An Internal Fraud Model for Operational Losses: An Application to Evaluate Data Integration Techniques in Operational Risk Management in Financial Institutions

by

Rocío Margaret Paredes Leandro

A thesis submitted in partial fulfillment of the requirements of the degree of DBA awarded by Pontificia Universidad Católica del Perú (Centrum Católica) and DBA awarded by the Maastricht School of Management (MSM)

June 2016

An Internal Fraud Model for Operational Losses: An Application to Evaluate Data Integration Techniques in Operational Risk Management in Financial Institutions

by

Rocío Margaret Paredes Leandro

June 2016

Approved:

Charles Vincent, Ph. D., Supervisor

Fernando D'Alessio, Ph. D., Committee Member

Beatrice Avolio, DBA, Committee Member

Sergio Chión, Ph. D., Committee Member

Luis Felipe Zegarra, Ph. D., Committee Member

Wim Naudé Ph. D., Committee Member

Accepted and Signed:_____ _____
Charles Vincent — June 2016

Accepted and Signed:_____ _____
Beatrice Avolio — June 2016

Accepted and Signed:_____ _____
Sergio Chión — June 2016

Accepted and Signed:_____ _____
Luis Felipe Zegarra — June 2016

_____ _____
Wim Naudé Ph. D. — June 2016
Dean of Maastricht School of Management

_____ _____
Fernando D´Alessio I. Ph. D. — June 2016
General Director of CENTRUM, PUCP

Abstract

The handling of external operational loss data by individual banks is one of the long-standing problems in risk management theory and practice. The extant literature has not provided a method to identify the best way to combine internal and external operational loss data to calculate operational risk capital. Hence, to improve the knowledge and understanding of internal-external data combination in operational risk management, this study applied a simulation-based evaluation of well-known data combination techniques such as the scaling, the Bayesian, and the covariate-base techniques.

This research considered operational losses arising from internal fraud in retail banking within a group of international banks that share data through an operational loss data exchange. One of the key elements of the simulation-based statistical evaluation was the development of a dynamic internal fraud model for operational losses in retail banking. The internal fraud model incorporated human factors such as the number of employees per branch and the ethical quality of workers. It also included the extent of risk controls set by bank managers.

There were two sets of findings. First, according to the simulation-based evaluation, the scaling technique was by far the less useful for estimating the appropriate operational risk capital. The Bayesian and the covariate-based techniques performed best. The Bayesian technique was the best for higher percentiles while the covariate-based technique was the best at not so extreme quantiles. The choice of technique therefore depends on the risk appetite of the financial institution.

The second set of findings relates to the model validation with hard data. Losses generated by the model in the banks across the world were associated with GDP growth and the corruption perception of the country where banks were located. In general, internal fraud losses are pro-cyclical and the corruption perception in a country positively affects the

occurrence of internal fraud losses. When a country is perceived as more corrupt, retail banking in that country will feature more severe internal fraud losses. To the best of knowledge, it is the first time in the operational risk literature that this type of result is reported.

To

God and my family

Acknowledgments

My first thanks go to Professor Charles Vincent, my thesis supervisor. I thank him for his support during the completion of my thesis. His patience and wit allowed me to improve the entire thesis methodology and many specific aspects of the modeling.

I also thank Professor Fernando D'Alessio for making it possible for me to engage in this long endeavor to obtain the DBA. Thanks for his guidance, support and feedback during the entire doctoral process.

My gratefulness to professors Sergio Chión and Luis Felipe Zegarra for providing detailed comments to previous versions of the thesis. I readily considered all the comments and learned more in the process. This thesis has undoubtedly benefited from their observations.

My gratitude also goes to professors Beatrice Avolio, José Carlos Véliz and Giovanna Di Laura for their patience and support with all the handling of the thesis process.

Last but not least, my thanks to all my peers at the DBA program. I share with them good and stressful times in the Centrum campus and we all benefited from all the discussions and the great intellectual environment of Centrum. I learned a lot from everyone.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The aim of this study was to contribute to the solution of one of the key problems in quantitative risk management. The problem is the lack of a method to identify the best way to integrate internal and external loss data for operational risk management purposes in financial institutions. Thus far, academic risk management literature has not addressed this specific but important problem. In general, the handling of external data by individual financial institutions is one of the less developed tasks in risk management, and this gap has hindered further advances in the practical endeavor of handling operational risks.

The fact that banks have made insufficient progress in handling external data is evident, for example, in the review of principles of sound operational risk management made by the Basel Committee on Banking Supervision (BCBS, 2014). According to the review, only 26 out of 60 important banks had fully complied with the principle of using internal and external loss data. In addition, a global risk management survey of 71 financial institutions in 2014 showed that only a third of respondents felt that their financial institutions' external loss event data were extremely or very well developed (Deloitte, 2015).

In order to improve the knowledge and understanding of internal-external data integration in operational risk management, a simulation study is applied to determine the best data integration technique among well-known integration techniques proposed in the literature. A "simulation study" is a research method widely used in statistics. The theoretical basis for the simulation study is outlined in Greene (2012, Chapter 15) and Voss (2013, Chapter 3). No other rigorous attempts to compare capital risk-estimation techniques have been made. Only two studies in the operational risk literature resemble what is attempted in this study. Teker (2005) and Jiménez, Feria, and Martin (2009) compared three broad operational risk capital estimation methods: The basic indicator approach (BIA), the standardized method (SM), and a specific advanced method approach (AMA). The

conclusion in both studies was that the application of the AMA delivers much lower operational risk capital levels than those obtained by applying BIA or SM. In the current study, a simulation study was implemented to compare internal-external data integration techniques within the AMA framework.

The evaluation method in the current study considered operational losses arising specifically from internal fraud in retail banking within a group of international banks that share data through an operational loss data exchange. One of the key elements of the simulation-based statistical evaluation is the development of an operational loss model for internal fraud losses in retail banking. So far, the operational risk management literature has not provided a quantitative model for describing internal fraud in the financial sector; this study is the first to contribute in this direction. The internal fraud model borrows insights from a number of disciplines such as corporate governance, behavioral economics, human resources, and operational risk. Some of the ideas in the field of people risk management (Blacker & McConnell, 2015) have also been incorporated.

Operational risk, the study topic, has increased in importance since 2004 when the Basel Committee of Banking and Supervision (BCBS) published the Basel II Accord. One of the key issues of the Accord is that every bank must hold capital to afford operational risks. Operational risks are the risks of losses due to errors or failures related to personnel, internal processes, systems, or external events (BCBS, 2006). To calculate the amount of operational risk capital, Basel II allows the use of an AMA, but the institution must have the approval of a local regulator to implement an AMA. The approach refers to a variety of quantitative methods that reflect the specific characteristics of a bank. For that purpose, a bank can build its own empirical model to quantify its operational risk capital. A key feature of an AMA is that it relies on both internal and relevant external data, in other words, data from within the institution and from other institutions and even countries.

Many banks have adopted the loss distribution approach (LDA) as an AMA to build their own operational risk capital models. The LDA is an actuarial method that models the frequency and severity of operational losses separately and combines them to estimate a probability density function (PDF) for the aggregate losses for a given period. The PDF provides valuable information to estimate the operational risk capital; in particular, it is used to extract a specific extreme quantile of aggregate losses (value at risk). In turn, this extreme quantile is used to set the operational risk capital to comply with Basel II.

Within the LDA framework, various possible ways exist to combine and use internal as well as external data. In this study, the intention is to analyze three techniques: Scaling, the Bayesian technique, and a Covariate-based LDA. Specifically, the aim of the research is to devise an internal fraud model to be used as an operational loss simulator and then to evaluate the aforementioned data integration techniques in terms of derived operational risk capital. The assessment is relevant because the capital amounts held by financial institutions in order to afford operational risk losses could vary significantly according to the chosen technique and may distort competition in the financial sector.

Operational losses due to internal fraud in the retail-banking segment of financial institutions that belong to the Operational Riskdata eXchange Association (ORX) are considered. The ORX is a global data sharing association whose members comprise the biggest banks in the world. The data gathered covers 52 member banks between 2006 and 2010 that engaged in retail banking operations. The simulation-based evaluation of the data integration techniques requires the setup of a data generating process, namely, a model, for the outbreak of operational losses due to internal fraud events.

In this chapter, the background to the problem, the problem motivating the research, the purpose of study, the significance of the problem, the nature of the study, the research

questions and hypothesis, the theoretical framework, the definition of terms, the assumptions, and the limitations and delimitations are presented.

Background of the Problem

In contrast to the quantification of other risks that affect financial institutions like credit or market risk, operational risk measurement lacks sufficient data. Data for market or credit risk are readily available and abundant. Asset price data are available on many private or public electronic platforms, and financial institutions usually maintain databases on credit defaults to manage their credit risk. In sharp contrast, operational losses are diverse, heterogeneous, and infrequently documented, so no record of an event may exist for many years within a financial institution; however, there are significant latent events. The occurrence of latent events may wreak havoc in a financial institution if the firm is not sufficiently prepared or insured. In fact, low frequency, high severity operational losses have caused many bankruptcies around the world (Chernobai, Rachev, & Fabozzi, 2007).

Operational risk is not a small risk. For example, Ames, Schuermann, and Scott (2015) showed that economic capital requirements for operational risk increased, in terms of share of total capital, from an average of 9% to 13% between 2008 and 2012. In addition, The Banker Database (2015) reports that operational risk capital represented 12% of total risk capital among the five biggest UK banks in 2014. For big European, non-UK banks, Carmassi and Micossi (2012) reported that operational risk capital ranges between 10% and 30%.

Deloitte's 2014 global risk management survey assessed the financial industry's risk management practices. The survey included responses from 71 financial service institutions around the world. According to the survey, all financial institutions calculate economic capital to assess their risk. From them, 72% calculate market risk capital, 68% calculate credit risk capital, and 62% calculate operational risk capital (Deloitte, 2015).

Financial institutions that are using the AMA have been urged to use five years of data (BCBS, 2011a). The amount of data that can be collected by one institution, however, is insufficient to account for the true nature of operational risk. Some operational risks may not have enough loss data within a five-year window precisely because of the low-probability, high-impact nature of operational risks. To estimate operational risk capital with the AMA properly, data about this type of loss are required.

To circumvent the scarcity of data, as Chapter 2 describes, many database initiatives have been initiated since the onset of this century. These initiatives have been devoted to gathering operational loss data so that banks can anonymously share the data. The collection of this data is an ongoing but promising process. As the pace of improvements in data storage and management as well as advances in quantitative techniques continue, these shared databases will become more important.

Many financial institutions worldwide have already started applying the AMA using these datasets despite the many unsolved methodological challenges (Chaudhury, 2010; Shevchenko, 2011). As suggested by Shevchenko (2011), "The development of a consistent mathematical framework for operational risk treatment, addressing all aspects required in practical implementation is a challenging task" (p. vii). One of the unresolved issues is the treatment of internal and external datasets. In this context, a key question arises: What is the best technique that a financial institution can use to combine or integrate both internal and external data to apply the AMA?

In the operational risk literature, a number of techniques to integrate internal and external data in a financial institution have been proposed, but the literature is silent about which of these techniques performs best. Knowing which technique performs best is important because the estimate relates directly to the amount of capital a financial institution must hold to face operational risk. Excessive operational risk capital is costly to a financial

institution and harms its efficiency. Too little capital endangers the firm's solvency if a low frequency, high-impact event occurs. This raises another important question: How is too little, too much, or just the right level of operational risk capital determined?

Failure to answer these questions, namely, what is the best data integration technique, and how to determine the right level of operational risk capital, may hinder further progress in the implementation of the principles of sound operational risk management with respect to the treatment of external data (BCBS, 2014). The aim of this study is to shed light on the means to answer the above questions.

Other unresolved issues in the AMA literature exist that are not a focus of this study. For example, the AMA relies on historical data, and therefore, it is backward looking. The data do not reflect future risks like pending mergers, system integration, changing regulations, or the introduction of new products. Aspects of the operational risk problem as depicted, for example, by Haubenstock and Harding (2003), are not the focus of this study.

Statement of the Problem

One of the AMA Basel II requirements is that any operational loss measurement system must include internal and relevant external data. In practice, many possible techniques to combine internal and external operational loss data to a financial institution are available. The current study includes the evaluation of three prominent data integration techniques: Scaling, the Bayesian technique, and a Covariate-based LDA. Each technique could lead to different quantitative results in terms of the operational risk capital required for a banking institution. Practical implementation of these techniques in financial institutions must consider that the respective costs can alter the efficiency and degree of competition in the financial sector. It is critical for financial institutions that apply AMA or are moving toward AMA for operational risk capital calculations to know whether one technique performs better than others do and under what circumstances.

In essence, the idea is to compare operational risk capital estimations based on three techniques documented in the literature. Each technique maps the operational loss data into specific numerical estimates (operational risk capital levels). The task is to compare these numerical estimates. Therefore, a suitable methodological approach for this problem is a simulation study, which is a well-known statistical procedure that allows identifying features of either estimators or methods. In general, a simulation study allows learning the properties of objects that are functions of the data. As Burton, Altman, Royston, and Holder (2006) put it: "Simulation studies use computer intensive procedures to assess the performance of a variety of statistical methods in relation to a known truth. Such evaluation cannot be achieved with studies of real data alone" (p. 4279). Simulations are widely used in theoretical and applied statistics in a number of disciplines such as medicine, biology, psychology, physics, management, and economics. The simulation methodology is outlined in Greene (2012, Chapter 15) and Voss (2013, Chapter 3).

An essential feature of a simulation is a data generating process or model based on a given theory. In the case of this study, an operational risk model capable of generating operational losses that bear the same features as real observed losses is applied. This model construction is a key contribution of the study.

No other study in the operational risk literature has addressed the problem of evaluating data integration techniques in a rigorous way. Only two studies resemble the current study. Teker (2005) and Jiménez et al. (2009) compared three broad operational risk capital estimation methods but did not perform a simulation and therefore did not make inferential assessments. In this study, the simulation allows for making comparisons of internal-external data integration techniques.

Purpose of the Study

The aim of this quantitative research is to determine which of three techniques, namely, Scaling, the Bayesian technique, and a Covariate-based LDA, to integrate internal and external operational loss data performs best in reflecting the true operational loss distribution of a financial institution as required by banking regulators around the world. Performance is measured by the comparison of estimates of operational risk capital associated with each technique. The estimation of operational risk capital is based on a specific extreme quantile of the cumulative density function of operational losses in a given institution estimated through an LDA associated with each technique. A dynamic internal fraud model for operational losses simulates the internal and external loss data necessary to perform these estimations. The purpose of the dynamic model is to capture the nature of internal fraud and the corresponding operational controls that would mitigate or avoid the monetary losses caused by insider fraud to the retail segment of banks.

The key variable in the model is the level of operational losses that take place in a number of financial institutions that share operational loss data. Operational losses, in turn, depend on factors such as the level of operational risk controls, the propensity of workers to commit fraud, and other observable key risk indicators (e.g., the level of assets and the number of employees in each bank). In addition, operational losses are affected by macroeconomic factors such as country of location or the level of economic activity in the country of location. The financial institutions under study refer to banks that belonged to the ORX during the period 2005-2010. The banks that belong to this association are located mostly in Europe, USA, Canada, and Australia.

Significance of the Problem

Since the regulatory operational risk capital requirement first appeared in June 2004, many techniques and procedures have been developed to comply with the AMA quantitative

requisite to integrate internal and external data. The possibility of evaluating alternative integration techniques has been limited due to the lack of unique international data collection standard (BCBS, 2011). In the years that followed the publication of *Operational Risk Supervisory Guidelines for the Advanced Measurement Approaches* (BCBS, 2011), banks have been required to collect data with homogeneous criteria for a minimum of five years; it is likely that banks will revise their historic data to satisfy the new standard. A recent review of the principles for sound management of operational risk (BCBS, 2014) showed that banks have made good progress in the use of their own internal data but have not made much progress in the use of external data.

This research provides an opportunity for a timely evaluation of the different data integration techniques with the aim of providing guidance for practitioners in the industry to face the task of external-internal data integration for operational risk management purposes. The study is important because the results may help clarify the benefits and costs of three common data integration techniques for operational risk management and operational risk capital.

The results of this research may be important for risk managers of financial institutions as well as regulators. Both financial institutions and regulators are interested in the best possible quantification of risk; if the three data integration techniques affect the quantification of operational risk, it is important to know the most appropriate technique of the three in order to achieve efficiency. Failure to set a correct amount of operational risk capital is unlikely to be efficient. Financial institutions would fail to achieve the best income from their asset portfolio if too much capital has been allocated to operational risk. Conversely, they run the risk of huge losses that may jeopardize their solvency if too little capital is allocated to operational risk.

In terms of method, this study was also unique in the sense that it is the first time a comparison of different internal-external data integration technique has been performed using a rigorous, inference-based statistical procedure. As with any simulation study approach, the study required determining a data generating process. In the study, this process was achieved by means of a dynamic model for internal fraud losses. Internal fraud risk management is a field that has not been a focus of study in business management despite the strong human resource component of the risk. Therefore, the model sought to bridge the gap between the quantitative operational risk management literature and the business management literature that specializes in human resource management.

Overall, the contribution of the study was twofold. First, the study provided a framework to improve the handling of external data by financial institutions. Incremental use of external data in financial institutions has been hindered by the lack of appropriate guidance. This study therefore contributed to the solution to one of the fundamental problems in applied operational risk management. Second, the study contributed to the understanding of internal fraud risk in retail banking by laying out a model for internal fraud outbreaks. This specific aspect of operational risk has been touched only scantly in the academic literature.

Nature of the Study

In the study, a simulation-based evaluation to compare different data combination techniques was performed. Most research in this area of operational risk management is devoted to designing techniques for data combination. Examples of research on data combination techniques include Dahen and Dione (2010), Hassani and Renaudin (2013), Lambrigger, Shevchenko, and Wüthrich (2008) and Wei (2007). Whether one particular technique performs better than another has been overlooked in the literature.

The research design applied simulation steps grounded on three parts. First, a dynamic model for internal fraud losses within a set of financial institutions that share their operational

loss data from the ORX data exchange as well as other key risk indicators was calibrated. Second, the dynamic model was used to simulate internal fraud loss data and the key risk indicators and then produce datasets that, at the aggregate, mimicked the ORX dataset.

The Monte Carlo simulation exercise consisted first on selecting each bank in the ORX dataset for the period 2006-2010. Second, each bank's own simulated data were combined with its external data. Third, the three data integration techniques under inspection were applied to each bank. Each technique produced a different level of estimated operational risk capital for each of the banks. The novelty of the research design was to compare these three estimates of operational risk capital in terms of the "true" operational risk capital implied by the dynamic model in each of the banks; more specifically, a data-generating process was simulated and the properties of estimators then studied. The dynamic model acted as the data-generating process and the estimators referred to extreme quantiles of the generated internal fraud loss data. This design is common in the econometric and statistics literature (Greene, 2012, Chapter 15). Therefore, the study belongs to the simulation-study approach widely used in the econometrics and statistics literature.

The comparison of the three integration techniques in terms of the final operational risk capital estimates implied was a natural choice because operational risk capital is the final product of an AMA. For practical purposes, these estimates are of the most interest to regulators and managers of financial institutions. Other means of comparisons could have been used, such as the shape of the loss probability density functions (PDFs) obtained by each technique or the shape of the densities only at the extreme loss values. Overall, PDF comparisons are not satisfactory, however. For example, one technique may generate a density of losses very similar to the true density of the losses but only when losses are small; when losses are large enough, the fit may be poor. If the study had been used to compare the performance of the PDF for all losses small and large, it may have picked this technique as a

good technique even though the fit at the far end of the distribution was poor. For most operational loss types, what matters is the behavior of the PDF at the far right-end of the distribution. The estimated level of operational risk capital, being an extreme quantile of the cumulative density of losses, is one aspect of the shape of the distribution at the extreme right-end.

Research Questions

The focus of the study is on operational risk management. In particular, it is the first study to address the problem of choosing the best technique for internal-external data integration by financial institutions that share operational loss data. The selection problem outlined above relies on a well-defined simulation. In broad terms, any simulation steps apply the model-simulation-comparison sequence. Hence, a benchmark model is the starting point.

A model for internal fraud outbreaks was built as a starting point. The model, in order to be used for simulations that mimic real world scenarios, needed to be contrasted with reality. This reality check proves problematic in the field of operational risk management because operational loss data are proprietary. There is no public data on operational losses; thus, studies based solely on proprietary data cannot be replicated. To contrast the model with reality is to check whether, taken together, the simulated model data agree with aggregate operational losses due to internal fraud in retail banking. In addition, the model results are deemed plausible if, at the same time, the simulated model data correlate with observed global variables as reported in the literature.

This process of model building and broad reality check regarding operational losses due to internal fraud in retail banking raises important question about internal fraud processes in financial institutions. The main research questions (MRQs) can be stated as follows:

MRQ1: If the model is capable of generating internal fraud losses that are similar to ones reported in the ORX database and produce correlations with macro environmental

variables that are similar to those reported in the literature, how are these losses related to the Global Financial Crisis that occurred in the middle of the period of study?

This question has been the focus of studies about operational losses in general but not for internal fraud losses in retail banking.

MRQ2: Given the same conditions as MRQ1, how are internal fraud losses related to perceptions of corruption in the country where the main headquarters of a bank is located?

Internal fraud losses before and after the Global Financial Crisis and the correlation of those losses against corruption perception indices, have not been used in model building or the calibration of parameters. In this study, these outcomes are independent results of the model and provide valuable information about the nature of internal fraud losses in a financial institution.

MRQ3: Regarding the selection of the best internal-external data integration technique, is there any technique that can be considered best practice to estimate a correct operational risk capital across all levels of risk tolerance?

Hypotheses

The research hypothesis is stated in terms of the differences between the operational risk capital estimates obtained using the three data integration techniques and the true operational risk capital implied by the dynamic model generating the data. In the study, the simulation approach documented in Voss (2013, Chapter 3) is applied. The operational risk capital is obtained by setting a high percentile value (risk tolerance), such as the 99.9 or 99.99 percentile of the operational loss density function. These percentiles are called extreme values or value-at-risk estimators (VaR).

Let $OpRK_S$, $OpRK_B$, and $OpRK_C$ be the levels of operational risk capital estimated by Scaling, the Bayesian, and the Covariate-based LDA respectively, and let $OpRK_{true}$ be the true operational risk capital implied by the dynamic model for internal fraud. The intention is to know whether any of the objects, $OpRK_S$, $OpRK_B$, or $OpRK_C$, are systematically closer to the true capital $OpRK_{true}$, given the specific risk tolerance of the risk manager. The evaluation is performed by drawing a huge number S of simulations of complete 5-year histories of operational risk events. Each simulation $j = 1, \cdots, S$ is used to generate operational risk capital levels($OpRK_S^j, OpRK_B^j, OpRK_C^j$). For each technique, the simulation approach implies the estimation of the root mean square error (RMSE) relative to the true operational risk capital level. The lower the RMSE, the better is the technique.

The working hypothesis can be stated in the following terms:

$H_0 1$: No change is evident in the pattern of operational losses before and after the Global Financial Crisis.

$H_0 2$: Neither the frequency nor the severity of internal fraud operational losses are correlated with the corruption perception index of the country where the main headquarters of the bank is located.

$H_0 3$: None of the three techniques is systematically better as compared to the others across possible risk tolerance values.

Theoretical Framework

The general framework for the study belongs to the simulation-study approach. This method is widely used in statistical theory and applied statistics in the fields of business, engineering, and the natural and social sciences. In general, two branches of simulation methodologies are of interest in this study. The first relates to process simulation methodologies pioneered by information system engineers (Banks, 1998; Zeigler, Praehofer,

& Kim; 1976). The second relates to simulation-based evaluation of estimators used by theoretical and applied statisticians as well as econometricians (Voss, 2013).

Process simulation methods, favored mostly by information system engineers, are used to imitate the operation of a real-world process or system over time to generate artificial data for decision-making and management purposes (Banks, 1998). For example, in software engineering, many techniques exist to predict the features of software projects like duration or effort. It is of interest to identify the most accurate prediction techniques, and a usual avenue is to use a simulation-based methodology, as did Shepperd and Kadoda (2001) as well as Aranha and Borba (2008). On the other hand, simulation methods are used to infer the properties of estimators or statistics that depend on the data (Greene, 2012; Stern, 2000; Voss, 2013). Statisticians and econometricians mostly use this aspect of simulation both at the theoretical and empirical level. Studies that use this method are called simulation studies or Monte Carlo studies.

