use more scaling factors than were possible in Dahen and Dionne (2010) thus making the scaling mechanism more efficient.

Provided the relatively young literature that exists on the topic of operational risk and the lack of internal data, to our knowledge this research is one of the first to use data of operational loss events that include the latest financial crisis, as well as to incorporate regime-switching.

The thesis is organized as follows. The next section gives a brief overview of the literature surrounding the fundamental components that constitute the model: scaling mechanisms and regimes. Section 3 exposes the theory behind our model while section 4 gives an overall description of the data. We will elaborate on the methodology used to estimate the model in section 5. Section 6 exposes and analyzes our empirical findings, followed by a brief conclusion in section 7.

2. Literature review

A popular branch of the literature falls into the extreme value theory (EVT) category; authors try to incorporate externally observed operational losses by assuming they all stem from a same distribution that represents the upper tail of a common loss distribution. Papers such as Chavez-Demoulin et al. (2005) attempt to fit severity using distributions such as the Generalized Pareto or Lognormal families. The difficulty resides in identifying the unique threshold for each loss observed in a different risk control environment to be able to parameterize such distributions. They are working on stochastic methodology to find the threshold after which EVT asymptotics can be fitted. More on the basic theory behind EVT will be presented in the literature review.

As mentioned previously, empirical and even theoretical literature surrounding operational risk is scarce; this statement holds doubly when we consider scaling models. This section will cover two papers on scaling models that deal with operational risk. The first is a short key article generally perceived to have influenced more sophisticated statistical models such as the second one, which elaborates a method with important implications relating to our own. As this is the first time endogenous regime-switching are incorporated in scaling models for operational risk, the literature review is extended to a research paper that deals with credit spreads.

2.1 Scaling models

2.1.1 Shih et al. (2000)

To understand the basis of our scaling model, it is important to mention a research paper that, although simplistic by nature, paved the way for further work in the field. This is the case of Shih et al. (2000) who decided to test the somewhat intuitive

relationship expected to be observed between the size of a firm and the magnitude of its operational losses. Using proxies such as revenue, assets and the number of employees, and losses from the PriceWaterhouseCoopers OpVaR database, they found that the logarithm of revenues showed the best relationship of all combinations, thus concluding that the relationship is non-linear. Conceptually, this translates into a bank having twice the amount of revenues (the proxy for size) as another does not, on average, suffer twice the amount of losses.

Having chosen the most relevant scale factor (log-revenues), they perform an OLS regression to explain log-losses. The results show R^2 and adjusted R^2 just over 5%, which is frankly quite low. Noticing that their residuals plotted a funnel shape, which indicates a linear relationship between the variability of losses and their variable, they decided to run a GLS regression to deal with heteroskedasticity. They run what is called a weighted least square regression by dividing both sides of the equation by the log-revenues. This yielded strong t-statistics for the intercept and variable, but low R^2 and adjusted R^2 of just under 10%. The positive coefficient of log-revenues (0.7) tells the story of a diminishing relationship between losses and size. The only methodological problem relates to the nature of the data used. They do not account for potential selection bias due to the fact the minimum loss amount reported is $1 million in their database.

The authors point out that other variables must surely be needed to explain the remaining 90% of loss variability, such as business line or the quality of the control environment. Yet the most important insight of their report was not so much the results as it was the last few lines where they suggest using this correlation to "scale-adjust" external data to the size of the firm being analyzed. This intuition served as inspiration for more applied statistical models in the operational risk literature. Our work will include the natural logarithm of a size variable as suggested by this paper in our scaling mechanism.

The basis for a scaling model lies in the existence of what is called a power-law. The notion finds its roots in the field of theoretical physics, where it is hypothesized that physical laws governing the smallest increments of length scales can be used to derive the laws of larger ones through the existence of a mathematical relationship that binds them together using a scaling exponent. This self-similarity, where a smaller increment or function manifests itself proportionately is caused by a scale-invariance property that can be observed empirically in many diverse fields.

What interests us is that this also relates to mathematics, where scale-invariance may possibly be observed between probability distributions. A necessary condition for linking seemingly unrelated dynamics such as these is their mutual proximity to some critical point. They are said to share the same critical exponents (such as the λ in the aforementioned example), which allows them to display equivalent scaling behaviour. An example of a common critical point can be the ebullition threshold shared by H_2O and CO_2 which may be used to explain the physical dynamics they share around this transition phase. Based on their proximity to some theoretical threshold separating a distribution with its upper-tail, researchers in extreme value theory explore the property of self-similarity to better model large and rare events that may be bound together by a power-law relationship, despite them being observed in different institutions as in our case.

2.1.2 Shih (2001)

Shih (2001) was the first article to suggest a theoretical structure for a scaling model that could incorporate external losses through a critical exponent based on a proxy for size. Although the author does not provide empirical evidence of this bold claim, he provides two very fruitful insights on how to go about applying it. He mentions that suitable scaling variables include a proxy for the size of the company or the business line, but also the risk control environment. Secondly, he assumes that

operational risk losses can be sectioned into two primary components: a common one that is identical to every bank or business line and an idiosyncratic one. This allows for an equal component between all banks or business lines that can be isolated, thus allowing losses to be compared between different institutions invoking the self-similarity concept discussed earlier that we will examine in our work.

2.1.3 Na et al. (2006)

The article by Na et al. (2006), whose methodology influenced ours, takes these ideas one step further by testing them empirically. They wish to prove that through the use of this hypothetical common component, a power-law relationship can be established between losses and a variable, which in this case is the size of the business line. To do so, they lay out the theoretical foundation of a scaling model. They use internal data that include bank losses from different business units and external data from ABN-AMRO for operational losses in different business line for the year 2003. They decided to separate the data per week, in order to analyze the frequency of operational losses, and to be able to calculate an aggregate loss. Given the aberrant lack of data, instead of comparing loss distributions, they turned to a mathematical approach by testing out how several assumptions hold up once losses are scaled.

The first assumption is that an operational loss from a given business line scales using an idiosyncratic component, in this case gross income(s_b), and is affected by its two main components in the form of this power-law:

$$L_b = (s_b)^{\lambda} \cdot \Psi \qquad \text{(a)}$$

where L_b is the operational loss per business line b=1, 2…, B, s_b is gross income of the business line from which the operational loss originates, λ is the scaling exponent they hope remains constant for different business lines, and a constant Ψ, common to all business lines.

If this power-law represents the real relationship between operational losses, then internal losses should derive from the same distribution as scaled losses. Na et

al.(2006) test the validity of their proposed scaling mechanism by running a linear regression on the means and standard deviations of each business line and the aggregate standard one mentioned before. The two first moments should scale the same way (same λ) between losses per business line and scaled losses from other business lines. This comes from the logical theoretical construct derived from equation (a):

$$\frac{L_1}{(s_1)^\lambda} = \frac{L_2}{(s_2)^\lambda} = \ldots = \Psi \qquad \text{(b)}$$

This demonstrates the relationship between operational losses in different business lines that find their equivalency through a proportion based on their business line size(s_b). Using the logarithm of equation (a), Na et al.(2006) run a linear regression on the means (and standard deviations) of losses and scaled losses between business lines and conclude that no power-law relationship could be concluded for the severity of operational losses since their scaling exponent (λ) is not statistically significant.

The focus of our research being the severity distribution, those results may seem discouraging, but many factors must be considered. The main one is the lack of data and what it could mean. Another consideration is their scaling variable itself. Unlike what was advocated in Shih (2001), they do not go as far as to consider the effects of different control environments particular to each business line, something that we will test in our work. Despite all this, Equation (b) is fundamental if externally available data is to be merged with internal data using a proportion that can also be based on other variables to control for, and it is the fundamental basis of our study.

2.2 Regimes

Since endogenous Markov regime-switching models have not been touched upon in operational risk literature, we turn to an innovative article that deals with explaining credit spreads in bonds. Please note that we will not explicit any of their theoretical point of views on variable selection or other specifics of credit spreads as it is not particularly relevant to our work in operational risk. Regime-switching models allow for greater flexibility in terms of the way explanatory variables interact with the dependant variable we wish to analyze.

2.2.1 Maalaoui Chun, Dionne, and François (2010)

What the authors in Maalaoui Chun, Dionne and François (2010) observed was that some research papers found different signs for coefficients that most experts would agree upon should be fundamentally obvious and could not quite grasp the cause. Worst of all was that the poor empirical results could not really be attributed to a lack of data or the quality of proxies used. This phenomenon has been dubbed the credit spread puzzle. Their paper attempts to explain such aberrations while improving the explanatory power of the overall model by exploring the presence of low and high regimes based on the mean and variance of the credit spreads.

On one hand, they elaborate on what they call a single-regime model. This is simply a multivariate regression that includes a dummy variable for when the data is in a high regime from a Markov switching model that will be explained in full shortly. The models take this general form:

$$Y_{it} = \beta_{it}^{0} + X'_{it}\beta_{it}^{1} + \beta_{it}^{2} regime_{it} + \epsilon_{it}{}^{1}. \qquad (c)$$

The single-regime model assumes the effects of the explanatory variables found in the X'_{it} vector on the dependant variable Y_{it} remain constant over the entire period

[1] The annotations have been modified for simplicity and comparability purposes from those originally found in the article (they explicit and compare three mixes of variables).

analyzed. In other words, the coefficients do not change from one regime to the other.

The regime-based model, as mentioned previously, allows for a greater flexibility of the explanatory variables' effects on the dependent variable by adding interaction components between the two regimes. The general model takes this form:

$$Y_{it} = \gamma_{it}^{0} + X_{it}'\gamma_{it}^{1} + \gamma_{it}^{2} regime_{it} + X_{it}'\gamma_{it}^{3} regime_{it} + \upsilon_{it}. \quad \text{(d)}$$

The result of the estimation (Y_{it}) is dependent on the regime at time t:

$$\begin{cases} low - regime: \widehat{\gamma_{it}^{0}} + X_{it}'\widehat{\gamma_{it}^{1}} \\ high - regime: Y_{it} = \left(\widehat{\gamma_{it}^{0}} + \widehat{\gamma_{it}^{2}}\right) + X_{it}'(\widehat{\gamma_{it}^{1}} + \widehat{\gamma_{it}^{3}}) \end{cases} \quad \text{(e)}$$

The interaction coefficient $\widehat{\gamma_{it}^{3}}$ allows us to verify if the same variables that explained the dependent variable change in any way when the risk is in a high regime. The intuition is that perhaps the dynamics differ when the process in a higher mean, higher volatility environment. Certain coefficients from the single-regime model (d) may change signs, have their explanatory power increase or decrease, as well as lose or gain statistical significance. This method allowed the authors to double the explanatory power of most models in credit spread literature, which is why we will incorporate the technique to see if it will help us cope with the crisis periods in our database.

2.2.2 Engel and Hamilton (1990)

The theory behind the functioning of the Markov endogenous regime-switching model is exposed in Engel and Hamilton (1990). The regime is the result of a latent variable s_t that may take on the value 1 if it is found in a low regime or 2 if it is in a high one. The dynamics of our dependant variable y_t are hypothesized to change from one regime to the next follow a normal distribution:

$$y_t \sim N\left(\mu_{s_t}, \sigma_{s_t}\right), \text{ where } s_t = 1, 2 \quad \text{(f)}$$

Furthermore, it is assumed the latent variable follows a two-order Markov chain for the states:

$$p(s_t = 1|s_{t-1} = 1) = p_{11}$$

$$p(s_t = 2|s_{t-1} = 1) = 1 - p_{11}$$

$$p(s_t = 1|s_{t-1} = 2) = 1 - p_{22}$$

$$p(s_t = 2|s_{t-1} = 2) = p_{22}$$

The process for s_t depends on past realizations of y and s uniquely through s_{t-1} as shown by the Markov chain just presented. It is important to note that only two probabilities are needed to calculate the entire chain.