The study conducted draws insights from both areas of the simulation literature. First, it set up and simulate a dynamic operational risk model capable of drawing internal fraud operational losses in the retail-banking segment of individual banks that share data through the ORX. The dynamic operational risk model shares some features of those described in Supatgiat, Kenyon, and Heusler (2006), Leippold and Vanini (2005), and Bardoscia and Bellotti (2011). Operational losses within financial institutions can be classified according to event type and business lines within a financial institution. Annex 8 and 9 of the BCBS (2006) provides the standard in the classification of business lines and operational loss event types. A summary of the classifications is found in Tables 1 and 2. There are seven types of operational loss events. All these events can potentially occur within eight business lines in a typical banking institution.

Table 1

BCBS Loss Event Type Classification

Event number	Definition
1	Internal fraud
2	External fraud
3	Employment practices & workplace safety
4	Clients, products & business practices
5	Damage to physical assets
6	Business disruption and system failures
7	Execution, delivery and process management

Table 2

BCBS Business Line Classification

Event number	Definition
1	Corporate finance
2	Trading and sales
3	Retail banking
4	Commercial banking
5	Payment and settlement
6	Agency services
7	Asset management
8	Retail brokerage

A particular combination of event type and business line is called a cell. The research was focused on just one cell: Operational losses due to internal fraud (event type) in retail banking (business line). *Internal fraud* is defined as operational losses due to acts that involve at least one internal party aimed at defrauding, misappropriating property, or circumventing regulations, the law, or company policies (BCBS, 2006). As Kochan (2013) suggests, internal fraud is one of the fastest-growing and most complex criminal threats to financial organizations. This type of threat from insiders takes various forms because fraud can occur at any level of the administrative ladder, from junior employees up to chief

executives. This specific operational event calls for a specific modeling setup that incorporates, for example, factors that shape the incentives of insiders to engage in fraud, like worker compensation, culture, or macroeconomic conditions (see also Jarrow, 2008). Therefore, the model developed departed from the traditional simulation models used by engineers in that human resource processes are modeled instead of machines or information systems processes. The model applied borrows insights from the human resource and organizational literature, specifically from people risk management theory as shown, for example, in Blacker and McConnell (2015).

Retail banking is a traditional, universal type of banking involving payment services (debit cards), short-term unsecured loans (credit cards), money management facilities (current accounts), savings, loans, and mortgages (Pond, 2014). Retail banking provides services to the public or "retail customers." According to ORX (2012), retail banking experiences the larger number of operational loss events, 59% for the period 2006-2010, and increasing to 65% in 2011. In addition, the gross losses in retail banking are the most severe of losses across business lines, representing up to 37% of total losses by business line.

Once the model was set up to simulate operational loss data in the internal fraud-retail banking cell for a window of five years for each bank belonging to the ORX Association, the data were ready to be recorded by each participating bank. However, not all the operational loss data had been recorded due to collection standards and threshold rules of data exchange associations or due to measurement errors and delays in reporting protracted events (Chaudhury, 2010). This reflects the information restrictions that characterize data recording and that make real loss data different from collected data. The collected data in each bank were delivered to the shared database. Therefore, after a simulation, each of the banks not only has access to its own five-year recorded operational loss dataset but can also use the entire ORX dataset. Due to heterogeneous nature of the financial institutions in the dataset,

however, the external data to a bank cannot be used as it appears. For example, a bank may be smaller, bigger, or in general, more or less risky than another bank. As such, the loss amounts or the frequency of losses may be different across banks (Shevchenko, 2011).

Hence, like in the real world, external data had to be integrated properly into the internal data to be used in the AMA for operational risk capital estimation. The focus of the study was on three techniques that can be used to integrate external data into the internal data for a specific bank: Scaling, the Bayesian technique, and a Covariate-based LDA. The application of each technique assumes, calibrates, or estimates parameters and values so that each technique is conditional to its own set of assumptions. All three techniques draw total operational loss forecasts for the year ahead in the specific internal fraud and retail-banking cell for a bank under study. The operational risk management problem of the bank is the identification of the level of losses deemed "catastrophic" in terms of a bank's normal business operations.

In contrast to what is common in the forecasting literature, the emphasis was not on the mean one-year ahead operational loss forecast but on the entire probability density of future outcomes. In the standard forecasting literature, the probability of future outcomes usually shows a normal distribution, and in that case, only the mean and standard deviations are of interest for practical purposes.

In the financial risk literature, the object of interest is the tail-risk, namely, the possibility that the outcome exceeds a given extreme level. Furthermore, the shapes of probability densities of financial outcomes are not symmetric but tend to be strongly asymmetric and feature higher probabilities at the extremes of the distribution than does the normal distribution. In the particular case of operational risk, the LDA permits drawing data from the probability density of operational losses and therefore estimates the VaR, which is the most extensively used tail-risk measure in operational risk. Each of the three techniques

that integrate internal and external data delivers a VaR estimate that could be compared. The regulatory capital charge for operational risk is given by the difference between a given extreme percentile and the mean expected loss.

With respect to the techniques evaluated, the Scaling technique is used to transform the external data to make it comparable to the internal data (Chaudhury, 2010; Cope & Labbi, 2008; Na, van den Berg, Couto, & Leipoldt, 2006) based on observable factors pertaining to external companies. Once external data were transformed and made comparable to the internal data, the pooled data could be used to apply the LDA approach. Bayesian techniques combine the internal loss distribution information and external loss distribution information by means of the Bayesian theorem (Lambrigger et al., 2007; Shevchenko, 2011). The resulting loss distribution was used to extract the operational risk capital estimation. A Covariate-based LDA estimated a parametric loss distribution for the entire data set, thus including internal and external data, but conditional on different factors across the financial institutions covered in the database. Conditioning factors or causal dependence models have been depicted, for example, in Cruz (2002), Supatgiat et al. (2006), Kühn and Neu (2003, 2004), Bardoscia and Bellotti (2011), and Leippold and Vanini (2005).

The three techniques are used by individual banks because their own internal data, covering a window of five years, are insufficient for VaR estimations. Ideally, a bank would want to replicate current five-year windows as many times as it could in order to have enough data to estimate tail-risk measures accurately, but that is impossible with real-time series data. However, the dynamic internal fraud loss model that draws the five-year window of data can simulate as many 5-year windows as necessary to replicate the current business conditions and therefore can mimic the true VaR specific to the conditions of the bank under study.

With VaR estimators corresponding to each technique under evaluation and the knowledge of the true VaR, the research implemented standard procedures to determine

whether any technique performed better in terms of the root mean square error distance from the true VaR. This method is akin to applying statistical procedures to know which possible estimators of a given parameter are closer to the true parameter, where the modeler sets the true parameter through simulation assumptions (Matzkin, 2003; Chen & Pouzo, 2012; Greene, 2012). As in any simulation, the parameter under investigation was a function of the data, and the modeler's choice of a true model generated the data.

Definition of Terms

Below some key terms are defined. Throughout this study, a number of abbreviations were also used. Appendix A describes all the abbreviations used in this study.

Basel Committee on Banking Supervision (BCBS). The BCBS is a forum that sets the guidelines for worldwide regulation of banks. It agrees on international regulations for the conduct of the banking industry (Goodhart, 2011).

Basel II is the set of standards proposed by the BCBS that lead to the capital adequacy of international active banks; the set of standards was published in June 2004 (BCBS, 2006).

Bayesian inference is a statistical technique used to incorporate expert opinions into data analysis and combine different data sources (Shevchenko, 2011).

Monte Carlo simulation involves the simulation of random samples from a known population (an explicit parametric model) to track the behavior of a parameter that depends on the data (Robert & Casella, 2005).

Probability density function (PDF) is a function that describes the relative likelihood of the occurrence of a random variable at a given point (Greene, 2012).

Value at risk (VaR) is an extreme quantile of a distribution of financial outcomes. It usually refers to the 99.9% or the 99.99% quantile (McNeil, Frey, & Embrechts, 2005).

Assumptions

The dynamic model for operational risk is simple but strong enough to model the broad nature of internal fraud operational losses and their management. The model is general enough to explain the time series dynamics of loss occurrence in each financial institution under study, and therefore, it can be used to generate loss event simulations due to internal fraud in retail banking.

The LDA assumes a set of parametric density functions for measuring the amount of loss and a set of parametric distribution functions for the frequency of losses. The choice of function depends on the risk type under consideration. In the research conducted, the standard functions used in the literature are assumed.

For the calibration of parameters, a number of observable variables are used. These variables reflect each bank idiosyncratic features. Some of them reflect the scale of banks such as the level of retail assets or the number of employees in retail banking operations. All these scale figures are taken from bank's annual reports. Most banks follow generally accepted accounting principles (GAAP). Therefore, the research assumed that all the information from bank's annual reports reflect the true scale and risk exposure of banks and hence, they can be used to discriminate among banks.

Another set of idiosyncratic variables refer to textual analysis also obtained from banks' annual reports. These textual variables reflect the number of times a word or combination of words appear in the annual reports as proportion of the total page numbers. The chosen words reflect how concern, banks are about human resources and risk in their corporate environment. The key assumption therefore is that the more concern is a bank about risk or human resources, the more they write about risk or human resources.

Limitations

A drawback of the research design is that a model is never as complex as reality. It is inevitable that there are aspects of reality that will be excluded from a model. Given that the purpose of the study was to compare data integration techniques, it is sufficient to know that the absent features of the internal fraud model were not correlated with the features of the techniques. Furthermore, the evaluation of models or statistical estimators using the Monte Carlo simulation of an assumed data generating process is an acceptable and standard practice in academia (see, for example, McNeil et al., 2005). Nevertheless, a limitation of any simulation study is that the comparisons are specific to model parameterizations. To achieve generality it would be necessary to vary the parameters in such a way that all the possible cases are taken into account. This is practically impossible in the model set up presented in the research because there are 52 internal fraud models with their own set of parameters. Taken together there are more than two hundred parameters.

This limitation of simulation studies is however overcome by the fact that the simulations consider many different banks or many different cases. The different sets of parameters reflect different types of risk exposures and idiosyncratic characteristics.

On the same vein, estimations of operational risk capital by applying the LDA approach also entails different possible choices of distribution functions either for the severity of operational losses or for the number of loss events in a given year. A general simulation-based framework would entail the combination of all possible distribution functions. However, the space of these distribution functions is large and therefore it is not possible to apply the approach to all the cases.

Model simulations are validated by contrasting the simulated internal fraud losses with macroeconomic variables such as GDP growth or other country level variables. These country-level variables are assigned to banks only considering the country where bank's

headquarters are located. A specific bank can have branches in a variety of countries and a segment of these branches can perform retail banking so it might be possible to consider all the countries where these banks operate. However, the dataset used in the research does not have detailed bank data by country.

Delimitations

The internal fraud model developed did not capture specific details of the complex operational risk environment that surrounds internal fraud events. The usefulness of the model relies on it serving as a simulation platform to allow practical comparisons among competing data integration techniques. Moreover, the research only assessed the technical aspects of the techniques. Other aspects, such as the different costs associated with the implementation of each technique, were not the focus of the research.

The results are specific to banks that belonged to the ORX within the 2006-2010 window. They might not reflect the risk or loss behavior of ORX banks afterwards. This is so because the period under study in this research includes first a mayor global financial crisis that affected global banks in particular and second a corresponding global recession. The macro environment and the regulatory environment that banks face for the period after 2010 is arguable different.

In essence, the research method applied in the study is descriptive because it seeks to answer which internal-external data integration technique is best but it does not delve onto how or why. The explanatory research is the natural next step in this research agenda.

Summary

In this study, the aim was to contribute to the solution to one of the most challenging problems in the practical implementation of operational risk management: The selection of the appropriate internal-external data integration technique among a set of alternative

techniques laid out in the literature. For financial institutions, the data integration problem has hindered progress in the use of external data for assessing operational risk management.

A novel procedure for the selection of the best data integration technique was used in this study. The procedure was based on a simulation study, which is a powerful method used in theoretical and applied statistics to study properties of objects that depend on data. Statistical comparisons about which of the data integration techniques produced estimators of operational risk capital closer to the true operational risk capital are made.

A key step in the application of the simulation was the construction of a dynamic model of operational losses resulting from internal fraud in retail banking in a number of financial institutions. A set of parametric equations, based on the operational risk and human resource management literature, mapped conditioning factors pertaining to financial institutions to outbreaks of operational losses. The parameters of the dynamic model could not be estimated because detailed operational loss data for banks are not publicly available due to the sensitive nature of the data. Instead, the model was calibrated in such a way that simulated frequencies of losses and aggregate loss amounts were similar to those published in the ORX report. In addition, the calibration considered that the correlations of the simulated losses with aggregate macro environmental variables were the same as the correlations reported in the literature.

The model developed in the study is the first that directly links human resource factors with operational losses due to internal fraud. The existing causal models in the operational risk literature are focused more on losses due to failures in information technology processes, and scant attention has been paid to the human aspect of operational risk. The model allowed uncovering two facts that have not been reported in the operational risk literature: The relationship between overall economic activity and operational losses due to fraud and the relationship between the corruption perceptions of countries and operational losses.

In summary, two important contributions to the risk management literature were made in the study. First, a rigorous statistical method was used to help select the best data integration technique and thus contribute to solving one of the key challenges that have deterred practical progress in external data handling by financial institutions. The second contribution was the development of a causal model of operational losses due to internal fraud in retail banking. The model developed shed new light on the nature of internal fraud losses in the context of operational risk management.

In Chapter 2, the details of the theoretical and empirical underpinnings laid out in the literature about various aspects of the research are provided. The dynamic operational loss model is discussed in dialogue with the literature on dynamic systems. The literature about data integration techniques is covered and comparisons are made to published research in the same field.

Chapter 2: Literature Review

Despite the financial importance of operational risk documented since the onset of the 21st century, research on operational risk has not matured at the same pace as credit or market risk literature (Chaudhury, 2010; Medova & Berg-Yuen, 2009). To date, the operational risk literature has been silent about the evaluation of alternative data integration techniques within the AMA framework. Two factors may have hindered research on this topic. First, data for empirical research purposes are restricted; the focus of empirical research has been on robust estimation methods, data combination techniques, or other issues in the implementation of AMA. This is the case, for example, in Aue and Kalkbrener (2006), Guillen, Gustafsson, and Nielsen (2008), Wahlström (2013), and Guegan and Hassani (2013) among others. The second limitation is that the practical implementation of AMA in financial institutions has been focused on the details of implementing existing AMA approaches such as loss data classification or the specific parametric functional forms for the frequency and severity in each operational loss cell. This was the case, for example, in Aue and Kalkbrener (2006). Little theoretical or applied literature within financial institutions has been focused on the evaluation or comparison of distinct data integration techniques within AMA.

Literature about the comparison of known approaches to calculate operational risk capital is scant but relevant to the research. Only two papers were located in a broad search of the literature, namely, Teker (2005) and Jiménez et al. (2009). In both studies, the operational risk capital delivered by applying the basic indicator approach (BIA), the standardized method (SM), and specific AMA approaches were compared. Teker (2005) performed the comparison with data for a Turkish bank while Jiménez et al. (2009) used data from a Spanish bank. The conclusions for both studies were that the application of the AMA delivers much lower operational risk capital levels than those obtained by applying BIA and SM. The results, as explained in both papers, reflects the risk-sensitive nature of the AMA approach.

Two banks equal in size, measured by assets, for example, are deemed to have similar operational risk capital levels when measured with BIA or SM. If one of the banks is more exposed to operational risk or practices poor operational risk controls, however, it will have higher measured operational losses, and hence, it will endure a higher operational risk capital level as measured by the AMA. The evaluation exercise applied in this study goes beyond these earlier comparisons in two aspects. First, instead of comparing the AMA with non-advanced approaches, distinct data integration techniques and their resulting operational risk capital levels under the AMA were compared. Second, the comparison was made through statistically robust tools; thus, the conclusions reached have broader validity.

In this chapter, the research design steps are developed. First, issues related to the nature of operational loss data are reviewed. Second, a brief review of the literature on the modeling of operational losses is presented. Third, the current literature about how to integrate internal and external data in operational risk modeling is surveyed. Last, a summary and conclusions that arise from the literature are provided.

Data in Operational Risk Modeling

The use of external data is an essential step when applying the AMA to a firm, but direct use of external data is ill advised. External loss data comes from individual banks; each bank is likely to have its particular risk profile and characteristics, such as the size of its revenues, number of staff, quality of staff, and level of control systems once an event occurs (Chernobai, Jorion, & Yu, 2011). Moreover, each line of business of a bank is expected to be different in terms of risk profile and internal characteristics.

As described by Baud, Frachot, and Roncalli (2002), Dahen and Dionne (2010), and Wei (2007), a number of drawbacks and biases in the use of external operational loss data can be listed. For instance, selection bias occurs because only very large losses are published. Control bias happens due to losses coming from heterogeneous banks with different control

environments. Reporting bias appears when data are drawn from different sources with variations in recording thresholds. Scale bias arises when losses come from banks of diverse scale (for example, assets, revenues, number of employees). Last, survival bias ensues because the losses of bankrupt firms are not recorded. In order to overcome these biases to some degree, it is necessary to rely on techniques that allow a financial institution to use available external data and combine them with its own data in order to perform a standard AMA.

The evaluation of data integration techniques needs, in theory, a complete dataset of internal and external data for a firm. Banks that perform the AMA generally share data through data exchange associations such as the ORX, Global Operational Loss Database (GOLD), or Operational Risk Consortium (ORIC), among others, but these datasets are not publicly available. To overcome this data limitation problem, a dynamic model for internal fraud event occurrences was used to simulate loss events that, in aggregate, look like real datasets.

In this research, the ORX dataset was used as the basis for model calibration. In the next section, the relevant literature that supports the choice of a simulation model to perform the comparison exercise of data aggregation techniques is reviewed.

Simulations of Operational Loss Events

Researchers in many areas of scientific knowledge rely on simulations when data are unobserved and a need to learn about the behavior of a complex system exists (Banks, 1998). This is the case, for example, in the ORX database; the detailed data are proprietary, but data summaries are published regularly. In addition, other specific information about banks that participate in the ORX consortium is available from different outlets, for example, financial service authorities and their corporate Web pages. The question is how to simulate a model that can generate operational loss events that share the same statistical properties of the

published summaries of the reported data. A number of research papers feature models that simulate operational losses within firms in terms of their specific characteristics and industry-wide features. This strand of the literature borrows heavily from both the business process simulation (BPS) and statistical mechanics literature. The unifying feature between these seemingly different fields is the use of network theory.

The unifying characteristics of the papers reviewed are the interconnectedness among risk types, business lines, and external risk indicators. The features arising from networks are not crucial for the current study. The focus of the research is on one cell in the event type and business line matrix, namely internal fraud in retail banking, so all notions of interconnectedness among the other networks can be excluded. The theories on which network models are founded and their functional forms, however, provide rich alternatives for the model.

Shepperd and Kadoda (2001) offered a first idea to motivate the research. Shepperd and Kadoda compared software prediction systems by using artificially generated data with known properties to explore software engineering dataset modeling techniques. Shepperd and Kadoda suggested the simulation of artificial data serves to provide researchers with the following:

> A great deal more control over the characteristics of a data set. In particular, it enables the researcher to vary one property at a time, thereby allowing a more systematic exploration of the relationship between data set characteristics, type of prediction system, and accuracy. By contrast, especially with smaller real data sets, the true properties may not be fully known. (p. 1015)

The research conducted followed the example of Shepperd and Kadoda (2001) but in an entirely different field: Operational loss events. The simulation-based approach stems from system dynamics simulation (SDS), a field developed by the computer engineer Jay

Forrester during the 1950s at Sloan School of Management at MIT. Primarily computer and engineering sciences use SDS, but it has also become important in business simulation. Kessler (2007), for instance, applied the approach to build a framework for operational risk management where banks behave as a dynamic system within interacting and complex domains.

In general, SDS is parallel to business process management (BPM) in the field of management. Ideas within the field of BPM are also useful in the development of the simulation modeled applied. Although the aim of BPM is to support the design, enactment, control, and analysis of business processes, it could also serve as a platform for operational risk management as part of the business process itself. An important element of BPM is BPS, which relies on discrete event simulation models. In the research conducted, it is precisely this type of discrete event model that was used. Jansen-Vullers and Netjes (2006) provided a general overview of the use of BPS. A direct application to operational risk modeling is to find cause-to-effect relationships (Supatgiat et al., 2006). The idea of the approach is to provide the risk manager with a tool to reduce and control operational risk.

Under the BPM approach, Cheng et al. (2007) proposed an approach to operational risk modeling based on the automatic development of a probabilistic network that mimics closely and in real time the operational business processes. This means that the parameters of the probabilistic network can vary in time as business processes adapt to new situations. Cheng et al. focused primarily on operational risk related to information technology. Aleksy, Seedorf, and Cuske (2008) tackled one important aspect of the BPS, the link between domain knowledge (what practical end-users need and know) and software development. Aleksy et al. proposed an approach for modeling that monitored and controlled operational risk in financial institutions based on an approach called JOntoRisk created by Cuske, Dickopp, and Seedorf (2005). Weiß and Winkelmann (2011) took a similar stance. Using the same BPM

approach, Cernauskas and Tarantino (2009) proposed linking BPM with engineering process control or statistical process control to perform the management of operational risk where both automated and personnel processes appear. The above models applied the BPM approach and focused on operational risk management within a business process. These models do not explicitly aim at generating operational loss statistics.

The SDS approach in engineering is akin to causal models, Bayesian models, and reliability theory in the statistical science. These approaches are what Finke, Singh, and Rachev (2010) called process-based models. Process-based models are related to functional dependence and functional correlation in statistical physics. Kühn and Neu (2003, 2004) used this approach. Kühn and Neu studied models that generate operational losses in banks through network dynamics that lead to the occurrence of risk events in an environment where banks make efforts to mitigate operational losses. Leippold and Vanini (2005) used functional dependence modeling to extend the work of Kühn and Neu (2003, 2004) in two dimensions. First, Leippold and Vanini (2005) explicitly used a networked process through graphs. Second, Leippold and Vanini included fixed and stochastic costs that arise in case of operational risk events.

Based on the aforementioned strand of the literature, Bardoscia and Bellotti (2011) modeled the amount of operational losses recorded at a certain time in a certain process. The approach of Bardoscia and Bellotti was an effort to model some general mechanisms behind the generation of operational losses. This set of models in the statistical mechanics tradition focuses more explicitly on the generation of operational losses; consequently, they proved useful for developing the methodology for the research strategy next chapter.

The research conducted took into account ideas from Fragnière, Gondzio, and Yang (2010) and Yang (2010) who claimed that the treatment of operational risks must also follow a managerial approach whereby the quality and quantity of the workforce represent a source

of risk. This idea is similar to that of Hatzakis, Nair, and Pinedo (2010). The importance of human capital in the operational loss process of financial institutions accords with the idea that the key process of a bank is the handling of information; banking is known to be a knowledge-intensive business process (Weiß & Winkelmann, 2011). This calls for a modeling approach that takes the quantity and quality of employees into account. This is for example the case of Blacker and McConnell (2015).

Therefore, in contrast to the simulation approaches applied in engineering and computing systems or the bulk of BPS applied for modeling operational loss events related to failure of processes and machines, this research was focused on simulations applied to human behavior and operational risk controls that generate internal fraud events. The idea is similar to the managerial approach in Fragnière et al. (2010) and Yang (2010); however, these two studies are focused on the optimal planning of workforce capacity and do not touch on factors that generate operational losses due to internal fraud.

The originality of the approach applied is that the study conducted relied on the specific modeling of factors that bring about internal fraud events, such as the ethical quality of workers, the workplace environment, and other elements pertaining to human resource management.

Quantitative Techniques to Combine Internal and External Data Sources

Due to the potential biases in the use of external data, sound techniques are required in order to integrate external and internal data. The aim of the techniques is to allow a specific financial intermediary the use of the combined datasets to implement the AMA for operational risk capital. As internal and external data collection have become more widespread over the course of a decade, and as the AMA is ever more used across banks, a number of techniques have been presented to the operational risk community. This literature review is focused on three broad techniques to be further expounded in Chapter 3: (a)

techniques based on scaling losses to make the data comparable, (b) techniques based on the Bayesian inference, and (c) techniques based on a covariate-based LDA.

Scaling technique research. With the onset of the last decade, financial firms started gathering operational loss data. In this environment, earlier research about the use of internal and external data maintained the assumption of data homogeneity across internal and external data. This is the case for Baud et al. (2002, 2003) as well as Frachot and Roncalli (2002). In these three studies, the reporting bias was considered the most important issue at the time of combining the data, and therefore, the authors concentrated on dealing with that bias. For example, Baud et al. (2003) applied a fair mixing assumption by which external data were treated as homogeneous. A model for the data generating process focused on incorporating external data was developed. In the model, the bias came simply from the fact that external data were truncated above a specific threshold. Frachot and Roncalli (2002) augmented Baud et al. (2003) in the sense that the internal and external data combination was used both for the estimation of the severity of losses and for the estimation of the frequency of losses.

The idea of scaling external data to make it comparable to internal data first appeared in Shih, Samad-Khan, and Medapa (2000). Once transformed and made comparable to the internal data, external data can be pooled and the LDA approach applied. The technique relies on the existence of a power law relationship between losses incurred in business units and their gross revenue. Shih et al. showed that the size of a firm is related to the magnitude of its operational losses; however, this relationship was not linear but logarithmic, and there is evidence of a decreasing relationship between firm size and observed operational loss severity.

Shih et al. (2000) split the nature of operational risk into two components: Global and idiosyncratic. The idiosyncratic component is supposed to relate to bank and business line-

specific factors such as the size and volume of business, and their effect might be modeled using a specific power law that takes the form of

$$L = R^{\alpha} \times F(\theta) \tag{1}$$

where L is the loss amount; R represents the total income of the firm where the loss took place; $\boldsymbol{\alpha}$ is a scaling factor (to be estimated), and $\boldsymbol{\theta}$ is a vector representing all the risk factors not explained by R. $F(\boldsymbol{\theta})$ is a multiplicative residual term that is not explained by changes in revenue (firm size). Shi et al. (2000) reported heteroscedasticity, the presence of which reduces the relationship between firm size and loss severities. Na et al. (2006) also analyzed the power law relationship between losses and business revenue. Na et al. used internal data from the ABN AMRO Bank and external data from ORX to extend Baud et al.'s (2002) work. Na et al. (2006) applied direct scaling of variables according to a power law regression for the aggregation of data originating from different sources. Dahen and Dionne (2010) used the Fitch's OpVar database to extend the power law regressions by incorporating geographical differences and specific lines of business. Dahen and Dionne's model allowed for comparing internal data within a firm in different lines of business.

Cope and Labbi (2008) provided a further development within the branch of scaling techniques for operational loss data. Cope and Labbi used ORX data and applied the scaling techniques introduced in Shih et al. (2000), Na et al. (2006), and Dahen and Dionne (2010) to model not the mean losses but the different quantiles of losses using quantile regressions. Cope and Labbi (2008) concluded that frequently large losses scale differently to small losses. Because of the regulatory focus on large losses, it is essential that scaling relations for extreme events be appropriately characterized.

Overall, the scaling technique assumes that a relationship between the scale of a bank and the severity of operational losses exists. This relationship may not be apparent due to the presence of heteroscedasticity and must be extracted with techniques that control for

heterogeneity across banks; one such technique is quantile regression. In some applications, such as in Aue and Kalkbrener (2006) and Wahlström (2013), this type of scaling is not relevant, but Wei (2007), Cope and Labbi (2008), and Ganegoda and Evans (2013) posited that scaling, if properly identified, is important.

Bayesian techniques research. Bayesian techniques combine internal and external data as well as expert opinions by means of the Bayesian theorem. Besides a brief mention by Cruz (2002), practitioners and researchers seldom use the Bayesian technique to combine expert opinions to assess the severity and frequency distribution for operational risk.

Lambrigger et al. (2007), Peters and Sisson (2006), Shevchenko (2011), and Shevchenko and Peters (2013) documented the Bayesian technique. This technique relies on estimating posterior predictive density functions for losses. The combination of prior information and data summarized in a likelihood function allows posterior predictive density functions to be calculated. It is always the case in Bayesian estimation that the posterior does not have a known form; therefore, it is necessary to use sampling algorithms to sample loss data from the implied posterior.

For data combination, Bühlmann, Shevchenko, and Wüthrich (2007) provided a method that relies on a full hierarchical credibility theory approach to estimate frequency and severity distributions of operational losses by taking into account internal data, expert opinions, and external data. However, the model can be too sensitive to the expert opinions used to estimate scaling factors for distribution parameters. To improve upon this feature, Lambrigger et al. (2007, 2008) extended the approach developed in Bühlmann et al. (2007) to provide a more robust inference to expert opinions. Shevchenko and Wüthrich (2006) presented further examples of the Bayesian inference technique for operational risk quantification.

In the literature on Bayesian techniques to perform LDA, Bühlmann and Gisler (2005) and Bolancé, Guillén, Gustafsson, and Nielsen (2012) considered the input of expert opinions to elicit Bayesian priors or to calibrate credibility models. This is also the case for Agostini, Talamo, and Vecchione (2010). Agostini et al. set up a model that integrates the operational VaR obtained from historical data with the VaR drawn from expert estimations. Agostini et al. performed the integration by using credibility theory. Along the same lines, Figini, Gao, and Giudici (2013) proposed using self-risk assessment questionnaires to elicit suitable priors for the parameters that govern the distribution of loss frequencies and the density of loss severities. They suggested that once prior distributions are combined with the density of the data, it is straightforward to perform predictive densities of frequency and severities that allow the LDA to be performed. Agostini et al. (2010) and Figini et al. (2013), however, did not integrate internal and external data in their research.

Two recent papers combine the three sources of information, namely, internal data, external data, and scenario analysis, which is akin to expert opinion. Hassani and Renaudin (2013) proposed a Bayesian cascade methodology that works in two steps; in the first step, scenario analysis serves to elicit prior distributions and external data inform the likelihood component of the posterior function. In the second step, the posterior thus obtained plays the role of prior distribution while the internal loss data inform the likelihood component of the final posterior. The latter posterior allows for generating the predictive density of the severity of losses to apply the LDA.

Ergashev, Mittnik, and Sekeris (2013) proposed a Bayesian estimation method for loss severities using the generalized Pareto distribution common in extreme value theory (EVT) to tackle the integration of alternative sources of information such as scenario analysis and external data. Ergashev et al. focused on the extreme values of loss severities because the shape of the distribution on the right-end extreme drives operational risk capital. In the

elaboration of the Bayesian technique in Chapter 3, the research applied the example of Lambrigger et al. (2007, 2008).

Covariate-based technique research. In general, the covariate-based technique deals with the incorporation of covariates in the estimation of parameters of the loss frequency and severity distributions. Paredes (2006) offered an early attempt to introduce covariates in the estimation of the parameters that govern the frequency and density of losses. Within the LDA, Paredes showed that the data determine the shape of the loss density and frequency distributions. Research that is more recent showed the importance of covariates to shape the frequency and severity distributions. Chernobai et al. (2011) used the conditional Poisson regression applied by Paredes (2006) to model the parameter that drives the frequency of losses. Chernobai et al. (2011) found that internal control improvements and management oversight mitigate loss event frequencies. Chernobai et al. also found that macroeconomic factors are not that important. Hemrit and Ben Arab (2012) used the same approach as Chernobai et al. (2011) to identify the determinants of the frequency and the severity of losses in the Tunisian insurance industry. One key result of that research is that the frequency of losses increases with the number of employees, but none of these studies integrated internal and external data.

Wei (2007) first explored the data integration idea by using credibility theory for the combination of internal and external data about the frequency of losses and proposed a method to consider heterogeneity in the treatment of severity data. The idea was to incorporate covariates to pin down mean severity density parameters that describe the mean of operational loss data. Following the example of Shih et al. (2000) and Na et al. (2002), Wei (2007) assumed power law forms. The use of covariates to shape density functions can be traced back to Smith and Shively (1995) and Rootzen and Tajvidi (1997).

Ganegoda and Evans (2013) modeled the severity of operational losses by using generalized additive models for location, scale, and shape (GAMLSS) developed by Stasinopoulos and Rigby (2007). Like Ergashev et al. (2013), Ganegoda and Evans (2013) placed emphasis on the scaling properties of the tail of the loss distribution. In the implementation of the covariate-based LDA, the study followed the example of Ganegoda and Evans.

Summary

The review of the literature was focused on the techniques and methodologies that might prove useful for answering the research question. The first part dealt with the literature about simulation-based tools that would allow operational loss datasets to be generated across financial institutions. The simulations to be performed in the research applied ideas to model operational loss events from Kühn and Neu (2003, 2004), Leippold and Vanini (2005), and Bardoscia and Bellotti (2011) and the ideas of incorporating human factors from Fragnière et al. (2010) and Yang (2010).

The model developed was used to evaluate different data integration techniques. For the most part, the literature has been silent about evaluating the worth of data integration techniques that have appeared in the operational risk literature. This research is the first documented attempt to provide such an evaluation using statistical tools. Achieving the aim involved followed a simulation study approach that allowed robust testing of the data integration techniques as suggested by Voss (2013) and Greene (2012).

Based on the review of the integration techniques, Chapter 3 includes the work of Shih et al. (2000) and Na et al. (2006) for setting up the scaling method. It will also incorporate the contribution of Lambrigger et al. (2007, 2008) for implementing the Bayesian data integration technique and apply Ganegoda and Evans (2013) for implementing the covariate-based data integration.

Conclusion

The literature reviewed in this chapter has shown that this research constitutes the first documented attempt to model internal fraud losses in a dynamic environment with a direct application to operational risk management. There are a number of models used for simulating operational risk that are focused on internal business processes and systems. These models draw from the literature on systems dynamics pioneered at the Sloan School of Management at MIT during the 1950s. In this research, the aim was to fill this gap in the literature by building a dynamic internal fraud model based on the systems dynamics tradition and incorporate distinct aspects of human resource behavior and management.

In this research, an internal fraud model is applied in a simulation to assess the merits of different data integration techniques. The operational risk literature shows various procedures for integrating internal and external operational risk data in financial institutions, but there is an absence of studies to determine which techniques perform best given a specified criterion. This research is therefore the first documented attempt to perform an evaluation of different data integration techniques.

Chapter 3: Methodology

The aim of this research was to develop a model to be used as an internal fraud loss generating process in order to perform a simulation study. The simulations allow to generate the true operational risk capital level as well as three capital levels associated to three internal-external data integration techniques. These three operational risk capital levels are compared to the true capital level. The methodology involved two main parts; first, build an internal fraud model to describe the occurrence of internal fraud events within financial institutions and second, determine which of three existing techniques to integrate internal and external operational loss data performs best in reflecting the true operational loss capital level as required by regulators.

The three data integration techniques under scrutiny within the LDA were Scaling, the Bayesian technique, and a Covariate-based LDA. These techniques deliver different VaR measures as operational risk capital indicators. The assessment relied on the comparison of each VaR measure against the VaR implied by the true data generation process.

The focus of this chapter is on the details of the research design required to carry out the study. In addition, the appropriateness of the design is described and the research questions and hypothesis stated.

Research Design

The research design has three key steps: (a) Data simulation through an internal fraud model, (b) technique implementation, and (c) statistical comparison of the results. A dynamic model for operational loss events simulates the internal fraud loss data as well as the key risk indicators and other factors associated with the loss event. The model calibration aims to mimic the first moments of total severity and the frequency of operational losses observed in the ORX database. The dynamic model draws on Kühn and Neu (2003, 2004), Leippold and Vanini (2005), Bardoscia and Bellotti (2011), Fragnière et al. (2010), and Yang (2010).

The step to implement the data integration techniques used the simulated loss data and loss event-associated risk indicators as inputs into the specific techniques. The Scaling technique implementation is based on the work of Shih et al. (2000) and Na et al. (2006). In turn, Lambrigger et al. (2007, 2008) informed the application of the Bayesian technique and Ganegoda and Evans (2013) the Covariate-based technique. The evaluation or comparison step follows the simulation-based procedure described in Green (2012) and Voss (2013).

The idea hinges on comparing the three data integration techniques by applying them to simulated data obtained with the dynamic internal fraud model of operational losses. The dynamic model is in fact a true data-generating process. The three techniques deliver statistical objects that depend on the data. These objects are the VaR loss levels, which are considered operational risk capital estimators.

In standard applications of simulation-based comparisons of estimators, the estimators are relatively direct functions of the data. In the operational risk context described in this study, the estimators refer to extreme quantiles, for example the 99.9 percentile, conditional on each technique. The estimators are compared to each other by means of the root mean square error relative to the true operational risk capital.

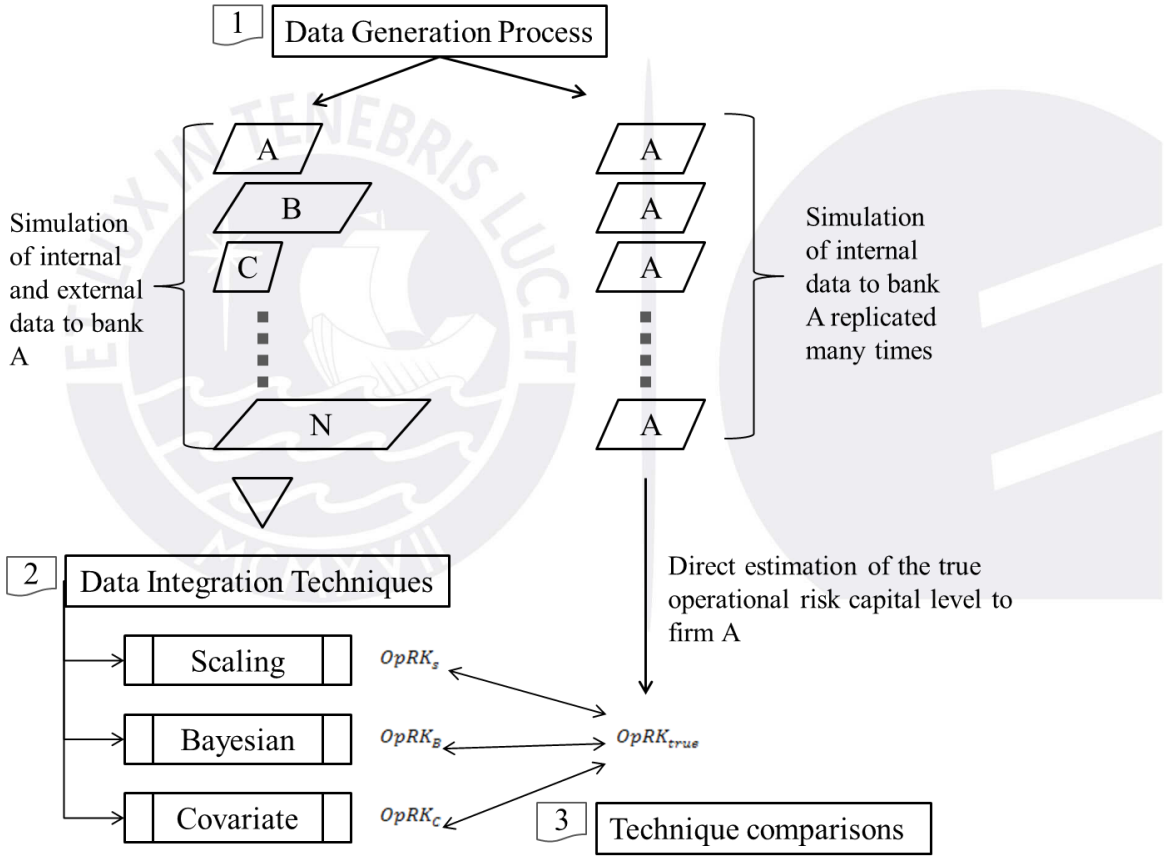
Elaboration of the Research Design

Figure 1 depicts the three steps carried out in order to accomplish the objective of the simulation study. The steps comprised the following:

1. Generation of data
2. Application of data integration techniques
3. Comparison and evaluation of data integration techniques.

Generation of data. The objective of the data-generation process was to simulate operational loss data and associated risk indicators for five years in a number (N) of financial institutions. In Figure 1, the simulated data for each bank is characterized by parallelograms

marked from A to N, representing each bank. Note that these parallelograms vary in size, reflecting the fact that data across banks are heterogeneous because some banks are bigger or riskier than are others are. Data simulations allow sampling two types of data as shown in Figure 1. The left-hand column of sampled data represents the data comparable to a shared internal and external data to a specific bank (e.g., bank A). The sampled data were comparable to a real existing dataset such as the ORX dataset. The right-hand column represents repeated hypothetical five-year loss samples in a specific financial institution or, alternatively, the losses in a span of sufficiently many years conditional on current financial institution indicators remaining similar.



Note: Numbered rectangles define the three main processes, parallelograms denote data; $\mathbf{OpRK}_i$ denotes the level of operational risk capital obtained by each technique *i*.

Figure 1. Research framework.

The BCBS (2011) suggested financial institutions use five years of data because this period is long enough for some rare operational losses to appear but short enough to reflect

the current situation and exposure to risk for a financial institution. The simulation on the right column of Figure 1 assumes that the current conditions and exposure to risk of a specific financial institution are valid for a sufficient number of years for the true nature of operational risk to reveal itself. The left-hand data simulation column represents only five years of operational loss events for bank A (internal data) as well as for a number of banks (external data). Bank A cannot pool the external data directly into its internal data to perform operational loss calculations.

The data generation process was implemented through a dynamic model of operational loss occurrence due to internal fraud in the retail banking segment of financial institutions in the vein of Kühn and Neu (2004), Leippold and Vanini (2005), and Bardoscia and Bellotti (2011). The model took into account the specific factors that trigger internal fraud losses in each of *N* banks that take part of the dataset pool. The possible dependencies that exist across time and banks also needed to be considered. Given that a key purpose of the research was to compare data integration techniques, the model, in the absence of specific operational loss data, was used to simulate the data to apply the integration techniques and perform the techniques comparison by means of a simulation study. As Figure 2 shows, once operational losses are drawn, it was also necessary to consider the loss recording processes by firms that take account of threshold levels (losses not recorded below a certain level) or measurement errors.

As explained above, the simulation process generated two types of data: The Type-1 simulation generated five years of data for a group of heterogeneous firms that share their data. The Type-2 simulation drew data for a specific firm along many years. The first simulation type is closer to reality, and the data generated needed a further data integration process. The second simulation type is referential. Firms, in theory, would want data for a huge number of years to be able to record even the rarest of operational losses, but such a

long period entails changes in the internal and external environments that banks face. This referential data simulation permitted uncovering the true nature of operational risk conditional on the current and specific bank environment. The operational risk capital estimated from this referential simulation is straightforward. There was no need for a data integration process because all the loss data come from the same bank. The Type 1 simulated data illustrated in Figure 2 are heterogeneous. These data need a further process of data integration.

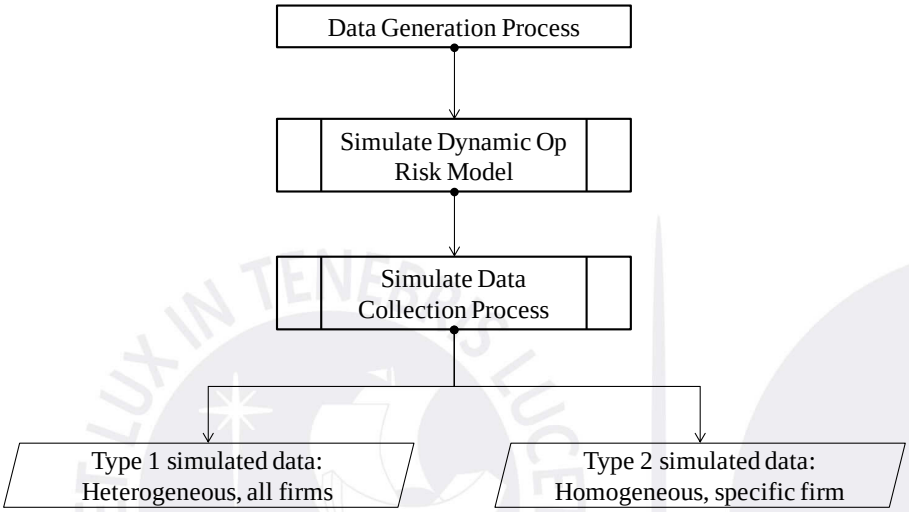


Figure 2. Components and outputs of the data generation process.

Application of data integration techniques. This was the second step in the research design. Data integration techniques allowed for appropriate incorporation of external data of bank A to estimate the operational risk capital level. The approach was consistent with the LDA framework. The simulated data were entered as input in the joint data integration and the LDA process. The output of this joint process was the PDF of operational losses that occurred for a given planning year. The PDF summarized the entire operational risk profile because it revealed the expected amount of losses, the variability of these losses, and more importantly, the loss amount above which there was a very small probability (e.g., 0.1%) of having an even larger loss. This extreme loss amount is called VaR. The VaR is a widespread risk measure to calculate the operational risk capital of a financial institution. There are a number of other related risk measures with better theoretical properties (McNeil et al., 2005),

such as the tail value at risk (TVaR) or conditional value at risk (CVaR), that, potentially, could have been used. In this study, the VaR, namely, the 99.9 or the 99.99 percentile, was used to obtain operational loss capital levels for each technique. Financial institutions have to report to their respective financial service authority on the level of operational risk capital they set aside to mitigate the impact of latent operational losses in a financial planning year.

The LDA process involves estimating the probability distribution of the frequency of losses that will occur during a planning year together with the estimation of the probability density of loss severities whenever they occur. Chernobai et al. (2007) and Shevchenko (2011) have detailed the implementation of the LDA framework.

The key idea in using the LDA is to estimate known parametric distribution functions for loss frequencies such as the Poisson or the Negative Binomial and parametric density functions for loss amounts. Estimation of loss amount densities is crucial in this framework because these loss amounts end up determining the overall operational risk capital. In other words, the estimation of operational risk capital levels is more sensitive to the loss amount PDF estimation than to the distribution of loss frequencies. This is the reason the current LDA literature has emphasized the estimation of the loss severity PDFs.

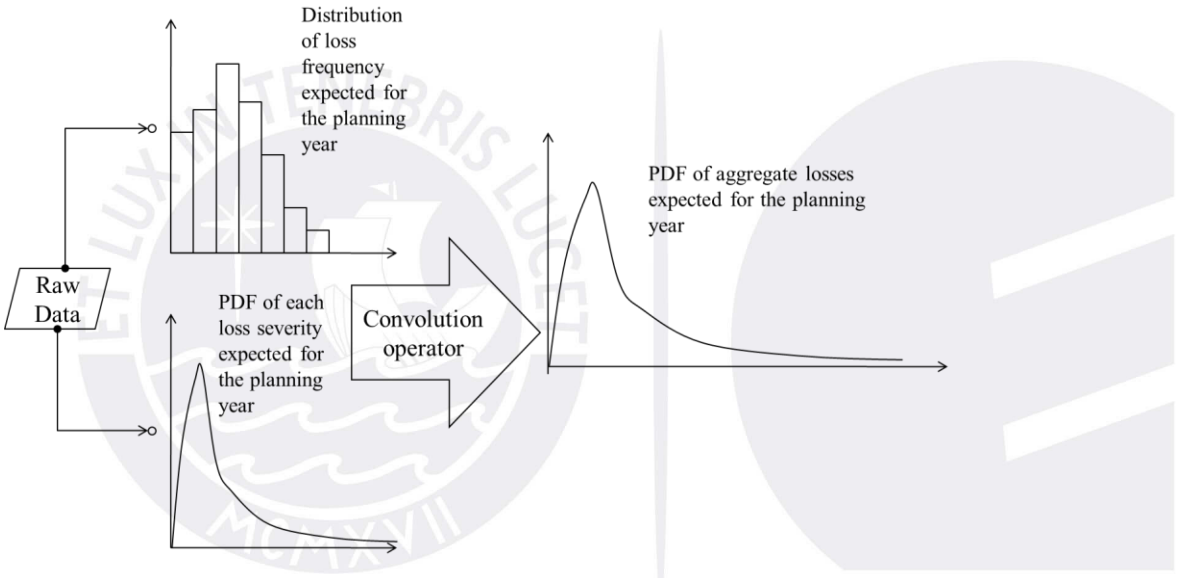
After estimation of both the loss frequency distribution and the severity density, the LDA hinges on applying a convolution operator to arrive at the PDF of the total operational losses expected over the course of the planning year. This convolution operator took into account that total annual losses, given by variable S_t, result from summing up the many losses that occur on a specified horizon. There are n_t losses in year t and let $z_{t,i}$ denote the severity of each loss in year t, then

$$S_t = \sum_{i=1}^{n_t} z_{t,i} \tag{2}$$

The loss frequency distribution draws the random variable n_t while the loss severity density draws each $z_{t,i}$. Panjer (2006) and Shevchenko (2011) described the convolution operator. Formally, the frequency of events n_t has a distribution function $p_n = Pr(N = n)$ while the loss severity $z_{t,i}$ has a density distribution and cumulative density functions denoted by f_Z and F_Z , respectively. Then, the cumulative density function of S_t is

$$\begin{aligned} F(S) &= Pr(w \leq S) \\ &= \sum_{n=0}^{\infty} p_n Pr(w|N = n) \\ &= \sum_{n=0}^{\infty} p_n F_Z^n(S), \end{aligned} \tag{3}$$

where $F_Z^n(S)$ is the *n*-fold convolution of the cumulative density function of S. Figure 3 shows the overall implementation of the LDA based on operational loss data.