We can summarize the probability law for y_t through 6 population parameters: θ= ($\mu_1, \sigma_1, \mu_2, \sigma_2, p_{11}, p_{22}$) which are sufficient to describe the different information we need such as the distribution of y_t given s_t, the distribution of s_t given s_{t-1}, as well as the unconditional distribution of the state of the first observation:

$$\hat{\rho} = \frac{(1-\widehat{p_{22}})}{(1-\widehat{p_{11}})+(1-\widehat{p_{22}})}, \quad \text{(g)}$$

and p(s_1=2; θ) = 1-ρ. All that is left is to estimate these parameters using maximum likelihood estimation. The joint probability distribution of our observed data given our sample (y_1, …, y_T) loss observations over T time intervals, and (s_1, …, s_T) matching states that could be high or low in each period looks like:

$$p(y_1, \dots, y_T, s_1, \dots, s_T;\ \theta) =$$

$$p(y_T|s_T;\ \theta) \cdot p(s_T|s_{T-1};\ \theta\text{the}\ ) \cdot p(y_{T-1}|s_{T-1};\ \theta) \cdot p(s_{T-1}|s_{T-2};\ \theta) \cdot \dots \cdot$$

$$p(y_1|s_1;\ \theta) \cdot p(s_1;\ \theta)\,. \quad \text{(h)}$$

The last term in equation (h) is the unconditional distribution of unobserved state of our first observation (g). The sample likelihood function is the summation of (h) over all possible values of the states vector:

$$p(y_1, \dots, y_T;\ \theta) = \sum_{s_1}^{2} \dots \sum_{s_T}^{2} p(y_1, \dots, y_T, s_1, \dots, s_T;\ \theta). \quad \text{(i)}$$

This would imply 2^T summations for each possible value of the θ parameters, which remain unknown to us. The authors decide to use the entire sample of ex post information in order to make an inference on the state that the process was in at some date t rather than just use past information and refer to it as being the smoothed inference:

$$p(s_t|y_1, \dots, y_T;\ \theta) \tag{j}$$

Equation (j) is the main difference between a mixture of normal distributions model, which supposed each draw to be independent, and the model put forth by Engel and Hamilton (1990).

An advantage of their model is the flexibility it provides by not imposing any restrictions on the parameters that will result from the maximum likelihood estimation. Those parameters are simply the first-order conditions for the maximization of summation (i) with respect to θ. The MLE $\hat{\theta}$ satisfy[2]:

$$\hat{\mu}_j = \frac{\sum_{t=1}^{T} y_t \cdot p(s_t = j | y_1, \dots, y_T; \hat{\theta})}{\sum_{t=1}^{T} p(s_t = j | y_1, \dots, y_T; \hat{\theta})}\ , \text{j=1, 2} \tag{k}$$

$$\widehat{\sigma^2}_j = \frac{\sum_{t=1}^{T} (y_t - \hat{\mu}_j)^2 \cdot p(s_t = j | y_1, \dots, y_T; \hat{\theta})}{\sum_{t=1}^{T} p(s_t = j | y_1, \dots, y_T; \hat{\theta})}\ , \text{j=1, 2} \tag{l}$$

$$\widehat{p_{11}} = \frac{\sum_{t=2}^{T} p(s_t = 1, s_{t-1} = 1 | y_1, \dots, y_T; \hat{\theta})}{\sum_{t=2}^{T} p(s_{t-1} = 1 | y_1, \dots, y_T; \hat{\theta}) + \hat{\rho} - p(s_1 = 1 | y_1, \dots, y_T; \hat{\theta})} \tag{m}$$

$$\widehat{p_{22}} = \frac{\sum_{t=2}^{T} p(s_t = 2, s_{t-1} = 2 | y_1, \dots, y_T; \hat{\theta})}{\sum_{t=2}^{T} p(s_{t-1} = 2 | y_1, \dots, y_T; \hat{\theta}) - \hat{\rho} + p(s_1 = 1 | y_1, \dots, y_T; \hat{\theta})} \tag{n}$$

This allows the data to dictate how the process moves. If (m) or (n) are large, regime switches will be less likely and frequent. The two first moments of the hypothesized distributions are pretty straightforward. If we could simply observe the state in which the process was in at each period, the probabilities in (k) and (l) would simply be 1 or 0, and we would simply have the average and standard deviations of each distribution

[2] As demonstrated thoroughly in Hamilton (1990) Appendix A.

separately. Instead, we have the smoothed probabilities that each observation is in one distribution based on our entire sample, which will be used to weight to construct an estimate of (k) and (l). Analyzing the Markov transition probabilities (m) and (n) in the same fashion, if the states could be observed, the probabilities being once again 1 or 0, we are simply left with, for a particular state, the number of times the process stayed in that same state as a fraction of the total number of times the process one period had been in that state one period before (s_{t-1}). Formulae (m) and (n) account for the smoothed probabilities with the addition of a slight modification for the initial condition (g) in the denominator.

Engel and Hamilton (1990) warn that singularities in the likelihood function arise when estimating the parameters for i.i.d. mixtures of normal distributions. An example they cite is when the mean of the first regime is assigned the value of the first realization in the sample ($\mu_1 = y_1$) and the variance of the regime is permitted to vanish (since we do not make any constraining assumptions). In this scenario, the likelihood function (i) diverges toward infinity[3]. For this reason, the authors utilize an adaptation of the EM algorithm[4] elaborated by Hamilton (1990)[5] which includes Bayesian priors to solve for the θ parameters. Their method calks the MLE as a special case where the diffuse priors $\nu=\alpha=\beta=0$. Formulae (k) and (l) are thus replaced by:

$$\hat{\mu}_j = \frac{\sum_{t=1}^{T} y_t \cdot p\left(s_t = j \middle| y_1,\ldots,y_T; \hat{\theta}\right)}{\nu + \sum_{t=1}^{T} p\left(s_t = j \middle| y_1,\ldots,y_T; \hat{\theta}\right)} \text{ , j=1, 2} \qquad \text{(o)}$$

$$\widehat{\sigma^2}_j = \frac{\beta + \left(\frac{1}{2}\right) \cdot \sum_{t=1}^{T} \left(y_t - \hat{\mu}_j\right)^2 \cdot p\left(s_t = j \middle| y_1,\ldots,y_T; \hat{\theta}\right) + \left(\frac{1}{2}\right) \cdot \nu \cdot \left(\hat{\mu}_j\right)^2}{\alpha + \left(\frac{1}{2}\right) \cdot \sum_{t=1}^{T} p\left(s_t = j \middle| y_1,\ldots,y_T; \hat{\theta}\right)} \text{ , j=1, 2} \qquad \text{(p)}$$

[3] Everitt and Hand (1981)
[4] Dempster, Laird, and Rubin (1977)
[5] Two-state, first-order Markov process with no autoregressive dynamics example p.52

The point is to indirectly guide the MLE estimates towards concluding that there is no difference between the regimes. This is done by replacing our original likelihood formula (i) with this one that includes the priors:

$$\log p(y_1, \dots, y_T;\ \theta)\text{-}(\ \nu\cdot \mu_1^2)/(2\sigma_1^2)\text{-}\ \alpha\log\ \sigma_1^2\text{-}\alpha\ \log\sigma_2^2\text{-}\beta/\sigma_1^2\text{--}\beta/\sigma_2^2 \qquad \text{(q)}$$

The EM algorithm begins by setting initial values for the population parameters θ. Using these parameters and the full sample of observations $(y_1, \dots, y_T)$, calculate the smoothed probabilities using an iterative processing of the data[6]. These probabilities will then be used to calculate (o) for both distributions, which in turn will be used to calculate (p), (m), (n) and by default (g). With our new vector of parameters θ, we can recalculate new smoothed probabilities utilizing the same algorithm as the first step. Each new iteration and recalculation of the weights will increase the value of the likelihood formula until a pre-determined convergence target specified in the code is reached, where we deem the marginal value added is negligible and the operation has reached the maxima.

We will test this method of using the conditional expectations for the unobserved scores in the EM algorithm with Bayesian priors as done first in Engle and Hamilton (1990) but applied in the context of a two-regime model as seen in Dionne, François and Maalaoui Chun (2010) to see whether operational risk contains high and low regimes and if the dynamics of it change whether we are in one or the other of those regimes.

[6]Algorithm for calculating the smoothed probabilities is exposed in Appendix B of Hamilton (1990).

3. Model

The model is above all a scaling mechanism allowing for the integration of external operational loss data that stem from other bank holding companies with heterogeneous control environments. The normalization formula used to scale the losses is identical to the one found in Dahen and Dionne (2010) with the exception that we include the effect of the specific low or high regime at each period. We basically explore the existence of a power-law relationship between the magnitude of operational losses and the size of the institution but include the bank's specific control environment as suggested by Shih (2001).

Similar to Equation (a) in Na et al. (2006), it is assumed that the loss amount can be broken down into two components: one that is common to all bank holding companies (in our case) as well as an idiosyncratic one. This implies that each operational loss from a particular bank at a given time can be decomposed in the following manner:

$$Loss_{it} = \psi \cdot Size_{it}^{\alpha} \cdot f(\omega, \vartheta_{it}), \qquad (1)$$

where ψ is the common component that remains constant for all time periods "t" and banks entities "i". This can represent the aggregate effect of macroeconomic, geopolitical, and other broad factors such as these which affect the population of bank holding companies as suggested in Dahen and Dionne (2010). Its other component is variable and specific to the institution where the loss was observed. It is comprised of a variable for the institution's size at the time of the loss and a function of ϑ variables specific to the loss event as well as to the bank's specific control environment with their relative scaling coefficients represented by the vector ω. The scaling exponent "α" associated with a size proxy for the bank institution will serve as the indirect factor that links together a bank's internal loss distribution with the one that can be fashioned from the external database with our normalization formula. We take the log-transformation of this equation to linearize it:

$$Log\ (Loss_{it}) = \text{Log}\ (\psi) + \ \alpha Log(Size_{it}) + \ Log(f(\omega, \vartheta_{it})). \qquad (2)$$

From the idea that all banks share a common component in their operational losses comes a similar equality as the one observed in Equation (b), but in this case it is between losses from different bank institutions instead of business lines. This leads to the normalization formula that will be applied to external operational losses in order to find their internal equivalent if they were to happen to the bank we are analyzing. By isolating ψ in Equation (1), we find an equality between the hypothetical loss that would be found in the internal database called $Loss_i$ and the one observed in the external database, called $Loss_e$:

$$\frac{Loss_i}{Size_i^{\alpha} \cdot f(\omega,\vartheta_i)} = \frac{Loss_e}{Size_e^{\alpha} \cdot f(\omega,\vartheta_e)}.$$

Theoretically, this equality between our internal rescaled losses and any loss observed in the external database holds through the existence of a ratio, as described by the self-similarity property of power-laws. This ratio is found by simply isolating $Loss_i$ in the last equation:

$$Loss_i = Loss_e \cdot \frac{Size_i^{\alpha} \cdot f(\omega,\vartheta_i)}{Size_e^{\alpha} \cdot f(\omega,\vartheta_e)}. \quad (3)$$

The coefficients necessary to implement the normalization formula (α,ω) are calculated using the methodology explained in section 5.