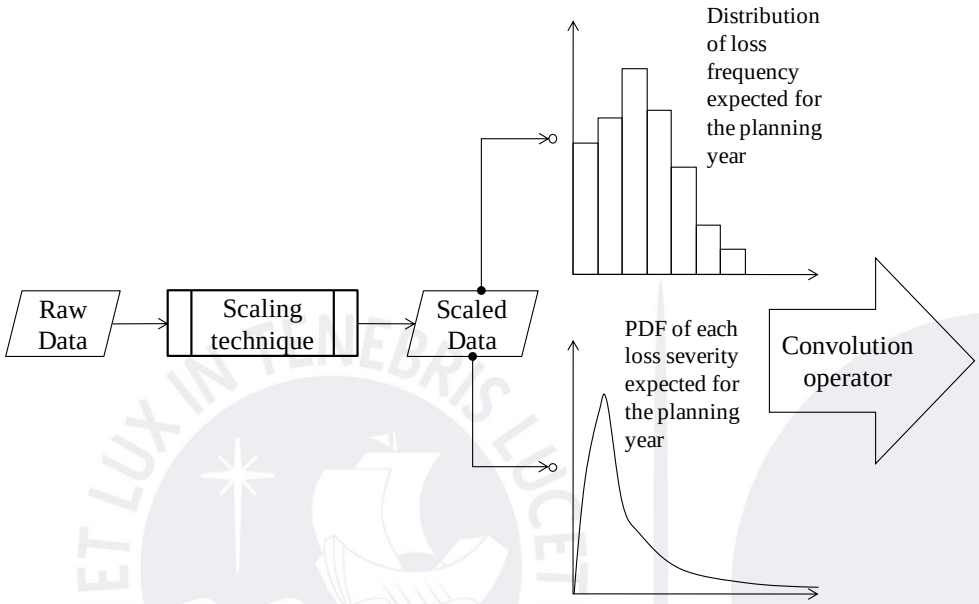
Note: The distribution and density functions are estimated with the raw data.

Figure 3. The standard LDA approach.

In this research design, this standard LDA approach worked only with Type-2 simulated data. Type-1 simulated data required a data integration process within the LDA. The different data integration techniques relate to different forms the LDA could be applied to heterogeneous data.

The Scaling technique. The Scaling technique transforms the external data to make them comparable to the internal data by means of some observable factors pertaining to external firms. After external data transformation by scaling, the pooled data enters the LDA

approach as depicted in Figure 4, where the Scaling technique process receives the raw data as input and generates scaled data as output. Papers like Shih et al. (2000) and Na et al. (2006) provided the Scaling technique framework. The LDA in this technique operates once the external data are scaled. In essence, the Scaling technique and the LDA are two separate processes. More details on the implementation of this technique are elaborated upon later in this chapter.

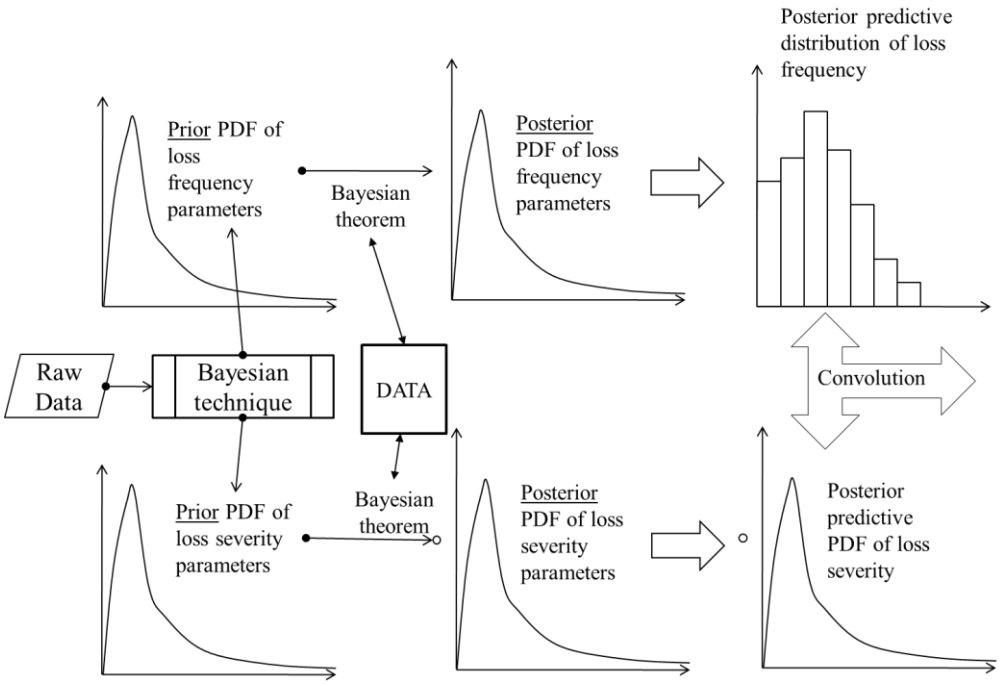


Note: The scaling technique modifies the raw data and obtains the scaled data as output.

Figure 4. The LDA approach with the Scaling technique.

The Bayesian technique. The Bayesian technique elicits a prior distribution of parameters that govern the loss frequency distribution and the loss severity PDFs as Figure 5 shows. Prior elicitation usually is based on subjective criteria such as expert judgments. In this study, the raw data were used to obtain empirical or objective priors distribution of parameters as in Lambrigger et al. (2007, 2008) and Hassani and Renaudin (2013). Next, the Bayesian framework combines the priors with the specific bank-level likelihood function of raw data to generate posterior densities of parameters via the Bayesian theorem.

The next step hinges on obtaining the posterior predictive distributions of the number of loss events for the next calendar year. Then, given the number of events expected to occur, the severities of each loss event are computed to obtain the total aggregate loss.



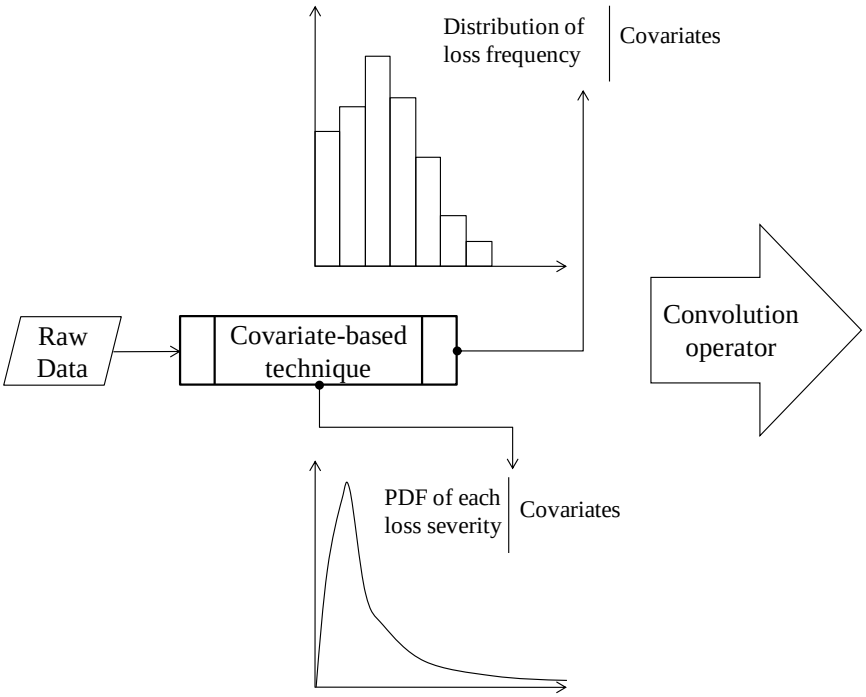
Note: The raw data were entered to elicit priors and obtain posteriors via the Bayesian theorem.

Figure 5. The LDA approach with the Bayesian technique.

Generally, only sampling techniques permit obtaining the posterior predictive density of aggregate losses for next calendar year. In a sense, the Bayesian technique affects the shape of the loss frequency distributions and loss severity densities in an informed way, disciplined by the Bayesian theorem. In this setup, there is no need to transform or scale the data.

The Covariate-based technique. The Covariate-based technique involves the modeling of the loss frequency distribution and the loss severity density in such a way that the shape and scale of these functions are affected by a set of covariates within the shared dataset. Ganegoda and Evans (2013) described a novel application of the technique; they used generalized additive models for location, scale, and shape (GAMLSS) as developed in Stasinopoulos and Rigby (2007).

Figure 6 shows that the Covariate-based technique modifies the shape of the loss frequency distribution and loss severity density to be used in the convolution. As in the Bayesian case, using this technique does not require transformation of the original external data.



Note: The set of loss data and covariates are used in models that affect the scale, shape, and location of the distributions and densities that enter the convolution operator.

Figure 6. The LDA approach with the covariate-based technique.

Comparison and evaluation of data integration techniques. The third and final step in the research design was the simulation study. In this step, a process was used to evaluate and compare techniques as depicted in Process 3 of Figure 1. After estimating the PDF of total losses under the three data integration techniques, it was straightforward to obtain VaR indicators or extreme quantiles (99%, 99.9%, or 99.99%). This step, statistically, compared these VaR measures associated with the Type-1 simulated data with the true VaR measure stemming from Type-2 simulated data.

For example, let $OpRK_S$, $OpRK_B$ and $OpRK_C$ be the levels of operational risk capital estimated by the Scaling, Bayesian, and Covariate-based techniques respectively, and let $OpRK_{true}$ be the true operational risk capital. The purpose of the research is to know whether any of the objects ($OpRK_S$, $OpRK_B$, or $OpRK_C$) was systematically closer to the true capital $OpRK_{true}$ given a specific percentile chosen by the risk manager. To perform the evaluation, a huge number S of simulations of complete 5-year histories of operational risk events are

drawn. Each simulation $j = 1, \cdots, S$ were used to generate capital levels $(OpRK_S^j, OpRK_B^j, OpRK_C^j)$.

Next, the root mean square errors (RMSE) can be computed according to

$$RMSE_k = \sqrt{\frac{1}{S}\sum_{j=1}^{S}\left(OpRK_k^j - OpRK_{true}\right)^2}, \text{ for k} = S, B, C \quad (4)$$

The RMSE is the distance of the capital levels associated to each technique from the true capital level. In Equation (4), the lower the level of $RMSE_k$, the better the results; furthermore, given that the mean square error can be decomposed into the squared bias and the variance of the estimators, it was straightforward to know the origin of systematic discrepancies among the techniques. In terms of formulas, the bias is defined by

$$BIAS_k = \overline{OpRK}_k - OpRK_{true}, \text{ for k} = S, B, C \quad (5)$$

where $\overline{OpRK}_k = \frac{1}{S}\sum_{j=1}^{S} OpRK_k^j$ is the mean operational risk capital level or technique *k* across simulated samples. A positive bias meant that the operational risk level estimated with a given technique was systematically larger than the true exposure to risk. The standard deviation of the estimator is

$$SD_k = \sqrt{\frac{1}{S-1}\sum_{j=1}^{S}\left(OpRK_k^j - \overline{OpRK}_k\right)^2}, \text{ for k} = S, B, C \quad (6)$$

A high standard deviation meant that the estimator of the operational risk capital level obtained by a technique was too uncertain. According to Voss (2013), the following approximation holds for the mean square error:

$$RMSE_k{}^2 \cong BIAS_k{}^2 + SD_k{}^2, \text{ for k} = S, B, C \quad (7)$$

This implied that the sources of the RMSE variation across techniques could be decomposed as bias and variability of the implied operational risk capital levels.

It is important to mention that this type of simulation-based comparison is widely used in academia to compare estimators. Estimators are statistical procedures that map data to

parameter estimates. The study carried out in this thesis is about the comparison of data integration techniques which are also estimators because they map data to VaR estimates. Traditional hypothesis testing cannot be done here in the sense that inferences from an observed sample to population statements are not possible because there are no observed samples.

Examples of applications of simulation or Monte Carlo studies abound in the literature. For example a quick search at top econometrics and statistical journals suffice to find research that apply simulation studies: Matzkin (2003) proposes a non-parametric estimation of random functions and compare the estimator with other estimators via bias, variance and mean square errors relative to true data generating processes. Chen & Pouzo (2012) propose estimators in the context of a conditional moment instrumental variable set up and compare the estimators also by using bias, variance and mean square errors.

More recently, Wang and Zhao (2016) propose a semi-parametric estimator of CVaR and compare this estimator with many alternative CVaR estimators by using a measure of relative integrated mean square error. Sarafidis (2016) proposes a new estimator of parameters in the context of dynamic panel data models where the errors are spatially correlated. Safaridis compares the proposed estimator against popular dynamic panel estimators by means of root mean square errors.

The research strategy used in this study is akin to the above papers because it compares techniques that deliver VaR levels and therefore a RMSE performance is the tool to be used.

Appropriateness of Research Design

The design outlined was appropriate to accomplish the goal of this research because a well-defined statistical procedure to answer the research question directly is applied to

discover if one particular data integration technique is better than the others for revealing the true operational risk profile.

The design was the optimum choice for this specific research. Following the example of statisticians and econometricians (Matzkin, 2003; Chen & Pouzo, 2012; Sarafidis, 2016; Wang and Zhao, 2016), simulation-based evaluation of estimators is optimal to infer the properties of estimators or statistics that depend on the data (Greene, 2012; Stern, 2000). In this study, the estimators were operational risk capital levels obtained with the data integration techniques, and the data referred to the operational losses.

Research Questions

The process of model building and broad reality checking with respect to operational losses due to internal fraud in retail banking raises important questions about internal fraud processes in financial institutions. The MRQs can be stated as follows:

MRQ1: If the model is capable of generating internal fraud losses that are similar to ones reported in the ORX database and produce correlations with macro environmental variables that are similar to those reported in the literature, how are these losses related to the Global Financial Crisis that occurred in the middle of the period of study?

This question has been studied for operational losses in general, but not for internal fraud losses in retail banking.

MRQ2: Given the same conditions as MRQ1, how are internal fraud losses related to perceptions about corruption in the country where the main headquarters of a bank is located?

Both the behavior of internal fraud losses before and after the 2007-2009 Global Financial Crisis and the correlation of those losses against corruption perception indices are not used in the model building or the calibration of parameters. Instead, these outcomes are independent

results of the model and provide valuable information about the nature of internal fraud losses in a financial institution.

MRQ3: Regarding the selection of the best internal-external data integration technique, is there any technique that can be considered best practice to estimate a correct operational risk capital across all levels of risk tolerance?

The simulated data generated by internal fraud model and the application of the three described data integration techniques allowed for testing the following key hypotheses:

$H_0$1: No change is evident in the pattern of operational losses before and after the Global Financial Crisis.

$H_0$2: Neither the frequency nor the severity of internal fraud operational losses are correlated with the corruption perception index of the country where the main headquarters of the bank is located.

$H_0$3: None of the three techniques is systematically better as compared to the others across possible risk tolerance values.

A Model for Internal Fraud Events

An essential element of the research design was the introduction of a dynamic simulation model for the occurrence of operational losses due to internal fraud within the retail-banking segment of a financial institution. In this section, the model is described in more detail. The aim of the model is to explain operational losses due to internal or insider fraud in the retail-banking context within each financial institution in terms of a set of conditioning factors. The main equation in the model is given by

$$l_{i,\tau} = \alpha_{i,0}\, ramp\left(\alpha_{i,1} + \alpha_{i,c} c_{i,\tau} + \alpha_{i,y} y_{i,\tau} + \alpha_{i,q} q_{i,\tau} + \xi_{i,\tau}\right) \quad (8)$$

where $l_{i,\tau}$ stands for an internal fraud loss in retail banking at bank i at moment τ. In the subscript, the Greek letter τ (*tau*) denotes moments of time during a given year t. In practice, it can represent days or hours within a year. The variable $c_{i,\tau}$ is the investment or

effort made by the bank to avoid the operational loss, or it can measure the level of internal controls. This variable can be measured as the share of monetary resources devoted to risk management and control and can be expressed as a percentage of operating costs. Higher standards of internal risk controls ($c_{i,\tau}$ high) imply that the likelihood of operational loss events is reduced. Internal fraud events are somewhat more controllable than losses originating from external sources (Chernobai et al., 2011). This control aspect of operational losses is outlined, for example, in Kochan (2013). The variable $c_{i,\tau}$ is also a measure of the control environment set by the organization in their people risk management effort (Blacker & McConnell, 2015).

The amount $y_{i,\tau}$ represents the scale of production in the business line; for retail banking, it can represent the number of transactions with bank clients, or it can represent the gross retail income. A higher number of transactions imply that the likelihood of operational losses increases. In the theory of people risk management, this scale $y_{i,\tau}$ level is a proxy of internal and external interactions, which give rise to operation loss risks due to fraud (Blacker & McConnell, 2015): The bigger the scale of the business, the higher the number of interactions. In an environment of increased employee interaction, fraud risks rise.

Variable $q_{i,\tau}$ measures the ethical quality of employees. High internal ethical standards mean that losses due to internal fraud are less likely to occur. The ethical quality of workers is different from the technical quality of workers, which is measured directly by worker productivity (e.g., gross income per worker). Therefore, the quantity and quality of human capital proposed by Fragnière et al. (2010) and Yang (2010) are key determinants of operational losses in the retail-banking segment of any bank.

From the variables explained so far, the volume of retail loans $y_{i,\tau}$ is directly observable. Information about this variable can be gathered from the annual reports of each of the banks in the ORX dataset. On the other hand, the level of controls $c_{i,\tau}$ and the quality

of employees $q_{i,\tau}$ are not directly observable. Specific feedback equations are required to model both variables to elicit their unobserved values and see how they quantitatively affect the generation of losses through Equation (8) above.

The last variable left to be explained in Equation 8 is $\xi_{i,\tau}$. This variable represents unknown factors or shocks that can potentially trigger losses. This random variable is assumed to be autocorrelated and heteroskedastic. The idea that loss shocks are autocorrelated and may exhibit volatility clustering is similar to what Chernobai and Yildirim (2008) and Guegan and Hassani (2013) suggested. In particular

$$\xi_{i,\tau} = \rho_i \xi_{i,\tau-1} + \sigma_{i,\tau}\mu_{i,\tau}, \mu_{i,\tau} \sim N(0,1), \sigma_{i,\tau}^2 = \beta_{0,i} + \beta_1 \xi_{i,\tau-1}^2 + \beta_2 \sigma_{i,\tau-1}^2 \quad (9)$$

The coefficient $\rho_i \in [0,1]$ measures the level of autocorrelation or the persistent nature of shocks that may trigger losses. The error term $\sigma_{i,\tau}\mu_{i,\tau}$ is heteroskedastic by virtue of the time-varying nature of the variance term. The variance term $\sigma_{i,\tau}^2$, also known as conditional variance, depends on past shock realizations as well as past variance itself. The way conditional variance behaves is called generalized autoregressive conditional heteroscedasticity (GARCH), as proposed in Bollerslev (1986).

Equation 8 also calls the function $ramp(.)$, which represents the mapping from operational loss factors to loss severities. This function has the feature of generating zero losses most of the time and positive loss severities at other times. The loss severities are correlated with the factors described in Equation 8. Kühn and Neu (2004) and Bardoscia and Bellotti (2011) used the same type of function. Formally, the ramp function is defined by

$$ramp(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (10)$$

To finish the description of Equation 8, all the coefficients $(\alpha_{i,0}, \alpha_{i,1}, \alpha_{i,c}, \alpha_{i,y}, \alpha_{i,q})$ vary across banks, but they are constant through time. This reflected the fact that operational

loss occurrences were sensitive to each of the factors described, which are idiosyncratic for each bank.

The levels of operational risk control, $c_{i,\tau}$, and the ethical quality of the workforce, $q_{i,\tau}$, need to be simulated due to their unobservable nature. The simulation of these variables was achieved by setting up model equations that captured their behavior.

First, the level of operational risk controls can be modeled by a feedback equation whereby the controls or efforts by risk managers to prevent or mitigate operational losses depend on the observable state of the system. Control is a fundamental aspect of risk management, the actual ISO standard defines risk management as "coordinated activities to direct and control an organization with regards to risk" (ISO, 2009).

The feedback from the observable state of the system to the level of controls is found in studies spanning a number of disciplines. For example, in the human resource literature, Lukic, Margaryan, and Littlejohn (2013) emphasized the process of learning from incidents as a key mechanism for improving management. In the management literature, AlHussaini and Karkoulian (2015) emphasized knowledge management in the efforts to mitigate risk in the banking industry. In the field of operational research, controls can be associated with process improvement (Mizgier, Hora, Wagner, & Jüttner, 2015). A comprehensive assessment of risk controls in organizations, as regards to people risk, is found in Blacker and McConnell (2015, Chapter 8) where the control process emphasizes the assignment of responsibilities at various levels of the organization.

In this study, the levels of controls are represented by a sufficient statistic denoted by $c_{i,\tau}$. The learning or improved control process depends on the level of risk. This idea is common in stochastic control environments and follows the example of Cooke and Rohleder (2005) who proposed a very general feedback model of operational risk. The study

incorporated this idea but was explicit about the level of risk that feeds back onto the control level. The feedback control equation can be expressed as

$$c_{i,\tau} = \frac{\rho_c(2c_i^*)}{1 + e^{\gamma_i\left(\frac{\hat{L}_{i,\tau-1}}{Y_{i,\tau-1}} - \lambda_i\right)}} + (1 - \rho_c)c_{i,\tau-1} \tag{11}$$

where c_i^* stands for the optimal level of controls associated with a benchmark loss ratio λ_i when the actual loss ratio is given by $\frac{\hat{L}_{i,\tau-1}}{Y_{i,\tau-1}}$. The control level at bank i depends on the observed key risk performance given by the ratio of cumulative average observed losses over the stock of retail loans. If the observed loss ratio is beyond the desired level, with $\gamma_i < 0$, control levels need to be adjusted upward. The degree of the actual adjustment depends on the parameter $\rho_c \in [0,1]$. The higher ρ_c, the quicker the control response is. In the opposite case, when ρ_c is small, the control level is governed by its previous value.

In Equation 11, $\hat{L}_{i,\tau-1}$ stands for the cumulative average observed losses up to the previous time, while $Y_{i,\tau-1}$ corresponds to the stock of retail loans granted in the same period. Capital letters stand for average quantities, while the circumflex ($\widehat{\ }$) denotes that the variable is the observable counterpart of an unobservable underlying variable. In the case of operational losses, this distinction is important. An observed loss amount in a bank i at time τ is $\hat{l}_{i,\tau}$ whereas the true loss is $l_{i,\tau}$. Equation 11 means that banks, which experience a history of large losses relative to other banks, will learn from the incidents and therefore increase their controls to levels above average. This idea is also suggested in Lukic et al. (2013).