4. Data

4.1 Description of datasets

This research analyzes data from Algo OpData provided by Algorithmics Inc. which collects public source information on operational loss events stemming from various categories of companies around the world. We utilize only events that fall into the "Business and Finance" category between the years 1994 and 2010, thus adding 7 years of information to the empirical work done by Dahen and Dionne (2010). The inclusion of loss data from the financial crisis of mid-2007 is in itself an innovation over most of the current literature. Algo OpData more specifically reports losses in excess of 1M$ and provides additional information on each event such as the specific business lines implicated and the subsequent risk type (secondary and tertiary ones are disregarded in our model) following the standard categories put forth by the Basel Committee, a brief description of each event, the parent company as well as an array of information on the institution suffering the loss, such as the location, year, amount of assets, employees, revenues, etc. Most of that additional information about the company tends to be unreliable and will be disregarded in our analysis.

The use of external data exposes this thesis to a multiple of potential biases. The first is that our sample of losses has a truncation point that might render it not representative of the population of operational losses. This potential selection bias must be accounted for. It is possible that some banks may not even be predisposed to have a loss of such amplitude, all the while having an operational loss nonetheless. To account for this, additional information on these institutions is necessary in order to create a sample selection model as described in Heckman (1979) and which will be further discussed in section 5. With that in mind, we restrict our analysis on the population of Bank Holding Companies (BHCs) from the United States with a minimum asset-base of $1 billion taken from the Federal Reserve of Chicago database. This gives us access to a plethora of variables on each institution. The full description of the variables retained for each econometric model is also included in section 6.

4.2 Unique challenges

Combining the two databases came with a set of considerable challenges. The Algo dataset does not include a uniform matching variable such as the unique CUSIP code of each institution. The only way to correctly attribute the operational losses from the Algo OpData dataset to the BHCs that suffered them and whose idiosyncratic information is found in the Federal Bank of Chicago database is through the firm name, and since the Algo name variables did not seem to be standardized, there was no straightforward method to go about matching them. It required a more sophisticated tool. For the period between 1994 and 2010, the Federal Bank of Chicago database includes 1,137 different BHCs having over $1 Billion in total assets spanning 7,038 annual observations while the Algo OpData database has 4,557 operational loss events of over $1 Million in the Business & Finance industry; an automated solution was needed. Compustat has developed a name-matching software based on an algorithm that finds equivalencies between abbreviations and other differences in spelling. One of its main advantages is an option that allows the user to adjust the degree of leniency of the exactitude of the match. The results were not convincing at any degree in our case.

We developed a method that substitutes out problematic punctuations and abbreviations that caused these aberrations. Examples include points, commas, parentheses and all their contents to name a few. This was performed on both datasets in order to reduce the expressions to a bare minimum. A fail-safe method was then performed to make sure the losses were not attributed to another BHC due to the fact the reduction technique had created the same expression for different institutions. When a match was found, it tests to see if an identical expression that originated from a different starting firm name exists. This entire process was performed iteratively with different variations of removed expressions in order to find the optimal combination which resulted in the most matched operational losses to BHCs. The method retained removed 30 expressions in all. The matches were

reviewed manually because of a small margin for error and yielded 623 operational losses found.

Another complication derived from the reporting methodology used by Algo OpData for operational loss events. When a merger or acquisition is recognized, the target company and all of its losses from its inception are retroactively introduced into the acquiring institution. Left alone, the use of our final database would overestimate the amount of losses suffered by parent companies while entirely underestimating the losses of target companies. Such rigour is especially important due to the panel nature of our data. Furthermore, target companies are generally smaller in size that their acquirers. These smaller companies are also the ones that would most probably be affected by a selection bias due to the $1 Million loss threshold, since they are the most likely not to be able to sustain a loss of such magnitude. The other risk is therefore that the procedure implemented by Algo OpData would overestimate selection bias risk if less operational losses are reported in smaller companies. An example of this problem is Merrill Lynch, an institution with a great deal of operational risk exposure being retroactively included under the parent name of Bank of America Corporation, with numerous loss events included falsely before their September 2008 merger date.

The risk can be narrowed down to matched losses that may actually belong to the company that was the target of a merger or acquisition at a later date, thus falsely leaving the BHC lossless when in fact a loss did occur, and adding a loss to the future parent company when there is none. To cope with this, we used the SDC Platinum database by Thomson Reuters which reports information on mergers and acquisitions as well as many other events globally such as syndicated loans, new issues, private equity, etc. By refining our search to mergers and acquisitions affecting bank holding companies in the United States for a period which spans until mid-2011 because that is the effective date our Algo OpData was assembled, which means a merger or acquisition even after our studied period can affect the data. Luckily, the SDC

Platinum database includes variables for CUSIP of both targets and acquirers. We then extracted the CUSIP of the BHCs that had matched losses (available on the Federal Reserve Bank of Chicago database) and cross-referenced them with the CUSIP of acquirers.

4.3 Setup

Two final databases will be considered in this study, an annual one ranging from 1994 to 2010, and a quarterly one between 2001 and 2010. The choice of period for the quarterly database is based on the introduction of the Basel regulation. A reporting bias is evident in the data before that period; mainly a clustering of events on January 1st and December 31st of every year before 2001. This does not affect the annual regression, but would render a quarterly analysis quite flawed. The added observations are needed to increase the number of periods for the Markov regime-switching algorithm, going from 16 annual periods to at least 40 acceptable quarterly ones. A key variable was reported cumulatively at each quarter and was annualized for regression purposes. We deal with real variables that were adjusted for inflation using the CPI-U less food and energy index[7].

4.4 Assumptions

We ensure that the losses reported in the database are accurate and are not based on rumours or estimations. On the other hand, a debatably strong hypothesis is that it contains the entire population of losses in excess of $1 million. Some would argue that smaller operational losses may be easier to mask than large ones and may have eluded public exposure altogether thus resulting in an under-reporting bias. This also leads to saying that all types of losses are equally probable to be found in the data; that some categories of risk that may be more damageable to a company's reputation are as likely to be found as another. It is also important to note that although this study does not cover a full approach for scaling as in Dahen and Dionne (2010), any

[7] From the Bureau of Labor Statistics website : http://www.bls.gov/cpi/

further work to develop the frequency model would have to make the assumption that there is no correlation between the loss amount and the probability of it being reported such that the two distributions (frequency and severity) are independent.

5. Methodology

The goal is to populate the values of our scaling exponents in Equation (3). We elaborate a two-step econometric model as proposed in Heckman (1979) to test for and deal with possible selection bias. The source of bias comes from the minimum loss reported of $1M in the Algo OpData database. Since our losses are not randomly selected from the population, we cannot simply base a model predicting operational losses from it without further consideration; it may create a rightward shift from the real mean of the data. Seen another way, our dependant variable $Loss_{it}$ is censored therefore we may have an overconcentration of zeros which comes from the fact that some losses under $1M that may exist in the population of operational losses are instead not reported.

5.1 Heckman first-stage

The first step is to formulate a model for the probability that a bank holding company suffers a loss (which in our case is over 1M$) and is calculated using the entire sample of data. The second step then calculates the magnitude of the loss, conditional on there being one, calculated using only a subset of sample data. The selection equation takes this form:

$$\begin{cases} z_{it} = 1 & if\ z_{it}^{*} > 0 \\ z_{it} = 0 & if\ z_{it}^{*} \leq 0 \end{cases}, \text{ where } z_{it}^{*} = w_{it}'\alpha + \epsilon_{it},$$

and the loss amount equation is estimated using GLS:

$$\begin{cases} y_{it} = y_{it}^{*} & if\ z_{it} = 1 \\ y_{it}\ not\ observed & if\ z_{it} = 0 \end{cases}, \text{ where } y_{it}^{*} = x_{it}'\beta + \upsilon_{it}.$$

Z is the realization of a latent continuous variable z^{*} that follows a dynamic measuring its exposure to operational risk in a way that may help explain the presence of banks in the loss population ($z_{it}^{*} > 0$) The probit model used is estimated using STATA software. The logit model did not yield visibly different results (regression results available upon request). The Heckman model is a Tobit type II model, but we cannot use a generalized Tobit type II in our case since we deal with

loss amounts conditional on there being a loss (severity), and the tobit model does not exclude them.

When a loss is observed, y_{it} is the observed realization of another latent variable y_{it}^{*}. The importance is that both error terms $(\epsilon_{it}, \upsilon_{it})$ be normally distributed, independent with a zero mean. The errors are assumed to have a correlation ρ . It is also assumed that they are part of a bivariate normal density. From known results in theory of continuous multivariate distributions[8], the inverse Mill's ratio, which is a monotone decreasing function of the probability that an observation is selected in the sample, can be calculated using the latent dynamics from the probit on all the data. Including the inverse Mill's ratio into the OLS regression helps produce non-biased estimates. If ρ=0, then OLS would have provided unbiased results to begin with. It is equivalent to saying that the inverse Mill's ratio included in the OLS is not statistically different from zero.

The variables we will include in the first-step probit are the same ones created from the BHC data in the Chicago Federal Reserve database found in Dahen and Dionne (2010). The selection model includes three explanatory variables to help identify factors that may explain the probability of an institution having a loss of over 1M$.

- Mean salary is calculated as the sum of salaries and employee benefits divided by the number of full-time employees. This is meant to grasp the level of sophistication of the institution. More so, it can be thought that higher average salaries may be related to higher quality employees, thus we expect it to have a negative impact on the probability of loss. The better quality of management argument may somewhat be counterbalanced by risk incentives underlying higher salaries.
- Bank capitalization is calculated as the capital divided by total assets. This variable is a measure of the bank's moral hazard. Banks that are more

[8] See Johnson and Kotz (1972)

capitalized are probably covering the extra risk they are exposed to. We can therefore expect them to have more extreme losses. It should therefore have a positive effect on the probability as the ratio gets bigger.

- We use annual growth of gross domestic product[9] as a proxy for the macroeconomic environment. Chernobai et al. (2010) argue that more losses are observed in an economic downturn. One of the reasons cited are the fact it becomes harder to mask a loss as cash becomes scarcer in the company.

Adding a size proxy would be quite a logical step in explaining the probability of a fixed-level minimum loss, but we opt not to do so because of correlation with the Inverse Mills ratio that is to be calculated from the residuals of this regression.

A new variable explored will be a dichotomous one we created for when the operational risk process is deemed to be in a high regime (higher mean and variance) by the Markov endogenous regime-switching model described in section 2. It is intuitive to think that if operational risk dynamics were in a higher mean and variance state there may be a larger probability of incurring a conditional loss of over 1M$.

The Markov endogenous regime-switching algorithm as described in section 2.2.2 is applied using the annual (1994-2010) and quarterly (2001-2010) datasets on the dynamics of our variable to be explained, the severity of operational risk losses per period (y_t). These are determined as the mean value of the natural logarithm of all operational losses that occurred during each period since the process does not operate on the identity of each institution that incurs a loss but rather the time period in its entirety. We iterate the smoothed probabilities (j) in order to find the first-order conditions (m), (n), (o) and (p) that satisfy the maximum likelihood estimation of our generalized objective function (q) that include the Bayesian priors in order to avoid the singularity mentioned by Engle and Hamilton (1990). This is done using GAUSS

[9] Found on the U.S. Bureau of Economic Analysis website: http://www.bea.gov/

software. When the MLE $\hat{\theta}$ parameters are found, we analyze the final smoothed probabilities (j) as provided by the algorithm and create a dichotomous variable for each period when that probability is equal to or greater than 0.5 ($p(s_1|y_1, \dots, y_T;\ \hat{\theta}) \geq 0.5$), where state 1 ($s_1$) is the high endogenous regime state.

5.2 Heckman second-step

Using the residuals from the Heckman (1979) first step probit regression, we create the Inverse Mills ratio that will be introduced in our second step regression. We create this variable manually using STATA rather than use an all-in-one Heckman process. The ratio is the fraction from the following property of the truncated normal distribution:

$$E(x|x > \alpha) = \mu + \sigma \frac{\phi\left(\frac{\alpha-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\alpha-\mu}{\sigma}\right)},$$

where x is a random variable, α is our constant truncation point (0), μ and $\sigma^2$are the two first moments of the normal distribution of our residuals.

For the second step of the Heckman (1979), we wish to determine the coefficients of formula (2) as was done in Dahen and Dionne (2010) while incorporating the regime-based model format adopted by Dionne, François and Maalaoui Chun (2010) to see whether it helps explain the dynamics of operational risk while in a high or low regime. We can rewrite formula (2) as:

$$Log\ (Loss_{it}) = \text{Log}\ (\psi) + \ \alpha Log(Size_{it}) + \ \sum_j \beta_j BL_{ijt} + \sum_k \gamma_k RT_{ikt}\ , \qquad (4)$$

where vectors α, β_j, γ_k are the vectors that will be used in the normalization formula (3) to scale external losses. We keep to the different variables covered in Dahen and Dionne (2010) to explain the variation in the logarithm of operational losses:

- Total Assets is the proxy used for size since it had the highest correlation with our dependant variable. This is very similar to the works of Shih (2001) or Na et al. (2006).
- A dichotomous variable for each of the business lines has been included in the model. These are the standardized classes put forth by the Basel Committee: retail brokerage, trading and sales, commercial banking, retail banking, agency services, corporate finance, asset management as well as payment and settlement. One will be excluded and used as reference to avoid a binary trap. The very different nature specific to each business line's activities justify well their inclusion in the model.
- Also in accordance with the Basel Committee's standard definition, we include the different risk types that can be encountered in operational risk: damage to physical assets, clients/products/business practises, employment/practises/workplace safety, external fraud, internal fraud and execution/delivery/process management as well as business disruption/system failures, omitting one for the dummy trap.

The difference with Dahen and Dionne (2010) is that we do not include dichotomous variables for time (years or quarters) because of collinearity issues that arise from the regime-based structure we employ. Equation (4) actually takes the shape of (d) with size, business lines, and risk types forming vector X'_{it} and the coefficients for scaling depend on the operational loss process being in a high or low regime.

5.3 Validation of the scaling mechanism

The power-law relationship described in section 3 can be verified empirically once we scale the external losses to the level of one bank using formula (3). Due to parsimony of our sample data, we restrict the scaling to observations that fall in the

observed interval (maximum and minimum) of operational losses of the analyzed bank holding company.

We will first use a one-sample Kolmogorov-Smirnov test[10] to verify that the internal and normalized data do not reject a specific reference distribution. We then proceed with a two-sample Kolmogorov-Smirnov test to conclude that they do (or do not) statistically differ from each other, mainly in their moments. If we can show that the two distributions do not differ statistically from each other, then we can conclude with greater certainty than by using the method proposed in Na et al. (2006) exposed in section 2.1.3 that the scaling method is acceptable.

A great advantage of the two-sample Kolmogorov-Smirnov test is that does not require a specific reference distribution to compare the two distributions. This is a very useful property since we are dealing with extreme value theory distributions and would rather like to verify, irrespective of a specific distribution, if the two are not statistically different.

The method analyzes the data as an empirical distribution function (EDF) and creates a cumulative distribution function of n steps:

$$EDF_t = \frac{number\ of\ elements\ in\ the\ sample \leq t}{n}.$$

These are individually compared to what the values should have been given the sample's first moments at each of those steps, more specifically, the supremum (least element) at each step, and the equivalent steps from the cdf of the distribution it is being compared to. The two-sample version of the test performs a similar process but uses the EDF distributions of both samples instead of the sample and a generated known type of cumulative distribution function.

[10] See Kolmogorov (1933) for the theoretical founding of this particular test.

The tests will be done using SAS software because of its superior flexibility over the STATA version of the Kolmogorov-Smirnov test. The p-values generated by the software are based on statistical values for the maximum differences calculated and exposed in Smirnov (1948).

6. Empirical Analysis

As briefly mentioned in section 5, we analyze a database of annual losses of bank holding companies between 1994 and 2010 as well as a quarterly database for added observations between 2001 and 2010. The bank holding companies included in the final databases are the ones with over 1B$ in total assets and values are in 2010 (or 2010 fourth quarter) USD.

6.1 Annual regression

The data lead us to believe the distribution of losses has a large right tail (positive skew) and would probably fall under those generally found in extreme value theory. Table 1 demonstrates how diverse the data is with maximum values that are exponentially bigger than even the median of the variables for loss or total assets, consequentially giving us very large standard deviations and kurtosis.

Table 1- Descriptive Statistics (1994-2010)

	Total Assets	Loss
Number of observations	7037	623
Average (M$)	79,651	102.74
Standard Deviation (M$)	305,140	636.79
Kurtosis	28.45	148.37
Skewness	5.22	11.76
Minimum (M$)	1,000	1.01
25th Percentile (M$)	1,691	2.81
Median (M$)	3,222	7.81
75th Percentile (M$)	12,002	31.72
Maximum (M$)	2,297,755	8624.64

The data give us an early indication that the probability of having a loss is fairly low, as only 125 banks of the 1,137 had losses over $1M between 1994 and 2010. The variables used in the first step probit regression are mean salary, real GDP growth, bank capitalization as well as the dichotomous variable created for the high regime. The first and last percentiles of mean salary were replaced by the inclusive value of the 1st and 99th, 33 and 208 respectively. This makes 141 values of the 7037 that were changed. Although the average of mean salary was 68,000$ per year, some bank holding companies were displaying nonsensical values such as 1,000$ and 3M$. Excluding them would have made us lose 19 of our 623 losses from JP Morgan, Goldman Sachs and Morgan Stanley. This is probably due to some bonus pay during the end of the crisis that was included in the salary and benefits variable used to calculate mean salary. We made sure to verify that this correction did not cause any major differences in the regression and have included an additional regression for comparative purposes in the Appendix.

As depicted by the shaded area in Figure 1, the high regime found in the natural logarithm of the average operational losses per year starts in 2003 and continues until the very end of the period studied. It is easily observable that the level of losses conditional that there be any (the severity) have risen significantly as opposed to the 1994-2001. Despite the fact we can see peak and trough dynamics with peaks easily distinguishable in 2005 and 2008, the level of losses have not since regressed back to the levels prior to 2001. Two vertical lines were included for the beginning and end of the last NBER recession (December 2007 until June 2009[11]). Dummy variables would have been too late to capture the rise in loss levels and would not have included the 2003-2007 rise, or the one visible at the end of the period; NBER announcements that came around a year later (December 2008 and September 2010) would have been even less precise.

[11] As advocated on the website of the National Bureau of Economic Research: http://www.nber.org/cycles.html

Figure 1- Average Annual LN (operational losses) between 1994-2010
Note: shaded area corresponds to the high endogenous regime while the two vertical lines correspond to the beginning of the latest NBER recession (Dec. 2007) and its end (June 2009)

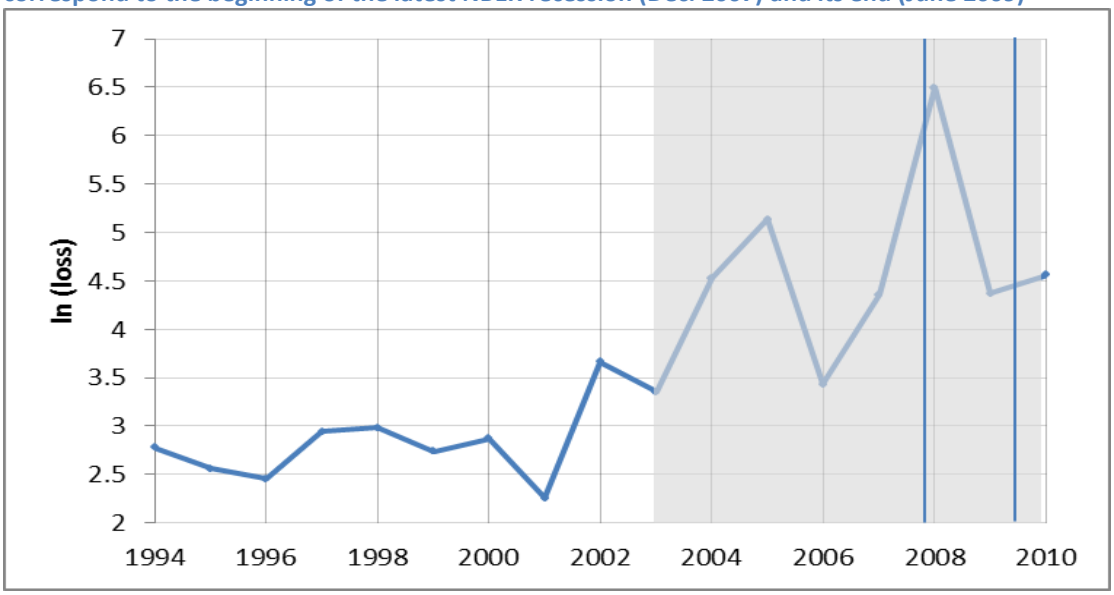


Table 2- Markov endogenous regime-switching parameters (Standard errors)

Parameters	
μ_1	2.775 (0.205)
μ_2	4.488 (0.420)
σ_1	0.320 (0.155)
σ_2	1.205 (0.607)
p_{11}	0.937 (0.076)
p_{22}	0.932 (0.083)
$\hat{\rho}$	0.516

Table 2 displays the parameters that populate the MLE $\hat{\theta}$ as a result of the Markov endogenous regime-switching algorithm. Since it was run on the natural logarithm of the mean of all annual operational loss events observed, the parameters are also in LN form, as is Figure 1. The $\hat{\rho}$ indicates the first observation probably came from the first distribution, which happens to be the lower mean and variance one that spans the period between 1994 and 2003. The high p_{11} and p_{22} values testify to the reluctant behavior of the process to revert back to historical level of losses.

As for the second step of the Hamilton (1979) model, the OLS will be comprised of the natural logarithm of assets, dichotomous variables for each business line and risk type, an interaction variable between the high regime and the business lines and risk types as well as the inverse mills ratio calculated from the probit residuals which is supposed to correct for potential selection bias.

The retail banking business line accounts for 38% of all operational losses observed while commercial banking accounts for 20%. Those two categories also had the biggest increase of severity in losses during the 2008 crisis. Dahen and Dionne (2010) had chosen the Payment and Settlement business line as a reference category to omit since it had by far the highest average losses in their database, thus expecting negative coefficients for business lines in comparison. Although the average losses remain high in that category, the results are a lot more mitigated with the addition of new losses between 2004 and 2010 as seen in table 3.

More than 51% of losses are of the clients, products and business practises risk type, which also accounts for 89% of almost all operational loss severity in our database. The losses are also on average a lot bigger when they stem from that type. Business disruptions have only been observed two times, and for minimal loss amounts. It will serve as our omitted category; we expect positive coefficients for the other categories. See statistics in table 4.