The second key variable that needs modeling is the quality of the workforce $(q_{i,\tau})$, which refers to ethical traits that drive worker behavior toward the bank. It measures the propensity of workers to commit fraud. For example, an employee can be extremely knowledgeable of internal processes at the bank and so be highly productive, but good knowledge of internal processes may make it easy to commit fraud (Cummings et al., 2012).

The equation that describes the ethical quality of workers is

$$q_{i,\tau} = \frac{\rho_q(2\bar{Q})}{1 + e^{\delta_i(a_{i,\tau} - \bar{A}_\tau)(eb_{i,\tau} - \overline{EB}_\tau)}} + (1 - \rho_q)q_{i,\tau-1} \tag{12}$$

where $a_{i,t}$ stands for measured technical quality (labor productivity), $\bar{A}$ is the cross-bank average labor productivity, $eb_{i,\tau}$ is the number of employees per branch at bank i, and $\overline{EB}$ is the cross-bank average of employees per branch. Given that $\delta_i > 0$, the sign of the impact effect of an increase in technical quality is given by

$$\frac{\partial q_{i,\tau}}{\partial a_{i,\tau}} = \begin{cases} < 0 & if \quad eb_{i,\tau} > \overline{EB}_\tau \\ \geq 0 & if \quad eb_{i,\tau} \leq \overline{EB}_\tau \end{cases} \tag{13}$$

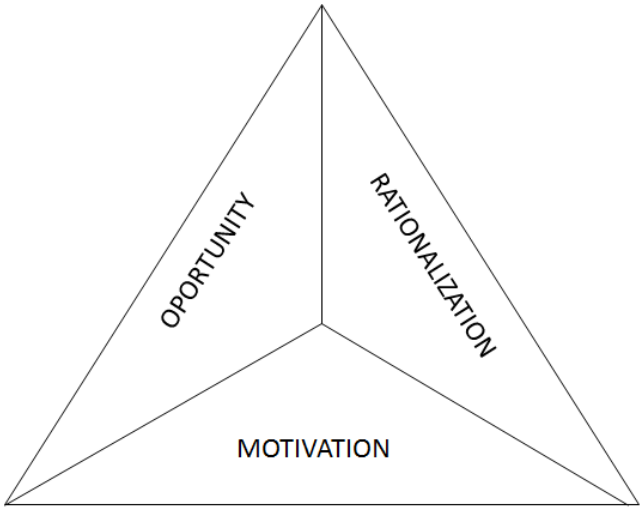
When there are few workers, increasing productivity is more likely associated with high ethical quality because it is easier for banks to screen workers before and after recruitment. When the number of workers is high, the workforce screening process is weaker. Due to the symmetry of Equation 12, it is also true that

$$\frac{\partial q_{i,\tau}}{\partial eb_{i,\tau}} = \begin{cases} < 0 & if \quad a_{i,\tau} > \bar{A}_\tau \\ \geq 0 & if \quad a_{i,\tau} \leq \bar{A}_\tau \end{cases} \tag{14}$$

which means that an increase in the number of workers harms the ethical quality of workers when the average technical productivity of workers is already high.

Equation 12 incorporates, in an explicit way, two concepts in the theory of people risk management. First, the basic fraud model based on the Cressey's fraud triangle (see Figure 7) asserted that fraud has three elements: Motivation or pressure to commit fraud, the opportunity to commit fraud, and the rationalization or justification that a fraudster makes to him or herself to commit fraud. An employee, often in dire financial straits, using its insider information about the firm's control system, redirects funds to other sources.

The insider information about the control environment is possible if the employee possesses knowledge about many processes in the bank. In this study, this knowledge is approximated by the technical productivity of workers in firm i at time t: $a_{i,\tau}$



Note: Adapted from *Other People's Money: A Study in the Social Psychology of Embezzlement* by D. R. Cressey. Copyright 1953 by Free Press.

Figure 7. Cressey's fraud triangle.

The second key fraud theory concept embedded in Equation 12 refers to interactions in the firm. In the words of Blacker and McConnell (2015): "Inappropriate interactions between individuals inside and outside of the firm give rise to People Risk" (p. 121). Blacker and McConnell underscore the qualitative nature of employee interactions. In this study, it is argued that the qualitative level of interactions (inappropriate or bad) is increased with the quantitative number of interactions that should be proportional to the number of employees scaled per branch $eb_{i,\tau}$ at bank i during period t.

Therefore, Equation 12 shows that the ethical quality of employees (inverse of the propensity to commit fraud) falls when both opportunities for fraud and the number of inappropriate interaction rise as suggested by the theory of people risk.

After losses $\{l_{i,\tau}\}$ for the set of banks $i = \{1, \cdots, N\}$ at high frequencies $\tau = \{1, \cdots, T\}$ are generated by the stochastic dynamic system given by Equations 8 to 12, the loss data have to be recorded and submitted to the pooled database. In practice, pooled operational loss data

in the financial system are gathered in various ways. The main form of data sharing is provided by consortium databases collected by multilateral interbank agreements. The data are dependent on participating banks that commit themselves to sharing their internal data. These consortium databases ensure that confidential data stay protected. The benefits for participating banks are the availability and usability of the data. The research was based on the summary statistics of the only publically available information disclosed, the ORX database (ORX, 2012).

Of note is that data recorded to build the loss datasets were not the same as the original loss data $\{l_{i,\tau}\}$ for a number of reasons. For example, the existence of recording thresholds indicated that only losses greater than a threshold level l_i^{min} were submitted to the dataset. Moreover, when a loss event occurred, banks did not necessarily know the exact loss amount incurred. There is a natural lag between occurrence of an event that involves loss and knowledge of the severity of the event. The lag depends on the specific nature of the event. For the purposes of this research, it was assumed that the severity of the event was known at the same time as the occurrence but that the knowledge was imperfect and subject to measurement errors. Therefore, the observed dataset process implies

$$\hat{l}_{i,\tau} = l_{i,\tau} + \eta_{i,\tau} \tag{15}$$

where $\hat{l}_{i,\tau}$ is the observed loss severity, $l_{i,\tau}$ is the true unobserved loss severity (underlying loss) and $\eta_{i,\tau}$ is an unbiased, white noise measurement error distributed normally $\eta_{i,\tau} \sim N(0, \sigma_{\eta i}^2)$. The measurement errors are independent over time and across banks, but heterogeneity across banks is allowed. The losses submitted to the pooled dataset and used for internal purposes are described by

$$D_i = \{\hat{l}_{i,\tau} \mid \hat{l}_{i,\tau} > l_i^{min}\} \quad for\ \tau = 1, \cdots, T \tag{16}$$

In the data collection process, both $\sigma_{\eta i}^2$ and l_i^{min} were assumed to be exogenous. In fact, the value of l_i^{min} determined for all member banks of the ORX association is €20,000.

Once the internal fraud loss simulation model was established, the second aim of this study was to use the simulated data to compare three techniques that combine internal data to a bank i given by D_i and the external data available from the data exchange given with $\{D_j \mid j \neq i\}$. Before detailing the data integration techniques, it is important to explain the parameter calibration procedure to be applied to the simulation model.

Calibration procedure. The parameters that specify the dynamic operational loss model described above are not free. It was necessary to restrict the parameters to specific values. Standard estimation techniques such as linear regression cannot be conducted because there was no hard data and the model exhibited unobservable variables like control and quality of workers. This implied that the model parameters needed to be calibrated, not estimated. Calibration means that each of the parameter values have been chosen following heuristic principles. The simulation process was conditional on the value of these parameters.

The parameters were divided in four subsets. First are parameters that define the main equation for internal fraud loss events in Equation 8. Second are parameters that affect the feedback equation from operational loss performance to risk controls given by Equation 11, third are parameters that describe the evolution of workforce ethical quality given by Equation 12, and finally, a single parameter that defines the measurement error in observed losses is identified.

Some parameters are specific to banks (idiosyncratic), so they have the subscript i in their notation. Other parameters are common to all banks. The procedure for calibration of all parameters, general and specific, is described below. There are 13 parameters idiosyncratic to each bank (see Table 3). Given that there are up to 52 banks, it would be necessary to pin-down about 672 idiosyncratic parameters. Given the vast number of parameters to be

calibrated, a very simple shrinkage method was introduced to reduce the number of parameters to be calibrated based on the available information, the period of analysis, and the specific banks under study.

Table 3

Parameters of the Dynamic Model for Operational Losses

Parameter	Equation	Definition
$\alpha_{i,0}$	Internal fraud losses	Overall scale of losses
$\alpha_{i,1}$	Internal fraud losses	Constant within ramp function
$\alpha_{i,c}$	Internal fraud losses	Impact of controls on losses
$\alpha_{i,y}$	Internal fraud losses	Impact of gross operating income on losses
$\alpha_{i,q}$	Internal fraud losses	Impact of quality of workers on losses
ρ_i	Internal fraud losses	Autocorrelation of operational loss shocks
$\beta_{0,i}$	Internal fraud losses	Constant in conditional variance of operational loss shocks
β_1	Internal fraud losses	Influence of past shocks on conditional variance of operational loss shocks
β_2	Internal fraud losses	Influence of past variance on conditional variance of operational loss shocks
ρ_c	Loss control	Weight of new conditions to affect current controls
c_i^*	Loss control	Control level associated to desired operational loss ratio
γ_i	Loss control	Controls de sensitivity of control to the loss ratio gap from the desired ratio
λ_i	Loss control	Desired loss ratio
ρ_q	Ethical quality	Weight of new conditions to affect current ethical quality levels
δ_i	Ethical quality	Determines the sign of impacts from factors
$\bar{Q}$	Ethical quality	Level of average ethical quality across banks
$\bar{A}$	Ethical quality	Average labor productivity across banks
$\overline{EB}$	Ethical quality	Average number of employees by branch across banks
$\sigma_{\eta i}^2$	Measurement error	Variance of measurement errors

In essence, the shrinkage method used in this study took into account the idiosyncratic data that were collected for each bank. These data proxy the degree of riskiness and heterogeneity of each bank and are used to map the heterogeneous values of the model parameters.

The starting point for model calibration is the operational loss summary report presented in the ORX database (ORX, 2012) for the period 2006-2010. In this report, there are 4,357 internal fraud loss events recorded in the retail-banking segment; the gross amount of losses measured reaches €880 million. The summary pertains to losses reported to ORX during the years 2006-2010 by a number of active member banks during that period.

Table 4

Member Banks of the ORX Data Exchange by Country and Selected Dates

	Bank name	Country	jun-08	sep-09	may-10		Bank name	Country	jun-08	sep-09	may-10
1	Commonwealth Bank of Australia (CBA)	AUS	no	no	yes	27	Banc Sabadell	ESP	yes	yes	yes
2	National Australia Bank	AUS	no	yes	yes	28	Banco Bilbao Vizcaya Argentaria (BBVA)	ESP	yes	yes	yes
3	Westpac Banking Corporation	AUS	no	no	yes	29	Banco Pastor	ESP	yes	yes	yes
4	Bank Austria – Creditanstalt	AUT	yes	yes	yes	30	Banco Popular	ESP	yes	yes	yes
5	Erste Group Bank AG	AUT	yes	yes	yes	31	Banco Santander	ESP	yes	yes	yes
6	Fortis	BEL	yes	yes	yes	32	Banesto	ESP	yes	yes	yes
7	Banco Bradesco S/A	BRA	no	yes	yes	33	Caixa Catalunya	ESP	yes	yes	yes
8	Bank of Nova Scotia	CAN	yes	yes	yes	34	Caixanova	ESP	no	yes	yes
9	Bank of Montreal (BMO Financial Group)	CAN	yes	yes	yes	35	Caja Laboral	ESP	yes	yes	yes
10	Royal Bank of Canada (RBC)	CAN	yes	yes	yes	36	Cajamar	ESP	yes	yes	yes
11	Toronto Dominion Bank Group (TD BG)	CAN	yes	yes	yes	37	Skandinaviska Enskilda Banken (SEB)	SWE	yes	yes	yes
12	Danske Bank A/S	DNK	yes	yes	yes	38	Standard Chartered Bank	GBR	no	yes	yes
13	BNP Paribas	FRA	yes	yes	yes	39	Barclays Bank	GBR	yes	yes	yes
14	Credit Agricole SA	FRA	yes	yes	yes	40	HBOS PLC	GBR	yes	no	no
15	Société Générale	FRA	no	no	yes	41	HSBC Holdings plc	GBR	yes	yes	yes
16	Commerzbank AG	DEU	yes	yes	yes	42	Lloyds Banking Group	GBR	yes	yes	yes
17	Deutsche Bank AG	DEU	yes	yes	yes	43	Royal Bank of Scotland Group	GBR	yes	yes	yes
18	Deutsche Postbank AG	DEU	no	yes	yes	44	Bank of America	USA	yes	yes	yes
19	Bank of Ireland Group	IRL	yes	yes	yes	45	Capital One	USA	no	yes	yes
20	Intesa SanPaolo	ITA	yes	yes	yes	46	JPMorgan Chase & Co.	USA	yes	yes	yes
21	ABN AMRO	NLD	yes	yes	yes	47	National City	USA	yes	no	no
22	ING Group	NLD	yes	yes	yes	48	PNC Bank	USA	no	yes	yes
23	Rabobank Nederland	NLD	no	yes	yes	49	US Bancorp	USA	yes	yes	yes
24	Banco Portugues de Negocios	PRT	yes	yes	yes	50	Wachovia Corporation	USA	yes	no	no
25	First Rand	ZAF	no	yes	yes	51	Washington Mutual	USA	yes	no	no
26	Hana Bank	KOR	yes	yes	yes	52	Wells Fargo & Co	USA	no	yes	yes

Note: Adapted from "Use of External Data for Operational Risk Management" by J. Sabatini and S. Wills, presentation at the Use of External Data for Op Risk Management Workshop, Bank of Japan. Retrieved from https://www.boj.or.jp/en/announcements/release_2008/data/fsc0804a4.pdf. Copyright 2008. "Profile of ORX and a Case Study in the Use of Consortium Loss Data" presentation made by J. Sabatini at the First International Conference on External Data for Operational Risk, Associazione Bancaria Italiana. Retrieved from http://www.abieventi.it/documenti/2973/Sabatini-JPMorgan-Chase-ORX.pdf. Copyright 2009. "Quantifying operational risk," paper presented by S. Patel in the 2010 Seminar on Reinsurance, Casualty Actuarial Society. Retrieved from http://www.casact.org/education/reinsure/2010/handouts/CS14-PatelAppendix.pdf. Copyright 2010.

Table 4 shows the banks that reported losses to the ORX data exchange during the period 2006-2010. The table shows three benchmark dates where membership data were publicly available from ORX officials. New members constantly enter the association, and some members quit due to bankruptcies, mergers, or acquisitions. For example, Wachovia was a member of the association until acquired by Wells Fargo in 2008.

As Table 4 shows, member banks are gathered all over the world but belong mainly to advanced economies (Appendix B contains the full names of banks and their countries). One important feature of the period under study is the Global Financial Crisis of 2007-2009 that was particularly acute in some of the member banks in the dataset.

Given the information in Table 4, the model calibration implied a varying number of total banks in the sample. For example, up to end of 2008, there are $N = 39$ banks. In year 2009, 10 banks entered the data exchange association, and four banks quit the association, making $N = 45$ banks participating in the data exchange. By the end of 2010, one more banks entered the association, making $N = 48$ active banks. We track 52 banks in total and 35 banks that belongs to ORX the entire five-year period. The key workable assumption is that arrivals and departures from the association are set at the beginning of each year. In addition, the database contained only banks that operated a retail-banking segment. Therefore, some banks that belong to the ORX association but only perform investment banking or other lines of business were omitted from the database.

For each of the *N* banks and years under analysis, a set of variables categorized as key risk indicators or conditioning factors were gathered. These variables condition the occurrence of losses in the model or are useful devices to calibrate idiosyncratic parameters. The set of conditioning variables is described in Table 5. All the variables indicate idiosyncratic factors to each bank that proxy risk exposure such as number of employees or size of retail loans. These variables are useful devices to apply the shrinkage procedure because they discriminate among banks. For example, according to Cressey's fraud triangle explained before, banks that have higher employees per branch relative to the mean among banks might be deemed riskier than those that have lower employees per branch. Therefore, the dispersion of employees per branch across banks is useful to calibrate parameters across banks.

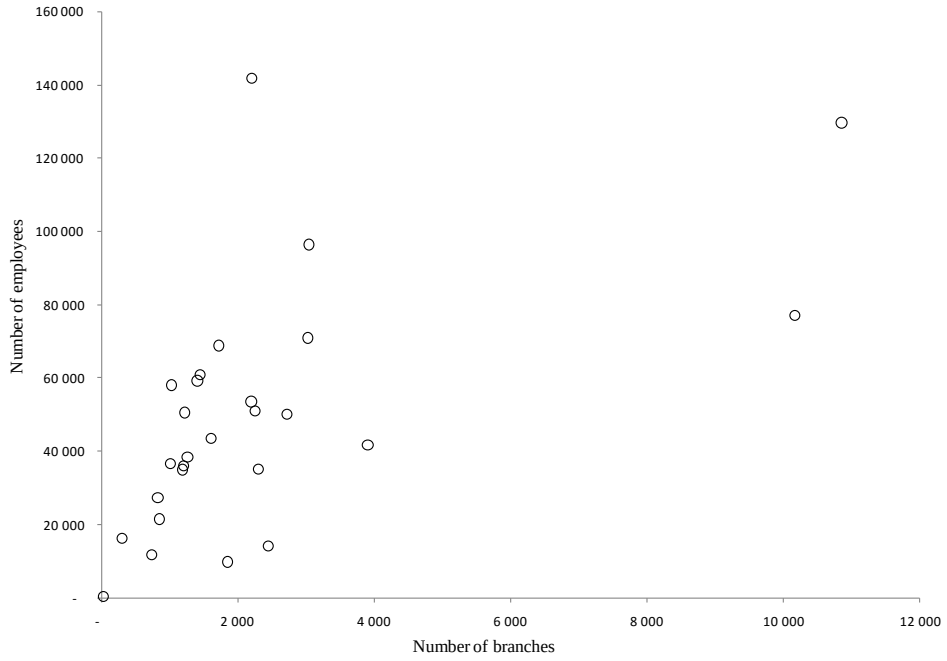
Table 5

Observable Variables that Condition the Simulation of Losses in Each Bank

Nomenclature	Description	Type
$e_{i,t}$	Number of employees	Idiosyncratic
$b_{i,t}$	Number of branches and offices	Idiosyncratic
$a_{i,t}$	Retail assets (millions of Euros)	Idiosyncratic
$y_{i,t}$	Retail loans (millions of Euros)	Idiosyncratic
$m_{i,t}$	Proxy for operational risk management awareness	Idiosyncratic
$h_{i,t}$	Proxy for human resource awareness	Idiosyncratic

In contrast to the model specification in the preceding subsection, the observed variables are indexed by time (t), where t stands for end-of year variables. Idiosyncratic variables for the years 2006 through 2010 were obtained from annual reports that member banks published on their Web pages. The values of interest were extracted from the descriptive information, balance sheets, and income statements contained in the aforementioned reports. These reports are publicly available as part of the information disclosure by banks directed to investors. The financial statements in these reports are compatible with sound regulatory and accounting practices and, on the majority of cases, they accord to GAAP.

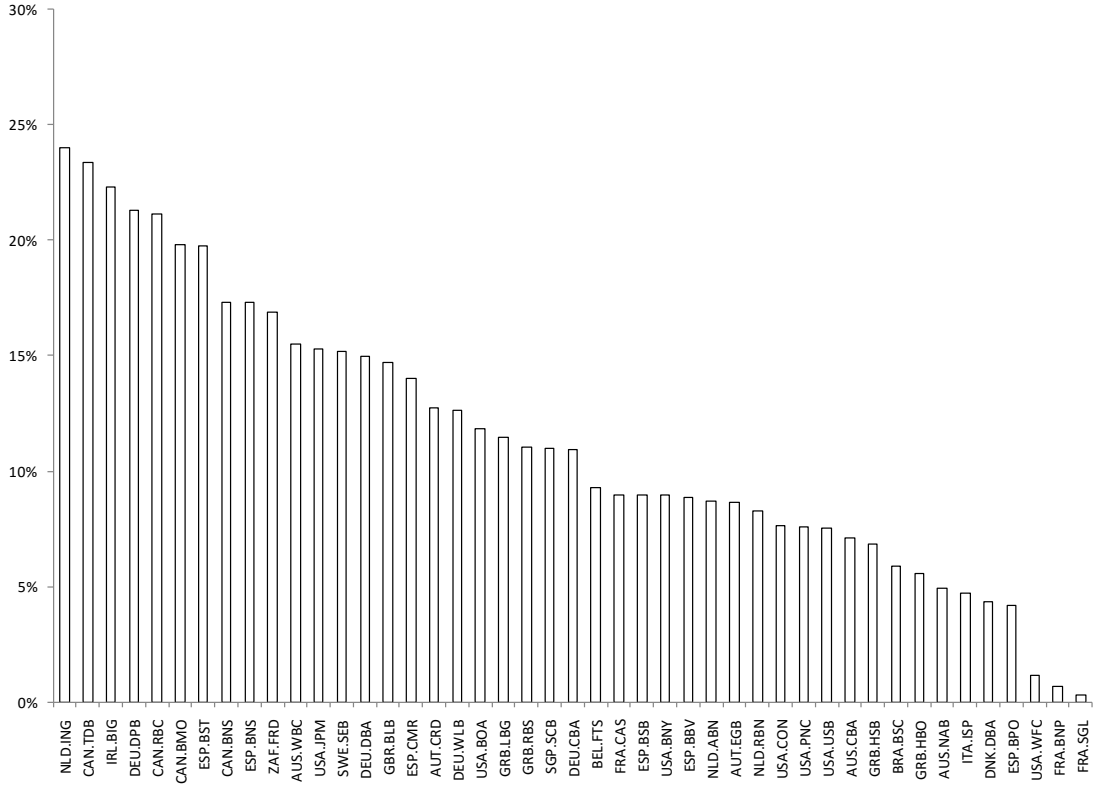
An example of the information recovered from these annual reports is given in Figure 8. The figure shows that the size of banks in the ORX dataset is heterogeneous. Each dot refers to a specific bank. The number of employees ranges from 10 to about 140,000, while the number of branches varies from 300 to about 10,000. In addition, both the number of employees and number of branches show some degree of correlation.



Note: Information extracted from bank's annual reports as of December 2006.

Figure 8. Number of branches and employees for banks in the ORX dataset.

In addition to the objective information included in the annual reports, proxy variables related to operational risk controls and the quality of human resources at each bank were constructed. Thus, variable $m_{i,t}$ measures operational risk management awareness implicit in the information shared with the public. This awareness proxy was obtained from textual analysis of annual reports; for example, the number of times a word or a phrase occurred within each report divided by the number of pages. An example of this type of textual information is given with Figure 9, which depicts the ratio of word counts of the expression "operational risk" as a percentage of the total number of pages. This variable could arguably reflect the extent of awareness of each bank toward operational risk management. The idea of extracting information from textual sources is not new in finance (Kearney & Liu; 2014). Textual information, also known as textual sentiment, reflects objective conditions within banks.



Note: For each bank, average percentages are reported.

Figure 9. Word counts of the expression "operational risk" as percentage of page counts in each report.

Other textual expressions that reflect operational risk awareness can be analyzed, for example, the use of the acronym AMA. To assess the validity of this type of proxy information, the textual indicators can be contrasted to objective measures like realized amount of total operational losses, as reported in Benyon (2008), for an important subset of banks in the ORX dataset in 2008.

The variable $h_{i,t}$ is intended to measure the awareness of banks about human resources. The annual reports also contain information about policies geared to improve the management and quality of human resources. So, human resource awareness could be obtained from textual analysis by extracting word counts of expressions such as "employees" or "human resources." The assumption was that these indicators reflect the quality of the workforce and are related inversely to the occurrence of internal fraud.