Table 3- Business Lines statistics (1994-2010)
Note : Results are in M$ according to the business lines in which the operational losses occurred. These include RBr: Retail brokerage, PS: Payment and settlement, CF: Corporate finance AM: Asset management, TS: Trading and sales, AS: Agency services, CB: Commercial banking, RB: Retail banking

LOSSES	RBr	PS	CF	AM	TS	AS	CB	RB
Average	10.36	83.19	533.23	85.15	104.68	48.04	28.15	106.455
Number	73	29	41	46	49	23	124	238
Std. Dev.	17.92	105.39	1,477	149.82	212.62	94.94	56.29	792.63

Table 4- Risk Type statistics (1994-2010)
Note : Results are in M$ according to risk types. These include DPA: Damage to physical assets, CPBF: Clients, products, and business practises, EPWS: Employment practises and workplace safety, EF: External fraud, IF: Internal Fraud, EDPM: Execution, delivery, and process management, BDSF: Business disruption and system failures.

LOSSES	DPA	CPBP	EPWS	EF	IF	EDPM	BDSF
Average	44.44	177.22	13.61	23.14	21.23	43.73	4.22
Number	3	216	19	60	73	32	2
Std. Dev.	47.13	879.07	22.60	45.02	54.19	130.84	2.16

We created the interaction variables between the high endogenous regime and the business lines and risk types from the matching periods. The coefficients are to be added to the ones found in the low regime for those observations as shown in (e).

Regression results for the 2-step Heckman (1979) process using the same variables as Dahen and Dionne (2010) are exposed in table 5 in order to get an idea of the marginal value added of our expanded dataset, which will be compared with our method exposed in table 6.

Table 5- Dahen and Dionne (2010) regression with new data

Variable	Coefficient (robust P-value)
Probit	
Bank_capitalization	-0.3954 *(0.375)*
Mean_salary	0.0131*** *(0.000)*
Real_GDP_growth	0.0285** *(0.008)*
Constant	-2.3607*** *(0.000)*
Pseudo R^2	0.0858
Regression Equation	
Constant	-0.7886 (0.248)
Log_assets	0.1748*** *(0.000)*
Retail Brokerage	-1.7072*** *(0.000)*
Trading and sales	-0.6566 *(0.106)*
Commercial banking	-0.6212** *(0.017)*
Retail banking	-0.8366*** *(0.009)*
Agency services	-0.3865 *(0.217)*
Corporate finance	0.3934 *(0.277)*
Asset management	-0.2045 *(0.640)*
Damages to physical assets	0.4279 *(0.788)*
Clients, products, and business practices	1.7749*** *(0.000)*
Employment, practices and workplace safety	1.3202*** *(0.000)*
External fraud	1.4589*** *(0.000)*
Internal fraud	1.0000*** *(0.001)*
Execution, delivery, and process management	1.3370*** *(0.000)*

(GLS second stage continued from table 5)

Year_1995	**0.2624** ***(0.349)***
Year_1996	0.3250 *(0.222)*
Year_1997	0.1990 *(0.467)*
Year_1998	-0.0078 *(0.975)*
Year_1999	0.1562 *(0.560)*
Year_2000	0.1011 *(0.605)*
Year_2001	-0.2716 *(0.330)*
Year_2002	0.4163* *(0.067)*
Year_2003	0.4981** *(0.037)*
Year_2004	0.6696*** *(0.001)*
Year_2005	0.0657 *(0.814)*
Year_2006	-0.2223 *(0.367)*
Year_2007	0.4330 *(0.123)*
Year_2008	0.8021*** *(0.004)*
Year_2009	0.3591 *(0.261)*
Year_2010	0.3452 *(0.212)*
Inverse Mills	-0.0238 *(0.906)*
R^2	22.31%

Table 6 – Our model

Variable	Coefficient (robust P-value)
Probit	
High Regime	-0.1713*** *(0.000)*
Mean_salary	0.0135*** *(0.000)*
Constant	-2.2589*** *(0.000)*
Pseudo R^2	0.0858
Regression Equation	
Constant	-0.8824 (0.230)
Log_assets	0.1720*** *(0.000)*
Retail Brokerage	-2.6821*** *(0.000)*
Trading and sales	-0.5010 *(0.540)*
Commercial banking	-1.8942*** *(0.000)*
Retail banking	-2.2585*** *(0.000)*
Agency services	-1.3648*** *(0.001)*
Corporate finance	-1.3648*** *(0.001)*
Asset management	-1.1750* *(0.071)*
Damages to physical assets	2.1324** *(0.024)*
Clients, products, and business practices	3.0503*** *(0.000)*
Employment, practices and workplace safety	2.7590*** *(0.000)*
External fraud	2.9259*** *(0.000)*
Internal fraud	2.3471*** *(0.000)*
Execution, delivery, and process management	2.7520*** *(0.000)*

(GLS second stage continued from table 6)

Regime · Retail Brokerage	1.0827** *(0.032)*
Regime · Trading and sales	0.0646 *(0.945)*
Regime · Commercial banking	1.6624*** *(0.000)*
Regime · Retail banking	2.0364*** *(0.000)*
Regime · Agency services	1.0864** *(0.048)*
Regime · Corporate finance	2.8226*** *(0.000)*
Regime · Asset management	1.2279 *(0.122)*
Regime · Damages to physical assets	-3.5635*** *(0.000)*
Regime · Clients, products, and business practices	-1.2180*** *(0.003)*
Regime · Employment, practices and workplace safety	-1.7415** *(0.013)*
Regime · External fraud	-1.6736*** *(0.000)*
Regime · Internal fraud	-1.4076*** *(0.000)*
Regime · Execution, delivery, and process management	-1.4504*** *(0.004)*
Inverse Mills	0.0703 *(0.682)*
R^2	23.20%

Table 5 exposes the model put forth by Dahen and Dionne (2010) with the addition of 323 new losses and 3,387 new observations between 2004 and 2010 to best grasp the marginal effect of the database. The probit variable coefficients have the same signs, as would be expected. It is clearer that bank capitalization is not a very good indicator of the probability of incurring a loss in our database. As for the GLS random effects targeted around the bank identities with robust p-values, many interesting differences become evident as opposed to their original regression spanning the 1994-2003 period available in table 13 of Appendix A.

The R^2 diminishes although that is not disastrous since variable significance is the ultimate goal for the scaling mechanism. The log assets variable remains positive as expected and statistically significant at the 99% level, which is crucial since it is our main scaling variable. The business line results are a lot less appealing, going from 5 statistically significant in the original article to only 3 in this regression. All their coefficients become lower, but that is to be expected since our reference category, payment and settlement is not necessarily the one with the largest average operational losses anymore as witnessed in table 3. The risk type variables become more significant, with 5 statistically significant variables versus 4 in the original article. This gives a final count of 10/14 variables available for the scaling formula as opposed to 9/14, which can be considered an amelioration. The inverse mills ratio is even less significant, thus concluding that selection bias is not really a problem for our regression.

The dichotomous variables for the years, not useful in Dahen and Dionne (2010) have taken on meaning. Years 2002, 2003, 2004 and 2008 become statistically significant, and contribute positively in the determination of loss levels, which is intuitive since they correspond to easily visible peaks in Figure 1. This does not replace the regime variable we are analyzing in the regression found in table 6, since they do not allow us to test for a change in dynamics of the operational loss process.

It indirectly makes the assumption that the dynamics of operational risk remain constant throughout the entire period studied, despite the peaks and higher levels observed post 2002.

Table 6 analyzes that possibility by using the interaction variables composed of the high regime dichotomous variable with the business lines and risk types. We do not include an interaction with log assets since we do not expect its effect to change from one regime to the other. The results are very pleasing when compared to those of table 5. The first thing we notice is that the real GDP growth and bank capitalization variables lose their statistical significance with the introduction of the dichotomous variable for when the process is in a higher regime, which is as statistically significant as mean salary. The mean salary variable remains positive and significant in explaining the probability of higher losses, possibly because of risk taking involved when large banks have a high relative percentage of qualified employees.

Besides the problematic business line of trading and sales, all other business lines and risk types are statistically significant at the 90% confidence level which is the one retained by Dahen and Dionne (2010) to be used for the scaling mechanism and most even at the 99% level. We observe that business lines are all negative and risk types positive during the low regime, which is somewhat expected but reverse during times of high regime. The business line coefficients, when the interaction coefficients are added to the low regime ones in order to get the high regime coefficients show that their effects almost completely disappear.

The business lines, although retaining their statistical significance quantitatively explain very little in the operational loss amounts when the process is in an endogenous high regime. The risk type coefficients also reverse, but on average remain positive in explaining the loss amounts during those same periods, making the drivers of explanatory power the log assets and risk types post 2002. Besides

explaining an additional 1% of loss amounts, the new model proposed allows us to utilize 13 of 14 potential scaling variables, which is a great deal better than the 9 available from the regression in table 5. It is important to understand that the Markov endogenous regime-switching algorithm is based on alarmingly low amount of observations: 16 yearly periods that include the average of the natural logarithms of operational losses. Although it is difficult to defend the use of this iterative process on such a small sample, we expect the relatively large standard deviation exposed in table 1 to help in determining concrete differences between the two distributions. We also expect the precision of this method to increase as time passes and more data on operational losses are collected. It would also be interesting to see if levels will ever revert to where they were prior to 2001. Results of the same regression performed on annual data that exclude the 1^{st} and 99^{th} percentiles rather than replace them as was done here is included in the Appendix in table 17 to show that there was no great distortion in using this dataset in particular. Besides a slight decrease in R^2, no large deviations in any of the variables can be observed.

Given the importance of the scaling variables and the relatively small size of our conditional loss sample, we proceed by testing the robustness of our variables during the second stage GLS process that is used in the normalization formula. Results are exposed in table 7. We test, as Shih (2000) had done to see if the size proxy has a statistically significant relationship with the loss amounts, in our case log losses. Unlike they had done, we include the inverse mills ratio in case of selection bias from the reporting of our loss database.