All the extracted information from bank's disclosures reflects the state of banks at calendar year-ends. Therefore, in order for these variables to be entered into the simulation model, it is necessary to perform simple linear interpolations to complete data for all moments of time τ between any consecutive years t and $t+1$. It is assumed that time τ will refer to business days within years. The existence of holidays was excluded in these calculations.

Five parameters affect the outbreak of operational losses in Equation 8. All these parameters are idiosyncratic; therefore, it was necessary to devise a way to calibrate all of them in a simple form. Let $\alpha_{i,j}$ be a parameter in Equation 8 for $i = 1, \cdots N$, and $j = \{0,1,c,y,q\}$. Then for each j, there is a mean parameter value taken from the cross section of banks. $\bar{\alpha}_j = \sum_{i=1}^{N} \alpha_{i,j}$. The parameters of interest were the $\alpha_{i,j}$ for each bank i. The idea is that by pinning down the value of $\bar{\alpha}_j$, it is possible to pin down the idiosyncratic variables $\alpha_{i,j}$ as well.

To complete the process, it was necessary to work with risk exposure indicators calculated from the data depicted in Table 5. This indicators are defined by $x_{i,t}$. This key variable can be given, for example, by the ratio of employees per branch or the technical quality of workers (labor productivity) measured as the ratio of total retail loans to the number of employees. These measures can be calculated for each financial institution and for each calendar year in the sample. For example, Figure 10 depicts a histogram of the number of employees per branch at the end of 2006 in all banks belonging to the dataset by the end of that year.

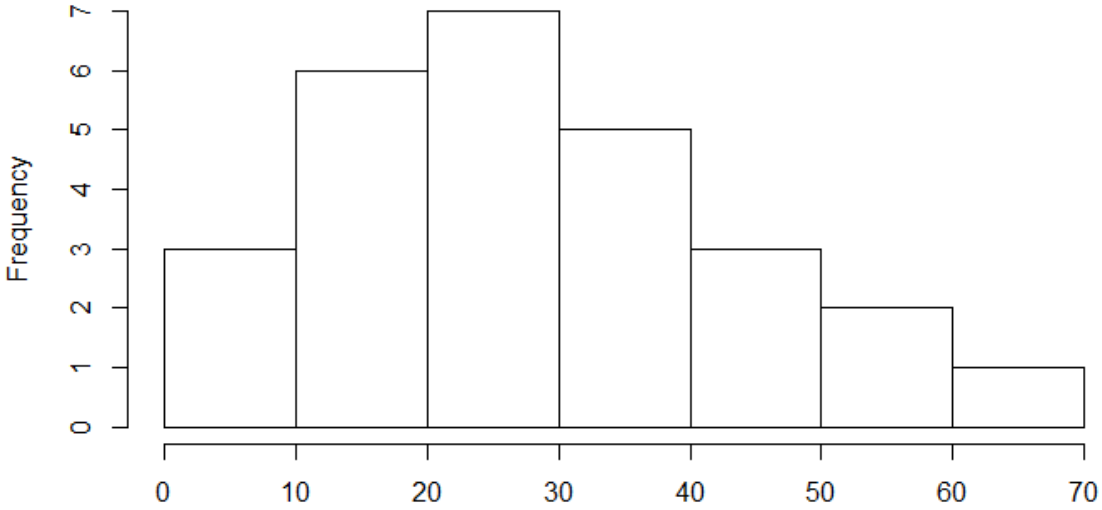


Figure 10. Histogram of the number of employees per branch across banks in the ORX database as of December 2006.

Next, let $\bar{x} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}$ be the overall average level of the risk indicator and $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{i,t}$ the average level of each indicator for bank i. If there are grounds to postulate direct proportionality between the coefficient $\alpha_{i,j}$ and the indicator $\bar{x}_i$, then each parameter can be pinned down according to

$$\alpha_{i,j} = \left(\frac{\bar{x}_i}{\bar{x}}\right)\bar{\alpha}_j \quad for\ i = 1, \cdots, N \tag{17}$$

For example, for parameters $\alpha_{i,0}$, $\alpha_{i,1}$, and $\alpha_{i,y}$, the choice of $x_{i,t} = \frac{e_{i,t}}{b_{i,t}}$ is a reasonable option. This means that loss event sensitivities are correlated with the ratio of employees to banks. In this case, only the parameter $\bar{\alpha}_j$ need be calibrated. Once its value is determined, Equation 17 fixes the distribution of $\alpha_{i,j}$ parameters across banks. The parameters $\alpha_{i,c}$ and $\alpha_{i,q}$ are likely to be inversely proportional to the ratio $x_{i,t} = \frac{e_{i,t}}{b_{i,t}}$. For example, controls are more effective when there are fewer people working at branches.

The adjustment can be made according to

$$\alpha_{i,j} = \left(\frac{\bar{x}}{\bar{x}_i}\right)\bar{\alpha}_j \quad for\ i = 1, \cdots, N \tag{18}$$

Parameters ρ_i and $\beta_{0,i}$ can be adjusted in the same fashion. In the case of the autocorrelation of shocks ρ_i, it is necessary to set bounds $\rho_{min} > 0$ and $\rho_{max} < 1$, such that the resulting operation $\rho'_i = \left(\frac{\bar{x}_i}{\bar{x}}\right)\rho$ can be further modified to become bounded within the range $[\rho_{min}, \rho_{max}]$. To do so, it is first necessary to calculate ρ'_{min} and ρ'_{max} with the parameters obtained given ρ and then to apply

$$\rho_i = \rho_{min} + \left(\frac{\rho_{max} - \rho_{min}}{\rho'_{max} - \rho'_{min}}\right)(\rho'_i - \rho'_{min}) \quad for\ i = 1, \cdots, N \tag{19}$$

The parameter $\beta_{0,i}$ controls for the unconditional variance of operational loss shocks in each bank, as shown in Equation 9. The dispersion of this parameter across banks can also apply the principle underlying Equation 17.

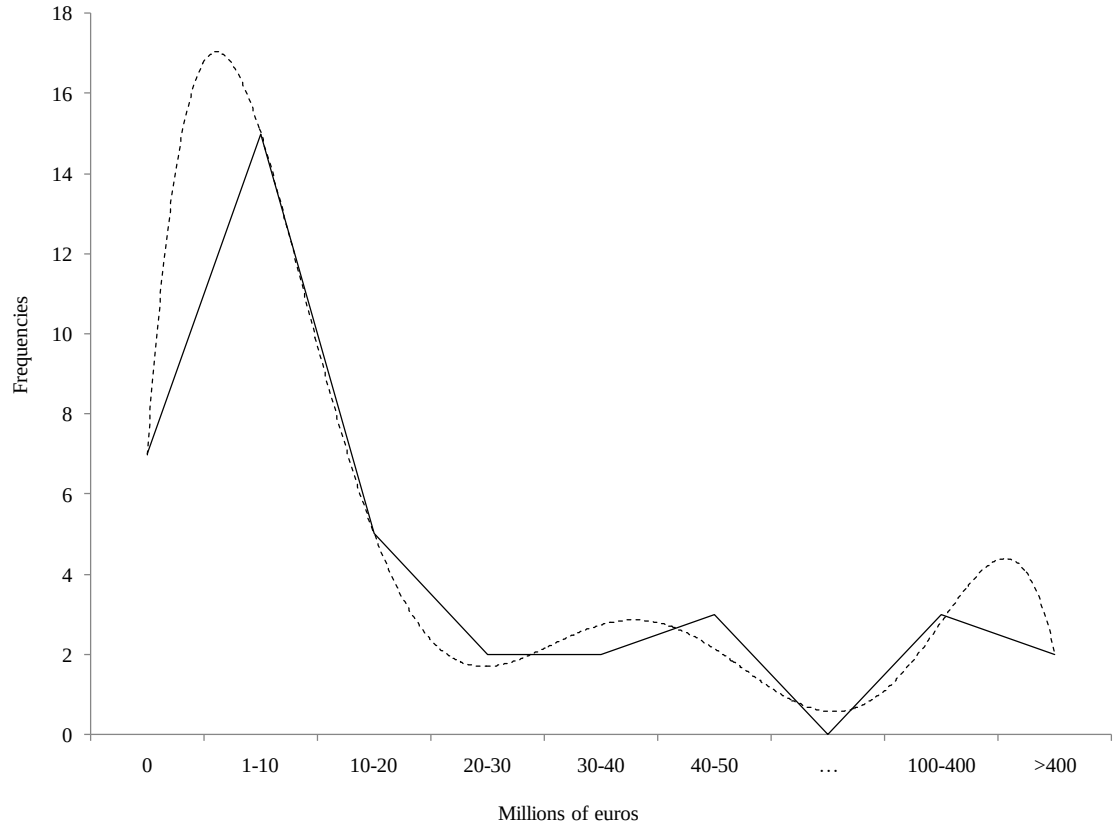
Three idiosyncratic parameters affect the level of controls $(c_i^*, \lambda_i, \gamma_i)$. The first parameter measures the long-run value of control levels. Current control levels may be stricter or easier than this long-run benchmark, which also needs to be on the range [0,1]. To calibrate the dispersion of this parameter, the indicator $m_{i,t}$ that measures operational risk management awareness can be used in the same fashion as the calibration of the dispersion of parameter ρ_i. Parameter λ_i reflects the operational loss level as a ratio of operating income that banks are ready to accept. Operational loss ratios larger than this benchmark λ_i prompt banks to increase their controls. Calibration of this parameter for each bank is problematic because the available information does not provide reasonable proxies for this ratio. Therefore, it was assumed that this ratio is similar across all banks in the sample. Its level was determined by the median ratio of cumulative operational losses to gross operating income for 2008 provided by Benyon (2008).

Parameter γ_i is a feedback adjusting parameter. From Equation 11, it is easy to note that for controls to tighten whenever the operational loss ratio increases, the parameter γ_i has to be negative. The greater γ_i is in absolute value, the greater the impact on control levels.

Again, it is reasonable to assume that the absolute value of γ_i is directly proportional to the level of operational risk awareness $m_{i,t}$ and thus, the calibration of the dispersion in γ_i will apply the same steps outlined above.

In terms of the ethical quality of human resources in a bank, the parameter δ_i measures how sensitive each bank's workforce quality is to the bank's size and labor productivity. Equation 12 assumes that size and labor productivity may be detrimental to workers ethical quality. Therefore, parameter δ_i is positive, and the higher it is, the more sensitive ethical quality becomes. The sensitivity may be related inversely to the human resource awareness proxy $h_{i,t}$ extracted from the data. Human resource awareness is related to degree of importance of the workforce in terms of well-being, compensations, and on-the-job training, for example.

The last idiosyncratic parameter calibration that needs a shrinkage procedure is the variance $\sigma_{\eta i}^2$ of the measurement error in recording the severity of operational losses. It is reasonable to assume that higher severity levels are associated with higher measurement error variances. To reflect this feature, the study used the aggregate operational loss figures reported in Benyon (2008). The losses reported by Benyon refer to the aggregate of all types of operational losses, not just retail banking. A summary of this data is depicted in Figure 11. The figure shows the existence of an extreme asymmetry of operational loss severities; in fact, two important modes appear for losses less than €10 million and for losses larger than €100 million. It is assumed that banks that face large loss severities are likely to have large variances in their measurement errors when they record operational losses. Therefore, the dispersion in $\sigma_{\eta i}^2$ will be calibrated by the dispersion of loss severities documented in Benyon (2008).



Note: For each bank, average percentages are reported. Ranges of losses are expressed in millions of Euros. The continuous line is the histogram; the dashed line is a polynomial smoothing of the original histogram. Data were obtained from "Top 100 banks—A new dawn for disclosure" by D. Benyon in *OpRisk & Compliance*, *9*(10), 22-29. Copyright 2008.

Figure 11. Histogram of aggregate 12-month cumulative operational losses for October 2008 in banks in the ORX database.

After the shrinkage procedure is executed, the space of parameters to calibrate shrinks to 20. The mean value $\bar{\alpha}_j$ of the parameters in the operational loss (Equation 8) was calibrated to generate loss severities compatible with the 2012 summary of the database for retail-banking losses due to internal fraud (ORX, 2012). The mean value of the parameter $\bar{\beta}_o$ was set to pin down the frequency of losses documented in the ORX summary. Parameters $\bar{\rho}$, β_1, and β_2 determined the clustering pattern of operational loss shocks. According to Chernobai and Yildirim (2008), internal fraud losses exhibit low clustering as opposed to other type of losses. Hence, the calibration assigned relatively low values to these parameters in the range 0.01 and 0.1.

Two parameters measure speed of response. First, ρ_c measures how quickly internal controls are implemented to achieve a new level after new conditions arise and are expected to persist. Second ρ_q measures how fast the average ethical workforce achieves a new level when conditioning factors change with the expectation that they last. With respect to ρ_c, in broad terms, control levels do not change from one day to another, and some changes necessary to implement control adjustments may require budgeting, planning, and extra human resources. In a given year, the average worst scenario would be to wait half a year to implement full changes. Thus, if the number of business days in a year is about 260 and the number of working days to implement a new long-run control level is 130, approximately 65 days are necessary for implementing half the changes. Due to the auto-correlated nature of controls in Equation 11 and a half-life of $\tau' = 65$ days, the parameter ρ_c is then set to the value $\rho_c = 1 - \left(\frac{1}{2}\right)^{1/\tau'} \cong 0{,}011$. Half-life properties of autoregressive models have been outlined in Andrews (1993). Changes in the average level of ethical quality of the workforce may take even more time, and therefore, the parameter ρ_q would have to be lower than the benchmark of 0.011.

Two parameters need to be determined in the control Equation 11, namely the mean value of long run control levels $\bar{c}^*$ and the the mean sensitivity of control to loss ratio $\bar{\gamma}$. In ethical quality (Equation 12), there are four parameters to be set, two of them being the average labor productivity $\bar{A}$ across banks and the average number of employees per branch across banks $\overline{EB}$, which are readily estimable from the data. The other two variables, the average ethical quality sensitivity to factors $\bar{\delta}$ and the average level of ethical quality across banks $\bar{Q}$, as well as the average variance of measurement errors in the data recording process $\bar{\sigma}_\eta^2$, are explained below.

All the variables yet to be explained, grouped with the vector $(\bar{c}^*, \bar{\gamma}, \bar{\delta}, \bar{Q}, \bar{\sigma}_\eta^2)$, can be set in a way to generate sequences of loss severities and frequencies that have realistic properties documented in the literature. For example, Cope, Piche, and Walter (2012) studied how operational loss severities are associated with macroeconomic variables. In particular, Cope et al. found that internal fraud losses occur in countries where ORX member banks work are strongly and positively associated with the countries' legal frameworks that favor insider trading and are negatively associated with country-specific constraints on executive power in banks. Therefore, the sequence of loss severities generated in the simulation model, conditional on the parameters chosen, have to mimic a good association with these macroeconomic factors. Data for these macroeconomic factors are available in Cope et al.

In terms of the frequency of losses and their possible determinants, Chernobai et al. (2011) found that there is a strong and robust association between monthly frequency of losses and firm specific variables related to broad risk management conditions. In particular, there is a positive association between frequency of losses and equity volatility and a negative association between frequencies and Tier 1 capital ratios. Therefore, the frequency of losses generated in the simulation was contrasted with these indicators. The data for Tier 1 capital ratios are available from annual reports.

Finally, Moosa (2011) reported a strong association between average severities in the USA and the unemployment ratio to show that operational losses may be linked to the overall state of the economy. Thus, another important aspect in the calibration procedure was letting the average loss severities identified for countries in the simulated dataset be negatively associated with each country's GDP.

Overall, the procedure described helped calibrate the entire set of parameters in a meaningful and realistic way. The data simulated with the model was found to be compatible with micro and macroeconomic features and comparable with real datasets.

Data Integration Techniques

The key contribution of this research is the comparison of the three data integration techniques. This section elaborates on the techniques under evaluation.

Scaling technique. Following the example of Shi et al. (2000) and Na et al. (2006), operational losses depend on idiosyncratic and common factors

$$\hat{l}_{i,\tau} = \left(y_{i,\tau}\right)^{\psi} H(X_{\tau}) \tag{20}$$

The common component X_t refers to the statistical influence on losses caused by general factors such as macroeconomic, geopolitical, and cultural environments, and general human nature, among others. The idiosyncratic component $y_{i,t}$ is assumed to be deterministic and refers to more specific factors such as the size, income, or number of transactions in Bank i. Bank i cannot use losses $\hat{l}_{j,t}$ directly, but given the proportionality between losses given by $\frac{\hat{l}_{j,\tau}}{(y_{j,\tau})^{\psi}} = \frac{\hat{l}_{i,\tau}}{(y_{i,\tau})^{\psi}}$, it has to modify external lossess via the formula

$$\hat{l}^{s}_{j,\tau} = \left(\frac{y_{i,\tau}}{y_{j,\tau}}\right)^{\psi} \hat{l}_{j,\tau} \tag{21}$$

The parameter ψ can be estimated through simple OLS regression. Once the data were scaled, it could be used in the LDA process as described in Equations 2 and 3. To illustrate the process, Table 6 shows a hypothetical data matrix for an individual bank. Events are dated to the day and comprise the recording of the amount of losses in Euro in each event and the value of risk exposure indicators at the time of each event.

If the bank were to use only the data as outlined in Table 6, it would have only five observations for the frequency of events per year to estimate the parametric distribution of loss frequencies needed in the LDA. In addition, it would only have 12 loss severity events to estimate a loss severity PDF. With so few observations, the estimation of the densities and distributions to apply the LDA would be highly unreliable.

Table 6

Example of an Operational Loss Data Matrix in an Individual Bank

Year 1	Year 2	Year 3	Year 4	Year 5
Event 1	Event 1	Event 1	Event 1	Event 1
Event 2		Event 2	Event 2	Event 2
		Event 3		Event 3
		Event 4		
2	1	4	2	3

Note: All events are dated on the day of recognition and imply a severity amount. The last row provides the count of events per year.

With the Scaling technique, pooling the loss frequency information of *N* banks over a span of five years permitted up to 5×*N* data points to estimate the distribution function of counts. For example, with *N* = 41, the data sample reaches 205 and thus the estimation of a Poisson or Negative Binomial distribution for the frequency of losses is more reliable. The estimated distribution of loss event frequencies is denoted by $\hat{p}_n$. Moreover, all loss severities, previously scaled according to formula in Equation 12, could be pooled to have a larger dataset to estimate a PDF severity, which can be a Weibull, Lognormal, or other plausible known density. The estimated density of loss severities is denoted by $\hat{f}_z$.

Once both the frequency distribution $\hat{p}_n$ and the severity density $\hat{f}_z$ were estimated, the following Monte Carlo steps were in place to generate draws from the total loss per year:

1. Draw the random variable n for the number of times a loss occurs in a year from the distribution $\hat{p}_n$.
2. Draw n number of loss severities from $\hat{f}_z$: $z_1, z_2, \cdots, z_n$.
3. Calculate the total loss per year according to Equation 2: $S^{(1)} = \sum_{i=1}^{n} z_i$.
4. Repeat steps 1 to 3 until reaching a sufficient number J of simulated aggregate losses: $S^{(1)}, S^{(2)}, \cdots, S^{(J)}$.
5. Order the simulated aggregate losses from lowest to highest and obtain extreme percentiles from the simulated aggregate losses; for example, $OpRK_s^{99.9}$ used the 99.9 percentile for aggregate losses in applying the scaling technique.

Bayesian technique. Consider a random vector of internal data to a bank given by $X = (X_1, X_2, \cdots, X_D)$ with conditional joint density $h(x \mid \phi)$, where ϕ is a vector of random parameters and x is a realization of X. $h(. \mid \phi)$ can represent either severity or loss frequency data. Both x and ϕ are considered to be random variables with joint density, so if

$$h(x, \phi) = h(x \mid \phi)\pi(\phi) = \pi(\phi \mid x)h(x) \tag{22}$$

then, the Bayesian theorem implies:

$$\pi(\phi \mid x) \propto \pi(\phi)h(x \mid \phi), \tag{23}$$

where the symbol $\propto$ means the left-hand side of the equation is proportional to the right-hand side. $\pi(\phi \mid x)$ is the posterior density of parameters given the data, and $h(x \mid \phi)$ is interpreted as the likelihood of the data, and thus, it is a function of ϕ. This likelihood $h(x \mid \phi)$ typically refers to the Poisson or negative binomial distribution in the case of frequencies and to the Lognormal, Weibull, Pareto, or other densities in the case of severities. $\pi(\phi)$ is the prior density of the parameters. This prior density conveys all the information prior to the use of internal data; this information may be comprised of expert opinions about parameters, by external data to a firm, or by both.

The determination of prior densities means that it was necessary to elicit or estimate hyperparameters that defined the shape of these densities. Studies such as Lambrigger et al. (2007), Shevchenko (2011), and Shevchenko and Peters (2013) applied an empirical Bayesian approach to estimate prior densities based on external data, expert opinions, or both. Sub-Section 4.4 in Shevchenko (2011) provided a detailed explanation of the procedure to obtain the required prior.

Once the prior density was available, the Bayesian LDA necessitated building predictive densities for the planned year $(T + 1)$, based on the available internal data up to year T. The predictive densities are given by

$$h(x_{T+1} \mid X_T) = \int h(x_{T+1} \mid \phi)\pi(\phi \mid X_T)\, d\phi, \tag{24}$$

where $h(x_{T+1} \mid X_T)$ is the density of future values x_{T+1} conditional on the available internal data X_T up to time T. In the case of event frequencies in this study, this predictive density is denoted with $p_n(n_{T+1} \mid N_T)$, with N_T comprising the counts of events observed in all past years up to year T. Likewise, in the case of severities, the notation for the predictive density is $f_z(z' \mid Z)$, with z' representing the size of the next severity and Z the history of all severity levels observed in the bank. In practical terms, the Bayesian LDA could be implemented with the following steps:

1. Draw the relevant frequency parameter ϕ_n according to Equation 23. A sampling algorithm called the Gibbs sampler is necessary to perform this task, as detailed in Shevchenko (2011).
2. Draw the random variable n for the number of times a loss event will occur in the next year by using the likelihood $p_n(n_{T+1} \mid \phi_n)$. Shevchenko (2011) and Lambrigger et al. (2007) have provided details of how to draw random samples from $p_n(n_{T+1} \mid \phi_n)$.
3. Draw the relevant severity parameter ϕ_z according to Equation 23 and using the Gibbs sampler outlined in Shevchenko (2011).
4. Draw n random variables for the severity of losses from $f_z(z' \mid \phi_z)$. Shevchenko (2011) and Lambrigger et al. (2007) provided details of how to draw random samples from $f_z(z' \mid \phi_z)$. The *n* random variables are $z'_1, z'_2, \cdots, z'_n$
5. Calculate the total loss per year according to $S^{(1)} = \sum_{i=1}^{n} z'_i$.
6. Repeat steps 1 to 5 until reaching a sufficient number J of simulated aggregate losses with $S^{(1)}, S^{(2)}, \cdots, S^{(J)}$.

7. Order the simulated aggregate losses from lowest to highest and obtain extreme percentiles from the simulated aggregate losses. For example, $OpRK_B^{99.9}$ uses the 99.9 percentile for aggregate losses by applying the Bayesian technique.

Covariate-based Technique. This technique relied on estimating the distribution for the frequency of loss events and the density of severities using the entire internal and external data and conditioning the parameters on observed covariates.

If observations $\{\hat{n}_{i,T^*}\}$ are counts of occurrence of loss events $\{\hat{l}_{i,t}\}$ during a series of years T^* for all banks i, the assumption was that all these counts originated from the same parametric distribution f_n with parameters varying as functions of observed covariates. A possible form of this function was the Poisson distribution:

$$f_n(\hat{n}_{i,T^*}) = \frac{(\lambda_i)^{\hat{n}_{i,T^*}} \exp(-\lambda_i)}{\hat{n}_{i,T^*}!}, \tag{25}$$

where the key assumption was that λ_i depends on observable covariates

$$\lambda_i = \lambda(\bar{y}_{i,T^*}, \bar{e}_{i,T^*}), \tag{26}$$

where the variables on the right-hand side of Equation 26 represented averages of the observed covariates for years T^* and by bank.

In this study, the advice of Ganegoda and Evans (2013) was followed. Each loss event $\hat{l}_{i,T^*}$ comes from a density function $f(\hat{l}_{i,T^*}; \Theta_i)$ conditional on the set of parameters Θ_i. Ganegoda and Evans focused on a vector of two parameters $\Theta_i = (\mu_i \quad \sigma_i)'$. Each of these parameters was linked to covariates through link functions

$$g_1(\mu_i) = Z_1\omega_1$$

and

$$g_2(\sigma_i) = Z_2\omega_2,$$

where Z_1 and Z_2 denoted the covariates that affect the distributional parameters and the ω's were the corresponding sensitiveness coefficients. The covariates included internal

and external data. The idea was to estimate the maximum likelihood estimator of parameters ω_1 and ω_2 through

$$\max_{\omega_1,\omega_2} \sum_{j=1} f(\hat{l}_{j,T^*}; \omega_1, \omega_2) \tag{27}$$

After estimating ω_1 and ω_2, it was straightforward to condition the severity density function to the specific covariates of the financial institution under study and perform the LDA to obtain the quantity of interest $OpRK_C^{99.9}$.

Summary

In this chapter, the design for the research was elaborated upon by describing the steps taken in the implementation of the equation calibrations applied in Chapter 4. The research design applied a simulation-based approach common in the statistics literature (Greene, 2012; Stern, 2000; Voss, 2013). The approach involved three main implementation steps: (a) Data simulation through an internal fraud model, (b) application of data integration techniques, and (c) a simulation-study evaluation about which data integration technique delivered a level of operational risk capital closer to the true operational risk implied by the data simulation model.

Each of the steps implemented was covered in some detail. The data simulation model resembles work on dynamic operational risk modeling developed in Kühn and Neu (2004), Leippold and Vanini (2005), and Bardoscia and Bellotti (2011), but for the most part, incorporated innovative ideas to model internal fraud losses that have not yet been studied in the literature. A relevant element of the process was the calibration of model parameters to simulate operational loss data that looked like real data in terms of statistical correlations with observable variables. The data integration techniques applied the techniques discussed in the literature review. The Scaling technique was based on the work of Shih et al. (2000) and Na et al. (2006), the Bayesian technique on the work of Lambrigger et al. (2007, 2008), and the

Covariate-based technique on the work of Ganegoda and Evans (2013).

In the next chapter, the full implementation of the research design is documented. Being a quantitative design, the implementation relied upon software codes developed in the R programming language to carry out all the design steps described in this chapter. The R language is a free software environment for statistical computing and graphics (R Core Team, 2016).

The database of key operational risk indicators, as outlined in Table 5, for banks listed in Table 4 was used to test the model developed. All necessary data described in the calibration procedure were gathered in order to apply the calibration design outlined. The calibrated model was used to simulate operational loss data across all member banks for the years 2006 to 2010 and to simulate data for a specific bank. The five-year data across all banks were used in the data integration techniques outlined in this chapter to produce operational risk capital levels $OpRK_i$, for $i = S, B, C$. The simulation of a large enough operational loss data sequence for a specific bank helped to determine the true operational risk profile for that bank and thus helped determine the true $OpRK_{true}$.

Last, in Chapter 4, a simulation study to determine which of the data integration techniques are the most valid and reliable was applied. Justification for this step was explained in detail. The final step and the focus of the next chapter was answering to the research questions motivating this study. In particular, how are operational losses related to the 2007-2009 Global Financial Crisis? (MRQ1), how are internal fraud losses related to perceptions about corruption in the country where the main headquarters of a bank is located? (MRQ2), and is there any technique that can be considered best practice to estimate a correct operational risk capital across all levels of risk tolerance? (MRQ3).

Chapter 4: Results

In this chapter, the set of results obtained by applying the quantitative procedure outlined in the previous chapter is described. First, a thorough description of the dataset is given. Second, the results of the calibration procedure to set up the model to make operational loss simulations for internal fraud in retail banking are reported. Third, simulations that capture the loss profile are performed, given the environment and conditions that banks faced during the period 2006-2010 (See Table 4). These conditions are both idiosyncratic and global. Fourth, with the simulated data for each bank and their corresponding conditioning factors, the data combination techniques defined in Chapter 3 are implemented. Fifth, the most important results of the study, the comparison of the different techniques with reference to the benchmark simulation, is presented. Last, the questions are answered and a conclusion to the analysis reached.

The Data

Because the thesis is concerned about operational losses due to internal fraud in retail banking for a set of banks belonging to the ORX data exchange, it was necessary to work with quantitative data on this particular operational loss as well as all the possible risk indicators per bank and per country of each bank. These key risk indicators refer to idiosyncratic and macroeconomic or other factors that affect the particular type of loss under study.

Most but not all the data necessary to perform the analysis belongs to the ORX data exchange (Appendix C contains a full list of the data and their sources). Access to this dataset is not possible unless the research is conducted from within one of the banks that belong to the ORX exchange. The dataset is proprietary, which means that a researcher cannot make it public for scrutiny or replication.

Instead, the data gathered for the analysis performed relied entirely on public information. Specifically, the analysis was based on data downloaded from websites of each of the banks belonging to the ORX exchange that perform retail operations. All banks publish their annual reports and financial statements each year and sometimes more often. These reports do not contain information about operational losses but contain most of the idiosyncratic data needed to infer operational losses. In Table 7, the type of data collected from each of the reports or financial statements is summarized.

Table 7

Data Gathered from Public Sources about Banks in the ORX Exchange

Key	Concept	Measure
No	Number of bank	index
bname	Bank name	index
country	Bank headquarters' country	index
code	Country and Bank Code	index
year	Year	index
ccy	Report's Currency Code	text
branches	Number of branches and offices (Retail)	count
staff	Number of staff (total)	count
staff_r	Number of staff (Retail)	count
loans	Total loans to customers	Currency millions
loans_r	Total loans to customers in retail banking	Currency millions
assets	Consolidated assets	Currency millions
assets_r	Assets in retail banking	Currency millions
tier1	Tier 1 capital	Percent
nic	Net interest income (total)	Currency millions
nic_r	Net interest income (retail)	Currency millions

Note: The main sources are the public annual reports and financial statements posted on banks web pages; also, in some cases, Form 10K of SEC filings (EDGAR database) for banks that operate in the USA were used.

Figure 12 provides a brief description of the dataset. The figure shows pairs of scatterplots between the numbers of branches, the number of staff related to retail banking

operations, the level of retail loans, and the value of retail assets. All currency amounts were converted to millions of Euros. Data comprised five years for all 52 banks considered.

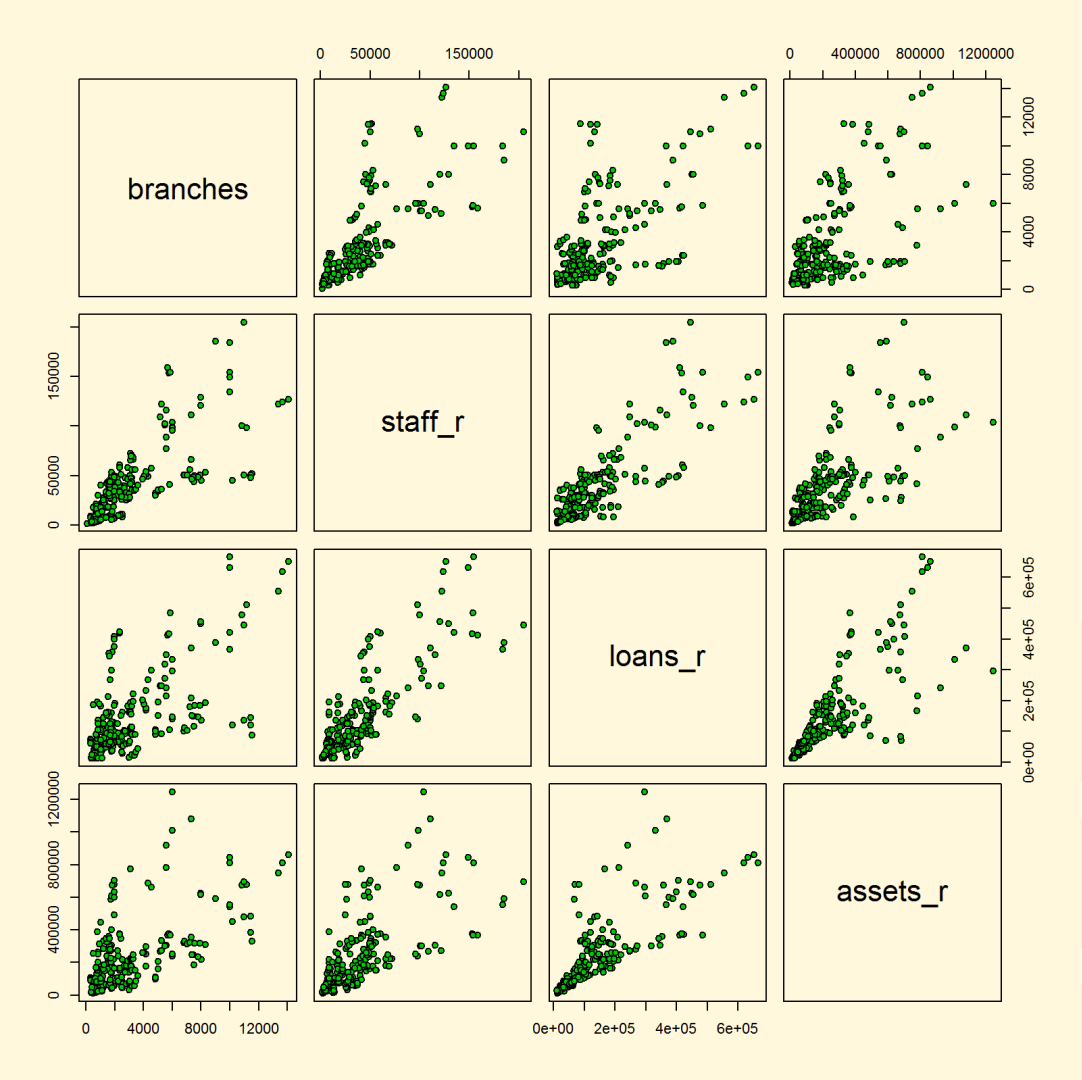


Figure 12. Scatter plot of bank data per year.

All variables considered in Figure 12 are indicators of the scale of operations in the retail banking segments of each bank. This explains the remarkable positive correlation that emerges (between 0.62 and 0.82). These scale indicators belong to the set of risk indicators that likely induce the appearance and severity of internal fraud losses (Appendix D provides more graphical description of these indicators). This idea is made operational later in the chapter through applying a quantitative model that mapped these scale indicators towards the outbreak of losses.

The calibration procedure needed more specific risk indicators. Therefore, the analysis

relied on other forms of risk indicators that could be collected from annual reports. Textual content was useful to calibrate sensitivity parameters in the operational risk model outlined in the previous chapter. Table 8 shows the textual context variables extracted from the annual reports. The variables refer to the number of instances a descriptive key word or phrase appeared within the entire text; also, the total number of report pages is recorded in order to calculate the ratio of instances to number of pages. These ratios give an indication about the relative importance of a key word that banks use in their public reports.

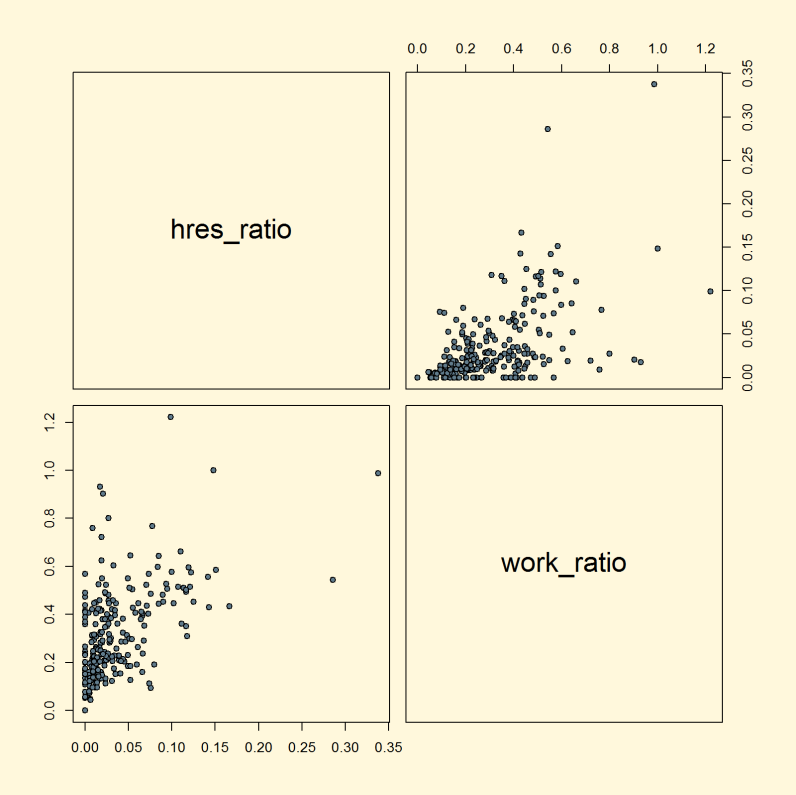
Table 8

Variables Contained in the Textual Database

Key	Concept
nbank	Number of bank in database
bname	Bank name
country	Country of bank headquarters
code	Bank Code = country code.bank
year	Year
orisk	"Operational risk" instances in annual reports
risk	"Risk" instances in annual reports
rman	"Risk management" instances in annual report
ama	"AMA" (Advanced Measurement Approach) instances in annual reports
hres	"Human resource" instances in annual reports
emp	"employee" instances in annual reports
Col	"colleague" instances in annual reports
workers	Sum of "employee" and "colleague" instances
npag	Number of pages in the Annual Report

Both panels in Figure 13 show scatter plots of textual variables. Panel A shows scatterplots of “human resource” paired with the sum of “employee” and “colleague” instances in annual report texts as a proportion of total number of pages in each report.

Panel A:



Panel B:

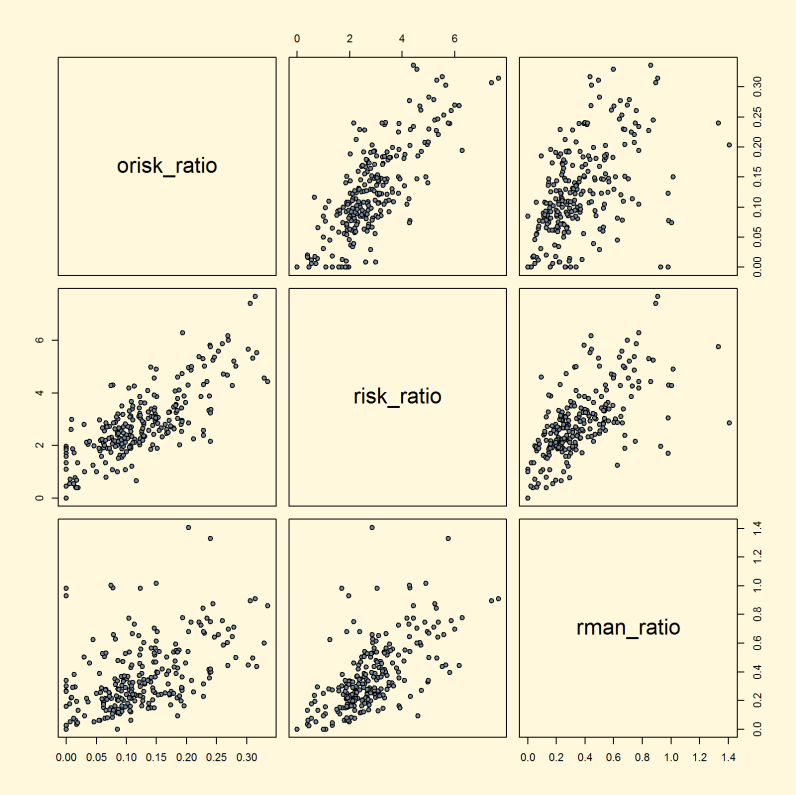


Figure 13. Scatter plot of textual context variables.

Panel A contains variables related to human resource management while Panel B contains variables related to risk management.

Panel B shows the paired scatter plot of the triplet of risk, risk management, and AMA instances as a proportion to total pages. All plots show positive correlations. This means that the calibration procedure worked with one variable from each panel. From panel A for example, the human resource ratio is useful to signal how important the concept of human resource management is for the bank. From Panel B, the calibration can use the risk management ratio, which proxies banks' awareness of risk management in the company. These two variables allowed the calibration of a number of parameters in the model.

In terms of global variables that affect all banks or groups of banks, the analysis incorporates variables such as growth rates of gross domestic products (GDP) of the countries to which banks belong. The dataset also contained a number of variables that could affect the outbreak of losses due to internal fraud such as the rule of law in a country or its corruption perception index.

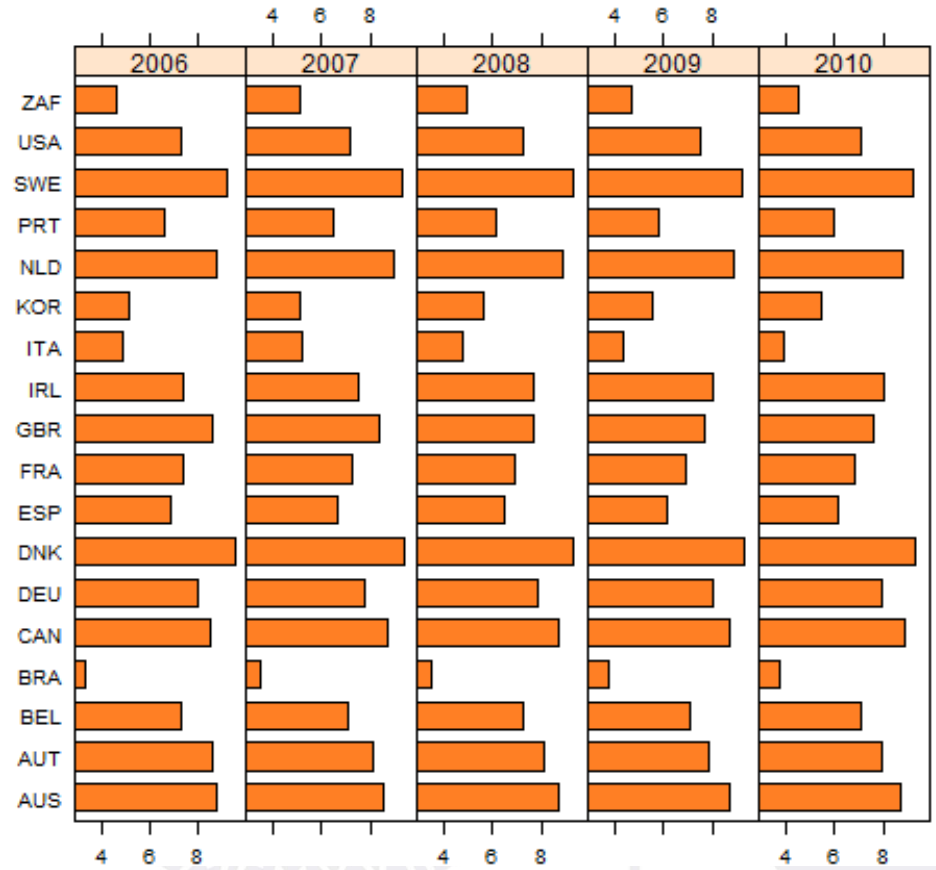
Table 9

Variables Contained in the Macroeconomic Database

Key	Indicator name
country_name	Country Name
country_code	Country Code
year	Year 2006-2010
gdp_growth	GDP growth (annual %)
crisis	Financial crisis dummy variable for 2007 and 2008
gover_effective	Government Effectiveness
reg_quality	Regulatory Quality
rule_law	Rule of Law
cont_corrup	Control of Corruption
cpi	Corruption Perceptions Index (CPI) score (2006-2010)

Figures 14 and 15 summarize the data for GDP growth and the corruption perception index. Figure 14 shows the corruption perception index (CPI) for each of the relevant countries during the years of analysis. In the data, Brazil and Italy are shown to have higher

corruption perceptions while countries like Denmark, Sweden, Netherlands, and Canada are seen to be less likely to be corrupt. The perception of corruption is possibly associated with real corruption levels, and the extent of corruption can affect the occurrence of fraud internal or external to banks because they are related to the cultural environment that rationalizes frauds according to Cressey's triangle.

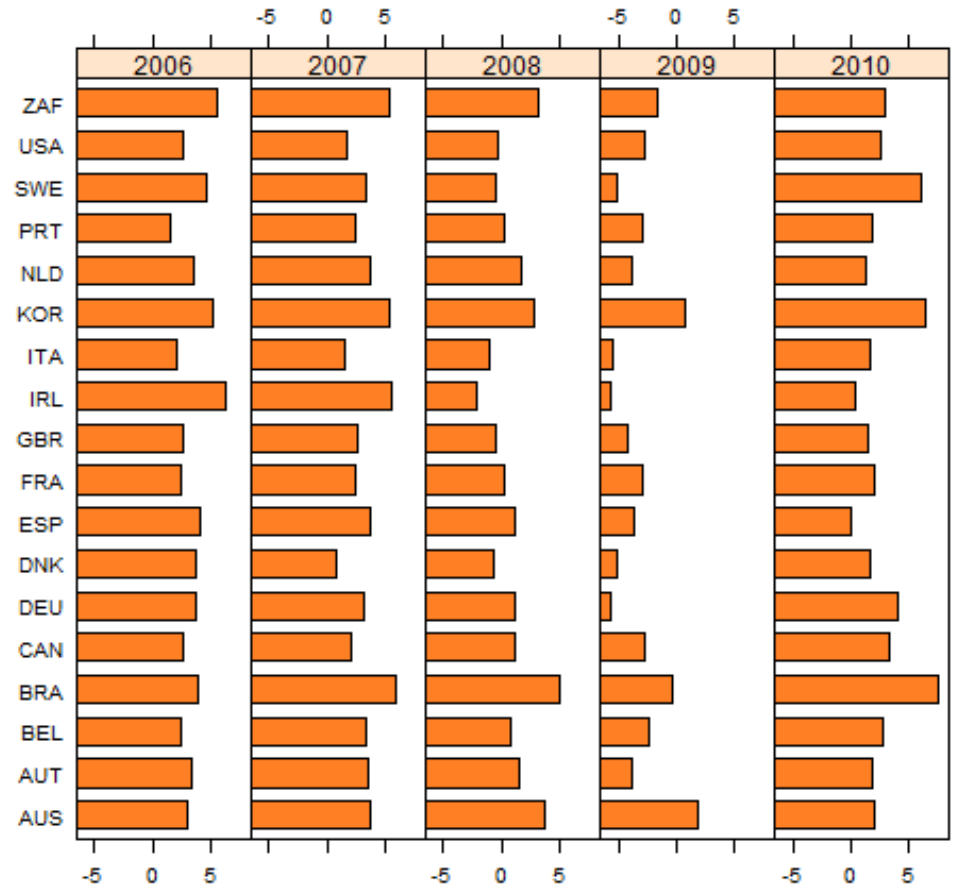


Note: The lower the index, the more a country is perceived to be corrupt. The data were based on figures released by Transparency International.

Figure 14. Corruption perception index (CPI) for countries where banks have their main headquarters.

Figure 15 shows GDP growth rates for the relevant countries and years under analysis. The association between the growth rate of economic activity as a whole and the occurrence of fraud events is not clear. Stewart (2016) shows evidence that fraud losses in US banks increase with the economic cycle while Abdymomunov, Curti and Mihov (2015), using operational loss data in US banks find that internal fraud operational loses are negatively associated with GDP growth. As will be seen in the development of the Scaling and the

Covariate-base technique, the association between economic activity and the outbreak of loss events is positive according to the simulations. This means that banks located in countries where there is strong growth are marginally more prone to have internal fraud losses, and those losses are likely to be more severe. Macro-level data do not explicitly help to calibrate the model for internal fraud loss outbreaks. Instead, the information provided by macro-level data is used in the second part of the simulation process, to condition the different data combination techniques to macro-level risk indicators.



Note: Data based on figures released by the World Bank.

Figure 15. GDP growth in countries where banks have their headquarters.