Table 7- Robustness tests

Regression Equation	***Model 1***	***Model 2***	***Model 3***
Constant	0.4933 (0.484)	-0.6867 (0.287)	-1.2587*** (0.032)
Log_assets	0 .1911*** *(0.000)*	0.1567*** *(0.000)*	0.1696*** *(0.000)*
RBr			-2.4947*** *(0.000)*
TS			
CB			-1.7067*** *(0.000)*
RB			-2.0737*** *(0.000)*
AS			-1.1814** *(0.016)*
CF			-1.7569*** *(0.003)*
AS			-0.9910* *(0.232)*
DPA		1.3452 *(0.427)*	2.4300*** *(0.009)*
CPBP		1.4261*** *(0.000)*	3.2528*** *(0.000)*
EPWS		0.6968*** *(0.005)*	2.9618*** *(0.000)*
EF		1.2298*** *(0.000)*	3.1252*** *(0.000)*
IF		0.6657*** *(0.001)*	2.5363*** *(0.000)*
EDPM		0.8669*** *(0.004)*	2.9530*** *(0.000)*
Reg · RBr			**1.1891*** ***(0.086)***
Reg · TS			
Reg · CB			1.7566*** *(0.000)*
Reg · RB			2.1387*** *(0.000)*
Reg · As			1.1906* *(0.100)*
Reg · CF			2.9273*** *(0.000)*
Reg · AM			1.3309 *(0.154)*
Reg · DPA		-2.6172 *(0.121)*	-3.7492*** *(0.000)*
Reg · CPBP		0.5349** *(0.020)*	-1.3178*** *(0.006)*
Reg · EPWS		0.0026 *(0.995)*	-1.8687*** *(0.001)*
Reg · EF		0.1771 *(0.435)*	-1.7571*** *(0.000)*
Reg · IF		0.3413 *(0.181)*	-1.4995*** *(0.001)*
Reg · EDPM		0.6069* *(0.074)*	-1.5494*** *(0.009)*
Inverse Mills	-0.2689 *(0.277)*	-0.1530 *(0.530)*	0.0813 *(0.650)*
R^2	6.16%	13.04%	23.01%

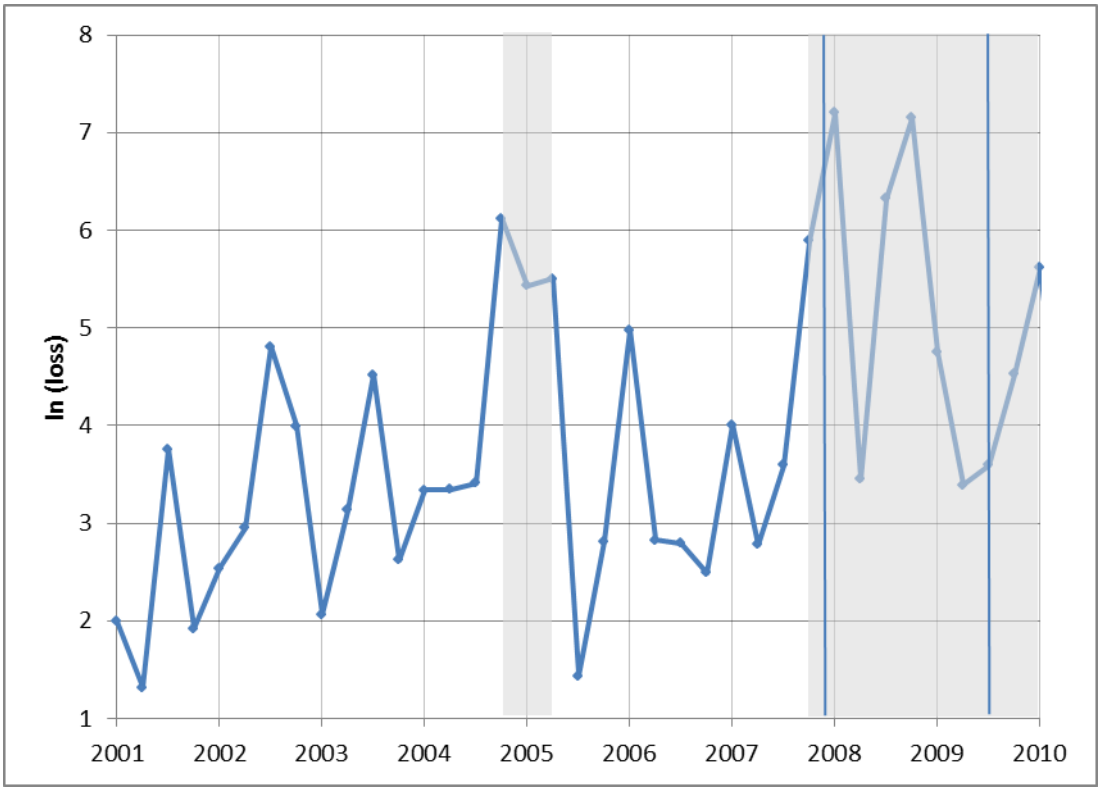
We see that log assets stays stable throughout the models studied. Model 1 shows us that taken alone, in concordance with the findings of Shih (2000), the size proxy does not explain a great deal of the conditional level of losses. Model 2 verifies how the inclusion of the risk type variables affect the regression. We see the explanatory value more than doubles while the log assets variable stays firm and maintains its high level of statistical significance. The low regime coefficients are evidently more stable than the interaction ones, although a couple remained significant even though more than half (348) of losses are found in the high regime. This could be caused by the large overconcentration of losses in the Clients, products, and business practices which is the most statistically significant regime interaction risk type variable as seen in table 3. This may explain why our model exposed in table 6, which includes the business lines that have much more evenly distributed losses, displays much more statistically significant results. Model 3 excludes the trading and sales business line as it had very inconclusive results as can be observed by the regressions in table 5 and 6. The results are very convincing as all variables except the regime interaction variable for the asset management business line (which was not statistically significant in our table 6 regression) are significant at the 90% confidence level. It is important to note that the explanatory value for model 3 resembles the one put forth in table 6 and the inverse mills ratio remains non-significant for all models analyzed.

6.2 Quarterly regression

As mentioned, we also analyze a regression using a quarterly database to get a better sense of the Markov endogenous regimes. Descriptive statistics of this subset of our database is found in the Appendix in tables 14, 15 and 16. Ideally, we would have had 68 observations, but as mentioned in section 4, there is an evident reporting bias prior to the Basel II and the official definition of operational risk. A great majority of losses were reported on January 1st or December 31st between 1994 and 2010.

The quarterly average natural logarithm of losses graph in Figure 2 shows additional evidence that simply putting a dummy variable for the NBER recession would not suffice in capturing the movement in the data. It would not have explained the high regime found between the third quarter of 2004 and the first quarter of 2005 lasting 3 periods as well as the apparent increase in average losses toward the end of the examined period which happens to be after the NBER recession ends.

Figure 2- Average Quarterly LN (operational losses) between 2001-2010
Note: shaded area corresponds to the high endogenous regime while the two vertical lines correspond to the beginning of the latest NBER recession (Dec. 2007) and its end (June 2009)



Although the added data should help in identifying regimes in a more precise manner and giving more credibility to the Markov regime-switching algorithm results (40 observations), some new issues arise. It is evident by the multiple peaks and troughs of Figure 2 that there is a more intricate bias that continues past 2001; probably deeply rooted and associated with accounting methods, mainly with the second quarters that seem persistently low. Losing the years associated with the low regime from the annual regression also remove any indication of the historically lower operational loss levels in the data. This is very evident by the results of the Markov regime-switching algorithm whose results are found in table 8. The means of the two theoretical distributions found are a lot closer to one another, which can be expected to mitigate the results of the overall regression.

Table 8- Regime parameters

Parameters	
μ_1	4.934 (0.681)
μ_2	3.133 (0.307)
σ_1	2.121 (0.891)
σ_2	1.050 (0.365)
p_{11}	0.810 (0.219)
p_{22}	0.885 (0.098)
$\hat{\rho}$	0.377

Proof of these concerns is also evident in the regression results of table 9. The explanatory power is relatively higher (R^2= 28.62%) than the annual regression but causality is mitigated by the loss of statistical significance of a few of the potential scaling variables.

Table 9 – Quarterly regression results

Variable	Coefficient (robust P-value)
Probit	
High Regime	0.0512 *(0.244)*
Mean_salary	0.0099*** *(0.000)*
Constant	-2.8183*** *(0.000)*
Pseudo R^2	0.0858
Regression Equation	
Constant	-2.8980 (0.008)
Log_assets	0.1664*** *(0.000)*
Retail Brokerage	-2.2330*** *(0.000)*
Trading and sales	-1.9061** *(0.010)*
Commercial banking	-1.0618*** *(0.002)*
Retail banking	-1.6019*** *(0.000)*
Agency services	-0.1577 *(0.724)*
Corporate finance	-1.5054** *(0.023)*
Asset management	-0.6275 *(0.148)*
Damages to physical assets	2.6451** *(0.023)*
Clients, products, and business practices	4.0049*** *(0.000)*
Employment, practices and workplace safety	3.6638*** *(0.000)*
External fraud	4.0193*** *(0.000)*
Internal fraud	3.7973*** *(0.000)*
Execution, delivery, and process management	4.2087*** *(0.000)*

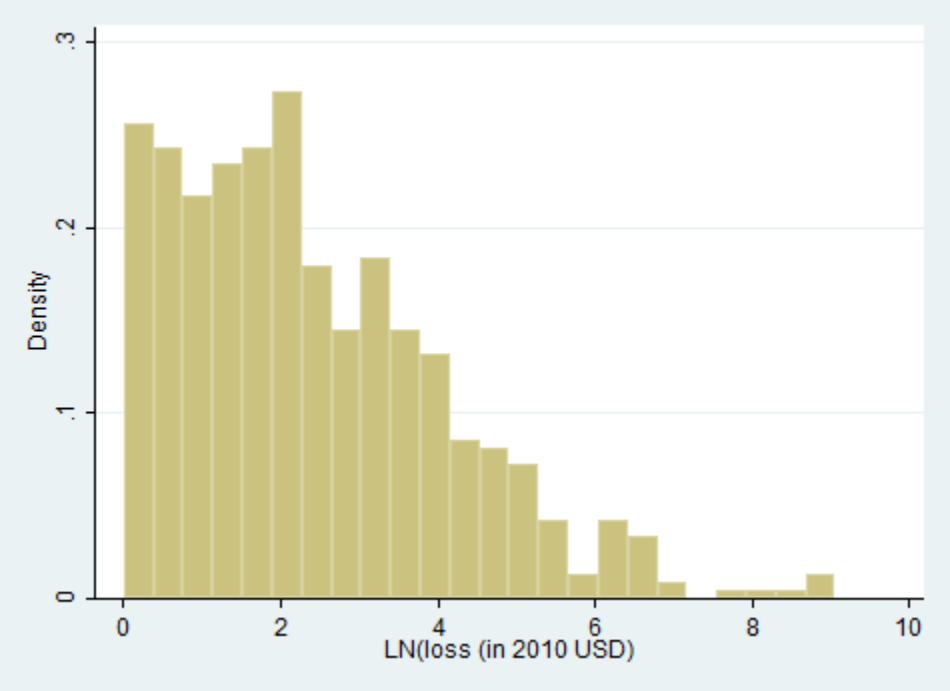
(GLS second stage continued from table 8)

Regime · Retail Brokerage	0.9529 *(0.221)*
Regime · Trading and sales	2.3052*** *(0.004)*
Regime · Commercial banking	1.2296*** *(0.020)*
Regime · Retail banking	2.2420*** *(0.000)*
Regime · Agency services	-0.2945 *(0.708)*
Regime · Corporate finance	3.9780*** *(0.000)*
Regime · Asset management	1.0687 *(0.154)*
Regime · Damages to physical assets	(omitted)
Regime · Clients, products, and business practices	-0.8844 *(0.159)*
Regime · Employment, practices and workplace safety	-1.6609** *(0.042)*
Regime · External fraud	-1.4815** *(0.027)*
Regime · Internal fraud	-2.1285*** *(0.000)*
Regime · Execution, delivery, and process management	-2.1077*** *(0.002)*
Inverse Mills	0.2490 *(0.326)*
R^2	28.62%

6.3 Testing the scaling mechanism

The log loss variable depicted in Figure 3 naturally takes the shape of what tends to look like a lognormal distribution. There is a large concentration of losses on the left side of the distribution, yet high losses, quite extreme in some cases (such as the >8B$ losses suffered by Citigroup Inc., Bank of America Corporation, and Wells Fargo & Company in 2008), persist and cause the right tail to remain much larger than say a normal distribution.

Figure 3 - Histogram of LN (losses) 1994-2010



Staying true to the comparison with the results found in Dahen and Dionne (2010), we analyze the scaling mechanism using US Bancorp as our reference bank holding company and test the goodness-of-fit of a lognormal distribution. Table 10 shows the statistics on the scaling that was performed. The coefficients used are the ones associated with our size proxy, business lines and risk types from Model 3 found in

table 7 that keeps only the statistically significant variables from our original Annual regression of table 6; recall that the coefficients differ depending on the regime.

Table 10 - Statistics on Internal and Scaled Annual Losses 1994-2010

	Observed losses in US Bancorp	Scaled losses within the same (Min-Max) interval as US Bancorp
Average (M$)	15.45	17.22
Median (M$)	6.25	8.40
Standard Deviation (M$)	23.85	20.19
Kurtosis	7.223	2.778
Skewness	2.61	1.871
Minimum (M$)	1.54	1.54
Maximum (M$)	96.11	96.11
Number of losses	19	224

Although the averages and standard deviations of the datasets seem somewhat similar, it is easily visible that this does not hold true for the other moments when comparing the two datasets. US Bancorp has not incurred many more losses since 2003 (19 vs. 15); period analyzed by Dahen and Dionne (2010).

Using a one-sample Kolmogorov-Smirnov test on the scaled sample with the lognormal distribution as the reference distribution does not prove conclusive. The average and standard deviation of the lognormal distribution are estimated at 1.95

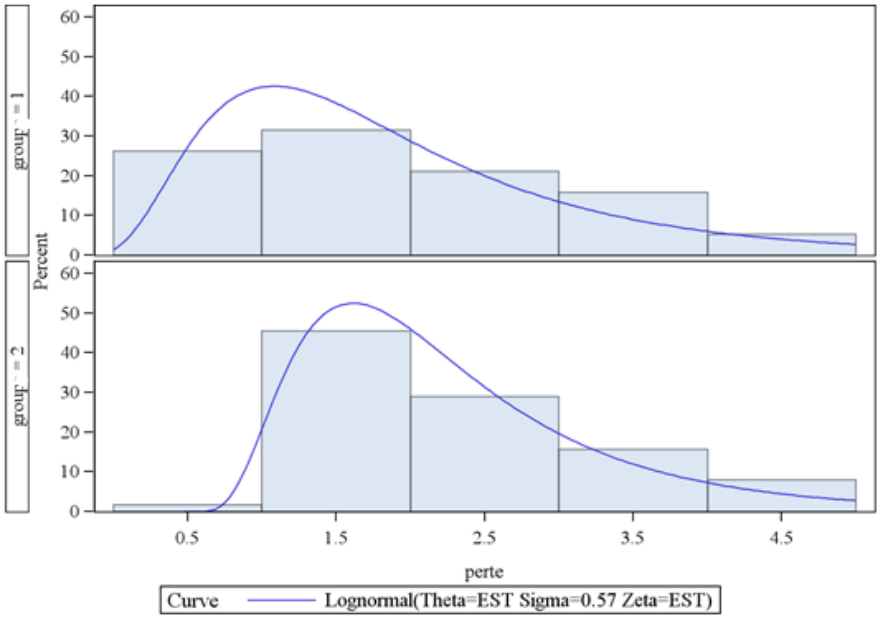
and 1.41 respectively. We can see the results of the fitting as Group 2 in table 12. Group 1 is the internal data; hence the less evenly distributed histogram under the lognormal estimated curve that serves as the EDF. The internal data does not reject the lognormal distribution as a possibility, but a well-known flaw with the Kolmogorov-Smirnov test lies in its difficulty to reject the null hypothesis of the distributions being equal when there are a small number of data points. It rejects the lognormal distribution for the scaled data ($p>0.5$). The results of the one-sample test in Dahen and Dionne (2010) had a p-value over 15%. This can probably be explained by the relatively high changes in standard deviations of operational losses that happened with the inclusion of the late 2007 recession.

Table 11 exposes the differences between the lognormal fitted distribution for the US Bancorp scaled losses and the theoretical EDF for the lognormal in greater detail, by analyzing steps at different quantiles. It is evident that the quantiles differ greatly toward the right tail. These results fall in line with the findings of Dahen et al. (2010) which empirically show that the lognormal right tail underestimates extreme losses as opposed to other extreme value theory distributions such as Pareto distribution.

Table 11- Quantile Analysis

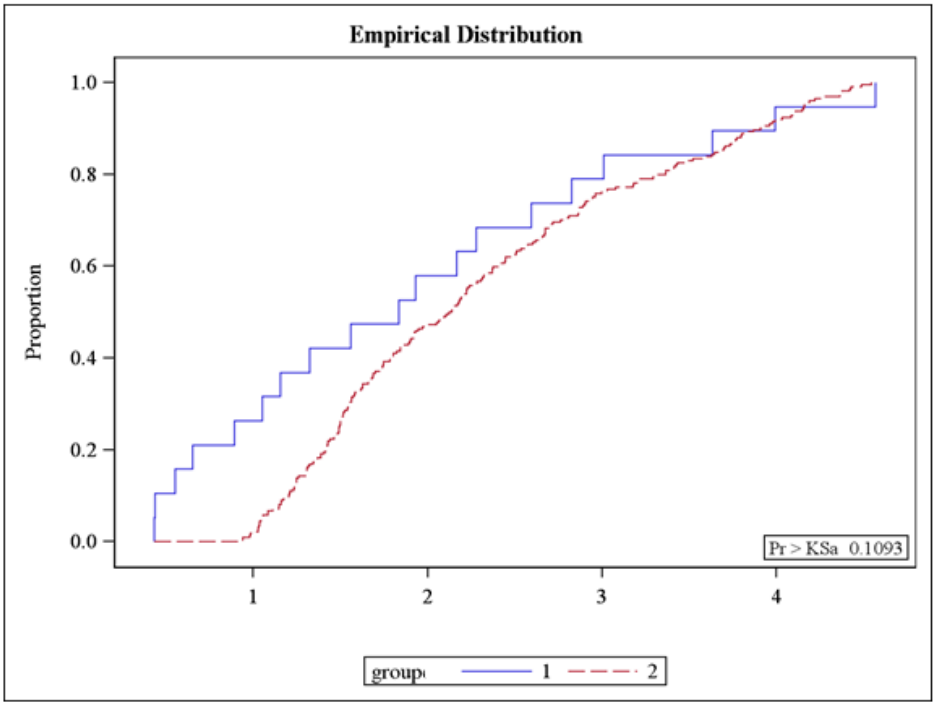
Percent	Observed (Scaled data)	Estimated (EDF)
5.00	0.43186	0.44097
10.00	0.43577	0.61521
25.00	0.89268	1.00034
50.00	1.83297	1.61761
75.00	2.82430	2.52428
90.00	3.98898	3.69869
95.00	4.56554	4.62248
99.00	4.56554	6.96691

Table 12- Kolmogorov-Smirnov Lognormal EDF (using LN (losses)
Group 1: External data normalized using US Bancorp's Minimum and Maximum observed losses
Group 2: Internal data of US Bancorp operational losses



As explained in section 5.3, this is not much cause for concern when it comes to assessing the pertinence of our scaling method. The two-sample test is much more important. Dahen and Dionne (2010) had found a p-value of 0.60 when comparing the internal distribution of US Bancorp and its 15 losses at the time with that of their scaled external dataset. In that respect, our results are very encouraging. We find a p-value of 0.1093. Figure 4 shows us the cumulative distribution functions of our two samples, where the steps discernible correspond to the 19 observations of US Bancorp. This result indicates that the scaling method developed in this study is appropriate to use when estimating internal operational losses for a bank holding company.

Figure 4 – CDF of both US Bancorp Samples used in two-sample Kolmogorov-Smirnov test



7. CONCLUSION

It is important not to lose sight of the context of scaling models for the severity of operational losses. As these can be potentially devastating, it is crucial to assemble additional data to aid in the future development of operational loss models. The scaling model proposed in this study is meant to make external losses useful to the bank holding company in need of observations to populate an operational loss database allowing them to get a better estimate of the all-important tails. The normalization formula copes with the potential selection bias and heterogeneity of the control environment of the banks that actually incurred these losses in order to scale effectively. The addition of data as well as the inclusion of Markov endogenous regimes added to the statistical significance and credibility of the scaling mechanism in question.

It is tempting to analyze the entirety of the data using a single regression since we have shown that there does not seem to be any danger of selection bias with the use of our external database. Doing so would result in explanatory power more than two times greater than those reported in table 6 ($R^2 \cong 70\%$). Our ultimate goal in this study was to lay the foundation (via our extended database and introduction of regime-based models) for the continuation of research towards a final distribution on which a VaR can be determined. This can only be accomplished once a scaling model for the frequency using models such as zero-inflated Poisson or Negative binomial are developed. Only then will the no-loss observations be analyzed appropriately in order to combine the two distributions (severity and frequency), a mathematical convolution, to create that final operational loss distribution.

There are surely distributions, popular in extreme value theory, that are much more representative of the nature of our data than the lognormal that we examined with a one-sample Kolmogorov-Smirnov test. The difficulty and mistake most make when applying such complex parametric distributions are violating or not even being able

to verify the very assumptions they are based on. These include distributions such as the Generalized-Pareto, G-and-H, GB2, Weibull, etc.

There is much room for improvement. It is necessary to analyze scaled datasets for the severity of operational losses more in depth than what was done in this study. As we have seen, although a quarterly database adds observations, they are probably riddled with a new bias due to loss reporting. It will also be interesting to see what additional information on these losses time will provide.

Despite the fact the high regime variables seem to help our scaling model greatly, there is a large nuance related to the normality assumption that is clearly violated in operational risk and many other processes that are based on the Markov regime-switching model in academia. It would be interesting to look into regime-switching models based on other overlapping distributions.

Appendix A: Tables

Table 13 - Original regression from Dahen and Dionne (2010)

Variable	Coefficient (robust P-value)
Probit	
Bank_capitalization	-1.2587 *(0.145)*
Mean_salary	0.0094*** *(0.000)*
Real_GDP_growth	0.0554** *(0.012)*
Constant	-2.0068*** *(0.000)*
Pseudo R^2	0.046
Regression Equation	
Constant	-0.1188 (0.941)
Log_assets	0.1482*** *(0.001)*
Retail Brokerage	-2.5840*** *(0.000)*
Trading and sales	-0.4807 *(0.604)*
Commercial banking	-1.7850*** *(0.000)*
Retail banking	-2.2250*** *(0.000)*
Agency services	-1.3789*** *(0.005)*
Corporate finance	-1.7572*** *(0.001)*
Asset management	-1.0412* *(0.088)*
Damages to physical assets	1.3779 *(0.365)*
Clients, products, and business practices	2.0372** *(0.018)*
Employment, practices and workplace safety	1.6798* *(0.057)*
External fraud	1.8783** *(0.030)*
Internal fraud	1.2612 *(0.137)*
Execution, delivery, and process management	1.8318** *(0.040)*
Year_1995	0.2661 *(0.383)*
Year_1996	0.3410 *(0.240)*

Year_1997	0.2381 *(0.485)*
Year_1998	0.0211 *(0.940)*
Year_1999	0.1828 *(0.531)*
Year_2000	0.2611 *(0.382)*
Year_2001	-0.4454 *(0.146)*
Year_2002	0.3000 *(0.328)*
Year_2003	0.3740 *(0.236)*
Inverse Mills	0.1929 *(0.705)*
Wald chi2(23)	166.89 *(0.000)*
R^2	29.58%

Notes: ***Coefficient significant at the 99% confidence level.
**Coefficient significant at the 95% confidence level.
*Coefficient significant at the 90% confidence level.

Their model is adjusted for heteroskedasticity. Omitted categories due to avoid dummy trap in the OLS are Year 1994, Payment and Settlement (Business Line), and Business disruption and system failures (Risk Type). The Wald $chi^2(5)$ and log likelihood for the probit model are 94.01 and -989.92 respectively. There are 3,650 observations used in the probit estimation and 300 conditional losses used in the OLS regression.

Table 14 - Descriptive Statistics from Quarterly database (2001-2010)

	Asset	Loss
Number of observations	17,122	404
Average (M$)	44,238	147.17
Standard Deviation (M$)	209,675	788.53
Kurtosis	66.30	95.55
Skewness	7.72	9.47
Minimum (M$)	1,000	1.01
25th Percentile (M$)	1,513	2.91
Median (M$)	2,594	9.93
75th Percentile (M$)	8,205	44.74
Maximum (M$)	2,479,088	8624.64

Table 15 - Quarterly database Business Lines

LOSSES	RBr	PS	CF	AM	TS	AS	CB	RB
Average	9.54	77.84	624.73	92.31	110	60.04	29.46	179.82
Number	48	28	35	39	45	13	64	132
Std. Dev.	17.32	103.1	1,595	160.73	222.2	118.17	45	1,056

Table 16 - Quarterly database Risk Types

LOSSES	DPA	CPBP	EPWS	EF	IF	EDPM	BDSF
Average	34.44	248.49	14.86	29.35	26.83	52.31	2.08
Number	3	216	19	60	73	32	1
Std. Dev.	47.13	1,064	25.24	57.61	65.21	145.81	-

Table 17- Our model excluding percentiles of Mean Salary

Regression Equation	***Model 1***
Constant	-0.6544 (0.360)
Log_assets	0.1654*** *(0.000)*
RBr	-2.6729*** *(0.000)*
TS	-0.5049 *(0.538)*
CB	-1.9309*** *(0.000)*
RB	-2.2279*** *(0.000)*
AS	-1.3476*** *(0.001)*
CF	-1.9451*** *(0.001)*
AS	-1.1637* *(0.076)*
DPA	2.1196** *(0.025)*
CPBP	3.0628*** *(0.000)*
EPWS	2.7542*** *(0.000)*
EF	2.9491*** *(0.000)*
IF	2.3477*** *(0.000)*
EDPM	2.7660*** *(0.000)*
Reg · RBr	**1.1005**** ***(0.034)***
Reg · TS	0.2640 *(0.788)*
Reg · CB	1.6949*** *(0.000)*
Reg · RB	2.0315*** *(0.001)*
Reg · As	1.0799** *(0.053)*
Reg · CF	2.7721*** *(0.000)*
Reg · AM	1.2321 *(0.123)*
Reg · DPA	-3.5219*** *(0.000)*
Reg · CPBP	-1.2278*** *(0.003)*
Reg · EPWS	-1.7706** *(0.015)*
Reg · EF	-1.6941*** *(0.001)*
Reg · IF	-1.4186*** *(0.000)*
Reg · EDPM	-1.5214*** *(0.003)*
Inverse Mills	-0.0268 *(0.906)*
R^2	22.78%

Table 18- Annual Loss breakdown

Year	Losses	Total Loss
1994	25	402.8
1995	23	297.24
1996	22	256.035
1997	26	494.75
1998	51	1004.4
1999	37	571.7
2000	35	617.6348
2001	18	171.523
2002	39	1525.347
2003	24	684.4344
2004	40	3708.72
2005	46	7832.93
2006	41	1267.95
2007	39	3043.92
2008	43	28433.06
2009	57	4506.16
2010	58	5551.8

Table 19- Annual Business Line breakdown

Year	RBr	TS	CB	RB	AS	CF	AM	PS
1994	0	1	3	16	3	0	2	0
1995	1	2	5	13	2	0	0	0
1996	2	0	5	13	2	0	0	0
1997	1	0	9	13	2	0	1	0
1998	7	1	22	16	1	3	0	1
1999	10	0	9	16	0	0	2	0
2000	4	0	8	18	0	3	2	0
2001	2	1	5	9	0	0	0	1
2002	5	0	6	19	3	0	1	5
2003	1	0	8	8	0	1	2	4
2004	2	4	5	11	3	5	8	2
2005	5	4	10	7	0	7	11	2
2006	4	5	6	15	2	5	2	2
2007	10	3	4	16	0	2	2	2
2008	2	6	8	16	0	5	4	2
2009	11	13	4	13	2	6	6	2
2010	6	9	8	18	3	4	4	6

Table 20- Annual Risk Type breakdown

Year	DPA	CPBP	EPWS	EF	IF	EDPM	BDSF
1994	0	13	0	7	5	0	0
1995	0	13	0	2	8	0	0
1996	0	10	1	7	4	0	1
1997	0	12	1	8	3	2	0
1998	0	15	3	19	8	5	1
1999	0	19	5	8	4	1	0
2000	0	22	4	2	7	0	0
2001	1	9	0	6	1	1	0
2002	1	12	3	12	7	4	0
2003	0	12	0	4	4	4	0
2004	0	22	2	4	8	4	0
2005	0	31	0	4	5	6	0
2006	0	17	8	5	8	3	0
2007	1	26	0	7	3	2	0
2008	0	23	2	5	10	2	1
2009	0	33	3	7	13	1	0
2010	0	32	1	6	14	5	0

Table 21- Average Loss Business Line breakdown

Year	RBr	TS	CB	RB	AS	CF	AM	PS
1994	0.00	0.01	0.05	0.12	0.49	0.00	0.45	0.00
1995	0.09	0.40	0.16	0.17	0.04	0.00	0.00	0.00
1996	0.22	0.00	0.22	0.24	0.05	0.00	0.00	0.00
1997	0.09	0.00	0.40	0.88	0.04	0.00	0.05	0.00
1998	0.16	0.01	1.41	0.57	0.05	0.06	0.00	0.52
1999	0.06	0.00	1.09	0.28	0.00	0.00	0.15	0.00
2000	0.04	0.00	0.18	1.18	0.00	0.19	0.02	0.00
2001	0.01	0.03	0.06	0.13	0.00	0.00	0.00	0.23
2002	0.04	0.00	1.10	0.89	0.15	0.00	0.28	1.41
2003	0.02	0.00	0.44	0.13	0.00	0.06	0.08	0.92
2004	0.07	0.25	0.22	0.54	0.24	6.15	0.80	0.06
2005	0.06	0.12	0.16	0.42	0.00	13.10	2.35	0.01
2006	0.23	0.06	0.41	0.35	0.99	0.11	0.02	0.55
2007	0.23	0.06	0.19	4.18	0.00	0.04	1.34	0.41
2008	0.07	3.33	0.12	34.64	0.00	19.14	0.43	0.41
2009	0.12	3.88	0.70	1.11	0.09	2.13	0.63	0.05
2010	0.06	1.91	0.45	4.97	0.07	1.73	1.36	0.48

Table 22- Average loss Risk Type breakdown

Year	DPA	CPBP	EPWS	EF	IF	EDPM	BDSF
1994	0.00	1.04	0.00	0.07	0.02	0.00	0.00
1995	0.00	0.29	0.00	0.13	0.43	0.00	0.00
1996	0.00	0.42	0.04	0.26	0.03	0.00	0.00
1997	0.00	0.83	0.17	0.42	0.02	0.02	0.00
1998	0.00	1.88	0.02	0.50	0.21	0.15	0.01
1999	0.00	0.91	0.05	0.44	0.18	0.01	0.00
2000	0.00	1.43	0.08	0.02	0.08	0.00	0.00
2001	0.23	0.09	0.00	0.14	0.00	0.00	0.00
2002	0.00	2.38	0.02	1.06	0.16	0.26	0.00
2003	0.00	1.38	0.00	0.07	0.08	0.13	0.00
2004	0.00	7.27	0.02	0.13	0.38	0.53	0.00
2005	0.00	15.75	0.00	0.04	0.38	0.04	0.00
2006	0.00	1.77	0.43	0.07	0.35	0.11	0.00
2007	0.00	5.85	0.00	0.30	0.11	0.18	0.00
2008	0.00	57.38	0.07	0.44	0.14	0.10	0.00
2009	0.00	5.80	0.02	1.22	1.67	0.00	0.00
2010	0.00	8.41	0.01	0.10	0.54	1.97	0.00

References

Acharya, V., Davydenko, S., Strebulaev, I., 2007. Cash Holdings and Credit Risk. Working Paper, University of Toronto.

Basel Committee on Banking Supervision, 2001. Operational risk-consultative document. Supporting Document to the New Basel Capital Accord.

Basel Committee on Banking Supervision, 2003. Third consultative paper. The New Basel Accord.

Chavez-Demoulin, V., Embrechts, P., Neslehova, J., 2006. Quantitative models for operational risk : extremes, dependence and aggregation. Working Paper, ETH-Zurich.

Chernobai, A., Jorion, P., Yu, F., 2010. The determinants of operational losses. Working Paper, Syracuse University.

Cummins, J.D., Dionne, G., McDonald, J.B., Pritchett, B.M., 1990. Applications of the GB2 distribution in modeling insurance loss processes. Insurance: Mathematics and Economics 9(4), 257-272.

Dahen, H.,Dionne G. 2007. Scaling models for the severity and frequency of external operational loss data. Canada Research Chair in Risk Management Working Paper 07-01.

Dahen, H., Dionne, G., 2010. Scaling models for the severity and frequency of external operational loss data. Journal of Banking & Finance 34, 1484-1496.

Dahen, H., Dionne, G., Zajdenweber, D., Summer 2010. A practical application of extreme value theory to operational risk in banks. Journal of Operational Risk 5, 2, 63-78.

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

Engel, C., Hamilton, J.D., 1990. Long swings in the dollar: Are they in the data and do markets know it? American Economic Review, 689-713.

Frees, E.W., Valdez, E.A., 2008. Hierarchical insurance claims modeling. Journal of the American Statistical Association 103, 1457-1469.

Hamilton, J.D., 1990. Analysis of time series subject to changes in regime. Journal of Econometrics, Vol 45, Issues 1-2, 39-70.

Heckman, J., 1979. Selection bias as a specification error. Econometrica 47(1), 153-161.

Johnson, N., Kotz, S., 1972. Distribution in statistics: continuous multivariate distributions, Wiley & Sons, New York, 333 p.

Kolmogorov, A., 1933. Sulla determinazione empirica di una legge di distribuzione. G. Inst. Ital. Attuari 4:83.

Maalaoui Chun, O., Dionne, G., François, P., 2010. Credit spread changes within switching regimes. SSRN: http://ssrn.com/abstract=1341870 or http://dx.doi.org/10.2139/ssrn.1341870

Na, H.S., Van Den Berg, J., Miranda, L.C., 2006. An econometric model to scale operational losses. The Journal of Operational Risk 1 (2), 11-31.

Shevchenko, P.V., Wuthrich, M.V., 2006. The structural modelling of operational risk via Bayesian inference: combining loss data with expert opinions. Journal of Operational Risk 1 (3), 3-26.

Shih, J., Samed-Khan, A.,, Medapa, P., 2000. "Is the size of operational loss related to firm size?" Operational Risk 2, 1-2.

Shih, J. (2001). On the use of external data for operational risk quantification. Technical Paper,Risk Architecture Group of Citigroup.

Smirnov, N.V., 1948. Table for estimating the goodness of fit of empirical distributions. The annals of mathematical statistics, Vol 19, No 2, 279-281.