A good measure of the fitness of the operational risk model to generate internal fraud losses is whether those losses are associated to the macro-risk indicators in the same fashion as documented in empirical research as shown for example in Chernobai et al. (2011), Moosa (2011), Cope et al. (2012), Abdymomunov et al. (2015) and Stewart (2016).

Calibration Results

Table 10 shows the value of the general parameters that set the behavior of the equations in the operational loss model. The setting of these parameters applied the calibration procedure outlined in Chapter 3. The target of the calibration was to allow the model to simulate losses as close to reality as is possible. The only reality check available was to mimic the mean frequency and severity of losses due to internal fraud in the retail segment across the banks belonging to the ORX data exchange for the period 2006-2010. Therefore, the calibration procedure used an optimizing framework to pin down the mean parameters of the loss equation described in Table 10.

Table 10

Calibration of Parameters

	Definition	Value	Equation
$\bar{\alpha}_0$	Mean scale parameter	0.400	Loss outbreaks
$\bar{\alpha}_1$	Mean constant within ramp function	16.810	Loss outbreaks
$\bar{\alpha}_c$	Mean impact of controls	-275.291	Loss outbreaks
$\bar{\alpha}_y$	Mean impact of gross operating income	-1.587	Loss outbreaks
$\bar{\alpha}_q$	Mean impact of quality of workers on losses	0.052	Loss outbreaks
$\bar{\rho}$	Mean autocorrelation of loss shocks	0.70	Loss shocks
ρ_{min}	Lower threshold for loss shocks autocorrelation	0.50	Loss shocks
ρ_{max}	Upper threshold for loss shocks autocorrelation	0.90	Loss shocks
$\bar{\beta}_0$	Mean constant term	0.20	Shock variance
β_1	Influence of past quadratic shocks	0.01	Shock variance
β_2	Influence of past variance	0.70	Shock variance
ρ_c	Weight of new conditions to affect controls	0.10	Loss control
$\bar{c}^*$	Mean long run value of control level	0.5	Loss control
$\bar{c}_{min}$	Lower threshold for long run control level	0.3	Loss control
$\bar{c}_{max}$	Upper threshold for long run control level	0.7	Loss control
$\bar{\gamma}$	Sensitivity of controls to the losses	-0.5	Loss control
λ	Desired loss ratio	0.0003	Loss control
ρ_q	Sensitivity to recent ethical quality	0.05	Ethical quality
$\bar{\delta}$	Determines the sign of impacts from factors	0.2	Ethical quality
$\bar{Q}$	Level of average ethical quality across banks	0.7	Ethical quality
$\bar{\sigma}_\eta^2$	Variance of measurement errors	0.01^2	Measurement error
l_i^{min}	Threshold level for operational loss reporting	20	Reporting

Note: The shrinkage procedure uses the parameters denoted with overbars. An optimizing search procedure determines the α parameters.

The optimizing framework hinged on minimizing the quadratic distance between the observed mean loss severity and the simulated mean loss severity. In addition, the optimization puts weight on the fact that almost all banks in the dataset must face losses. In reality, a bank with no operational losses in the pace of five years is rare. The parameters that affect banks in an idiosyncratic way were determined by the shrinkage procedure described in Chapter 3. Figures C1 and C2 in Appendix E depict the distribution of these parameters across banks. The idiosyncratic parameters calibrated through this procedure therefore serve as a useful device to control for the heterogeneity observed in the banks in the ORX sample.

Simulation Results

The analysis proceeded with the simulation of both operational losses and the rest of the risk indicators considered in the operational risk model outlined in Chapter 3. The simulation was performed with programs in the R software (R Core Team, 2016) written specifically for this study. The simulation hinged on generating 500 alternative histories of operational losses for the years 2006-2010 within the banks in the ORX database. The simulations considered the specific conditions banks have confronted during the five-year period in terms of their own risk exposure and the macroeconomic environment surrounding them. After the simulation, it was possible to calculate the gross amount of operational losses as well as the number of losses across banks. The 500 data points are drawn in Figure 16, where the straight lines mark the values reported in ORX (2012).

Each point in Figure 16 summarizes a possible five-year history of data in each of the 52 banks. Each bank has 500 possible histories of operational losses conditional to the circumstances in effect during those five years. Ideally, each bank can take their own 500 histories and combine them directly to estimate their true risk exposure. This type of exercise delivers the true operational risk capital for each bank as depicted on the right hand side of Figure 1.

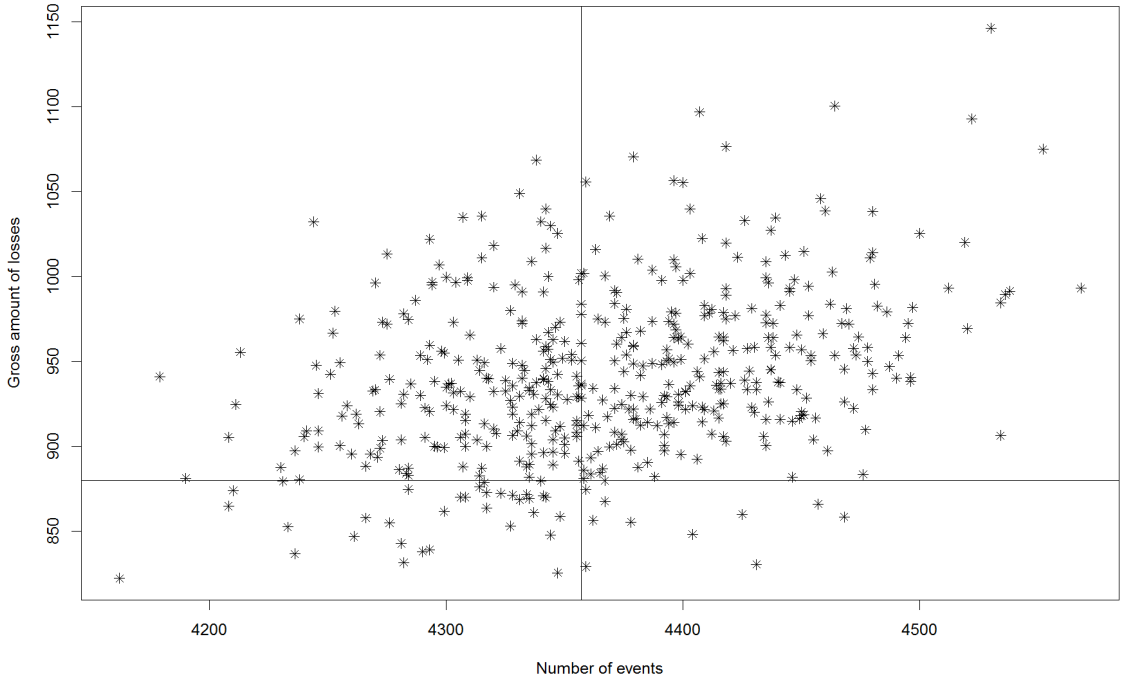


Figure 16. Summary of simulations and comparison to ORX report.

True operational loss. This step applied the AMA procedure described in Chapter 3. The analysis uses the fitdistrplus package in R described in Delignette-Muller and Dutang (2015). This R package allows for the estimation of frequency and severity distributions through maximum likelihood methods. Thereafter, this study will provide a detailed report of the results for the Commonwealth Bank of Australia (AUS.CBA) but the simulation-study exercise is performed for each possible bank in the ORX dataset. For all banks, the estimation of the true risk profile pools their 500 simulated internal histories and estimates the best distribution for the frequency of losses and the best density for the severity of losses.

For the frequency distribution, the Negative Binomial turns out to perform best in all the cases against the Poisson. Figure 17 depicts, on the left panel, the negative binomial density and CDF and the empirical distribution obtained for AUS.CBA.

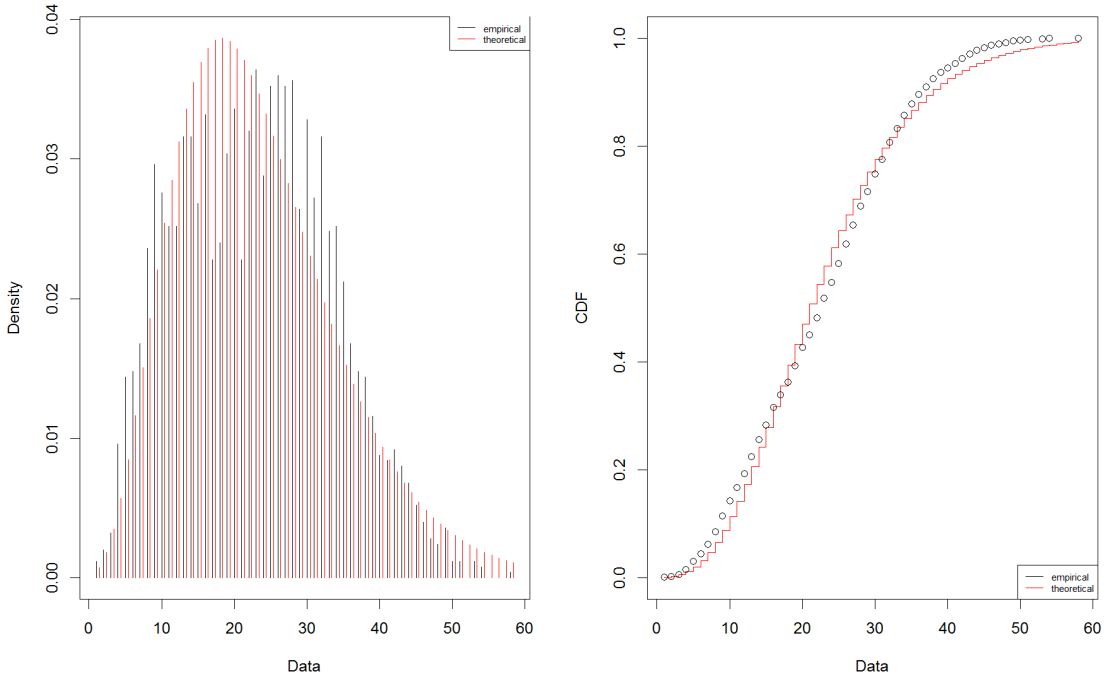


Figure 17. Estimation of the distribution of frequency

The right panel shows the cumulative probability distribution. The Negative Binomial distribution provides a good fit to the frequency data. The mean parameter is rounded to 23, which means that for AUS.CBA, there were 23 loss events due to internal fraud in retail banking in a typical year. The Negative Binomial distribution is a better fit than the Poisson distribution because the Poisson parameterizes the shape of the distribution with only one parameter. Table 11 shows the maximum likelihood estimation of the parameters of both distribution functions: the Negative Binomial and the Poisson.

Table 11

Maximum Likelihood Estimation of Frequency Distributions

	Estimate	Standard error
Parameters of the negative binomial distribution		
r (size)	5.042	0.176
μ (mean)	22.963	0.226
Parameters of the Poisson distribution		
λ (mean = variance)	22.964	0.096

The Weibull density better described the probability density for the severity of operational losses. Figure 18 shows the comparison between the Weibull, the Lognormal, and the Gamma densities. From the Q-Q plot and the other plots, it is observed that the Gamma density performs worst at the right extreme of the distribution. The Gamma density assigns too much probability at medium levels of severity and less probability at the extreme right. Both the Lognormal and Weibull behave similarly; they are closer to the empirical distribution than the Gamma is, even for extreme values. Overall, there is a marginal advantage for the Weibull as seen in the Q-Q plot.

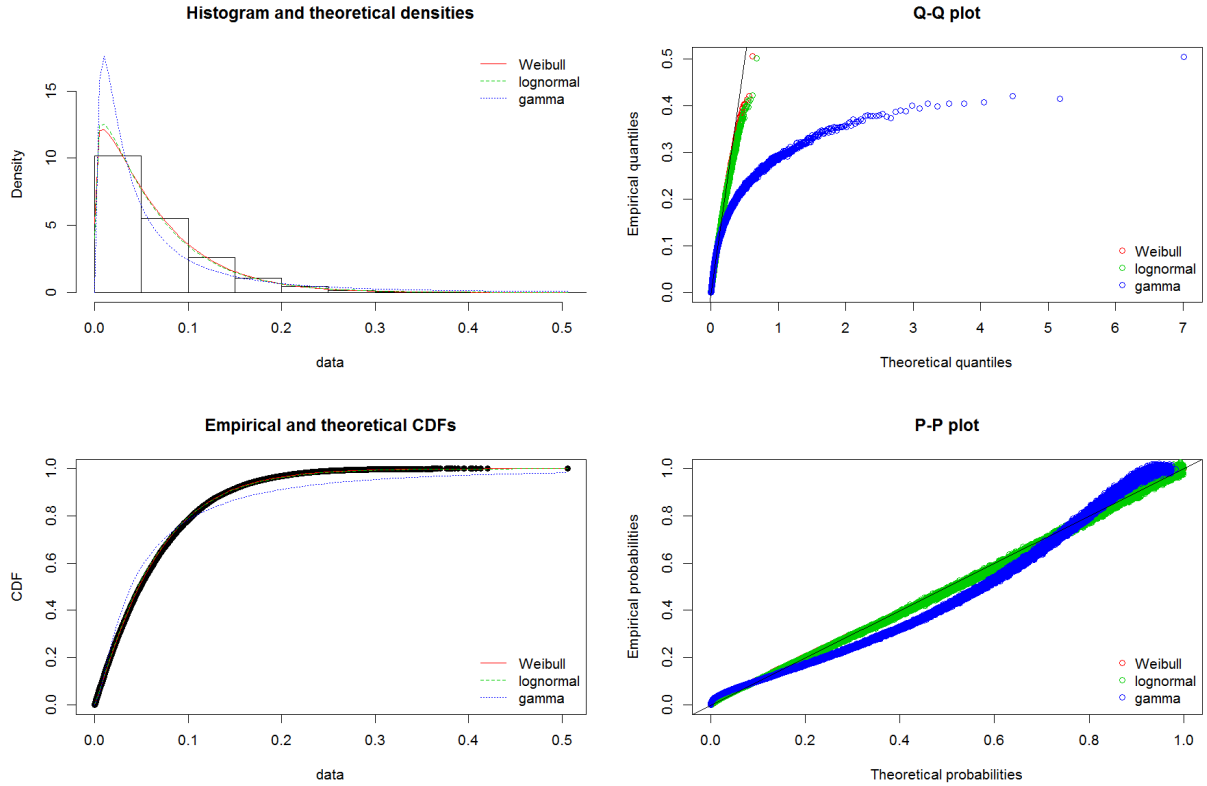


Figure 18. Estimation of the PDF for the severity of losses.

Table 12 depicts the parameters of the three density estimations. The analysis that follows used the shape and scale parameter of the Weibull density. The mean of the Weibull distribution is $\mu = scale \times \Gamma(1 + 1/shape)$ where $\Gamma(\quad)$ is the Gamma function. This means that if there was a loss event, its expected severity was 0,06 millions of Euros.

The implementation of the LDA considered the best performing models for both the frequency and severity of losses. In particular, the LDA used the estimated Negative Binomial distribution to model the frequency and the estimated Weibull density to model the severity of internal fraud losses.

Table 12

Parameters of the Severity Density Estimations

	Estimate	Standard error
Parameters of the Weibull density		
Shape	1.103	0.0036
Scale	0.066	0.0003
Parameters of the Gamma density		
Shape	1.138	0.0060
Rate	17.777	0.1165
Parameters of the Lognormal density		
Meanlog	-3.249	0.0050
Sdlog	1.210	0.0036

Implementation of the LDA was straightforward in the R software environment. The procedure consisted on applying the convolution operation explained in Chapter 3 by means of a Monte Carlo simulation. Figure 19 shows the graph of the empirical PDF and CDF of total annual losses obtained from the Monte Carlo procedure for 10,000 hypothetical cases. In fact, the procedure is used to obtain the entire forecast distribution of annual losses for planning year 2011 based on historical information of the last five years (2006-2010).

The calculation of operational risk capital always needs a huge number of simulations because risk management assumes a very high percentile of the distribution. To be able to obtain accurate estimates of extreme percentiles, the number of simulations needed is in the order of thousands. The forecast of interest for risk management is not the expected annual loss but an extreme value according to the risk tolerance of the risk manager.

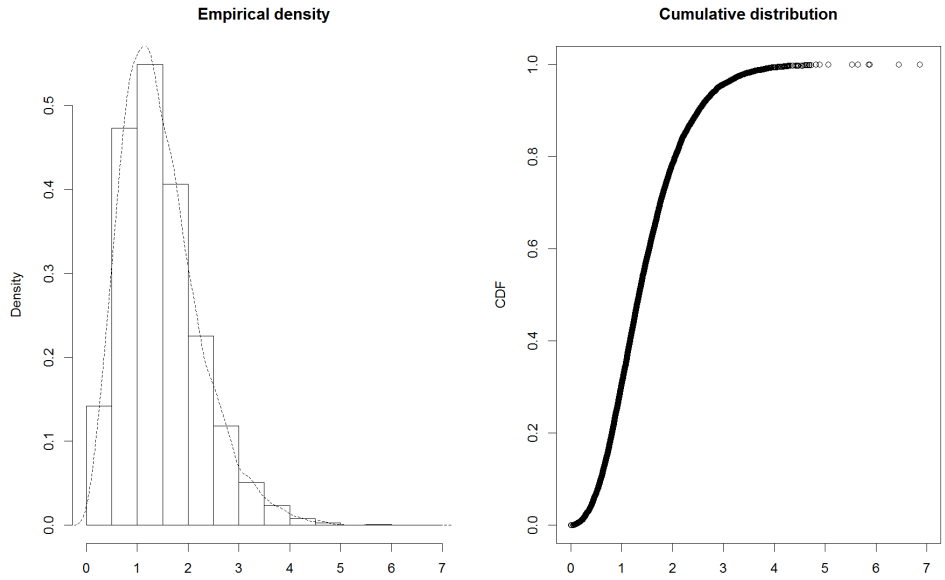


Figure 19. Empirical PDF and CDF of total annual losses.

The loss density estimated has implications for the operational risk capital because the procedure allowed extracting the least likely losses. For example, according to Table 13, a gross annual loss larger than €3.8 million is only 1% probable and a loss larger than 5.45 million is only 0.01 % probable. If a €5.45 million or larger loss actually occurs, it would be catastrophic for the bank, unless it has enough capital to cover up the losses. Risk managers at banks face a dilemma: If they hold low capital levels to face operational risk losses, they may improve their current profits because holding liquid capital is costly. However, a catastrophic event may hit the bank and make it insolvent due to lack of sufficient capital. On the other hand, if the bank holds high capital levels, it is prepared to face large but unlikely losses, but if the loss event does not materialize, current profitability is damaged.

Table 13

Operational Risk Capital Levels at Different Percentiles (Bank AUS.CBA)

Percentile	Value (millions of Euros)
99.00	3.77
99.50	4.08
99.90	4.87
99.99	5.45

The optimal operational risk capital level depends on the proper weighing of the above dilemma, which in turn, depends on the degree of risk tolerance in the organization. Therefore, Table 13 provides the possible choices risk managers could make.

This same exercise is performed for all the banks in the ORX dataset. Figure 20 shows the ratio of true operational risk capital levels or VaR at the 99.9 percentile as a percentage of total assets in each of the banks. The resulting percentages are small ranging from almost 0 to 0.06 percent. Bank IRL.BIG is not in the graph because its operational loss capital ratio reached up to 0.7 percent. Still, this operational risk capital levels are small because the present study only considers internal fraud losses in retail banking. It does not consider the entire operational loss spectrum.

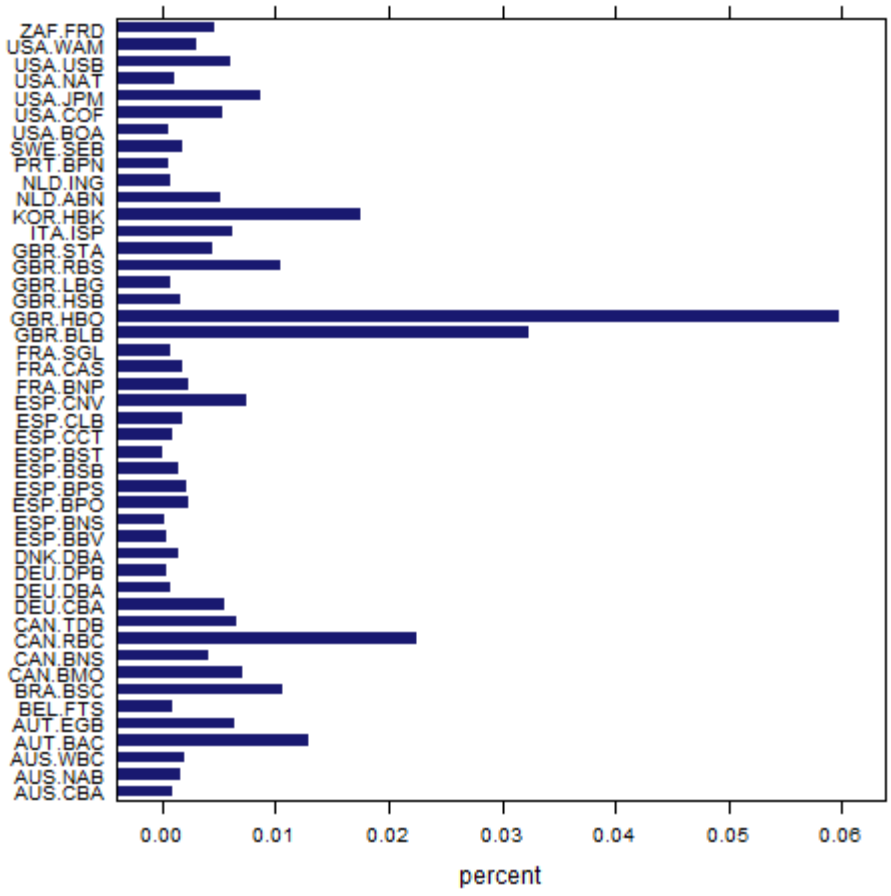


Figure 20: VaR levels at 99,9 percentile as a percentage of total bank assets.

Data sharing procedure. As explained in Chapter 3, simulations for one bank across possible histories are homogenous but simulations across bank for a specific five-year history are heterogeneous because each bank faces different risk exposures.

As noted in Chapter 3, banks need to share their data in order to process and use the data in their operational risk capital level calculations. The ORX data exchange is one of these sharing schemes for banks. Therefore, given the loss simulations and their associated risk exposures, the analysis applied Equation 15 as determined in Chapter 3 and gathered the value of all conditioning idiosyncratic factors as well as observed macro variables at the time of each loss event for all the banks.

Data Aggregation Techniques

As described in Chapter 3, there are three data aggregation techniques. This section deals with the implementation of these three techniques based on the shared simulated database generated in the previous section. For comparison purposes, this section will describe the procedure in bank AUS.CBA but at the end, summary results for all the banks will be shown. In each bank, the three techniques deliver operational risk capital levels which were compared to their true operational capital levels found in the previous section.

Scaling technique. In applying this technique, a regression equation as discussed in Chapter 3 first needs to be estimated. The technique follows the example of Shi et al. (2000) and Na et al. (2006). Table 14 shows the baseline regression results (see Equation 20) performed with data taken to the annual frequency. This means that before performing the regressions, it was necessary to calculate mean severity of losses, mean operational control levels, and mean values of the risk ratios, like employees per branch, for each year.

Results show that the ratio of employees per branch has a statistically significant effect on operational losses due to internal fraud in retail banking. The number of employees per branch does not directly affect losses in the model specification (Equation 8). The effect

of the number of employees per branch on the severity of losses may be due to the indirect effect that comes from the effect of the number of employees in the calibration of the parameters.

Table 14

Dependent Variable: Log of Operational Losses

Regressors	Estimates	Std. error	*t*-value	*p*-value	
Intercept	-5.06	0.22	-23.37	0.00	***
Log(employees per branch)	1.32	0.06	22.18	0.00	***
Corruption perception index	-0.01	0.03	-0.26	0.79	
GDP growth	0.04	0.02	2.35	0.02	*
Dummy for crisis (2008-2009)	0.17	0.10	1.80	0.07	.

Note: Significance codes: 0=***, 0.001=**, 0.01=*, 0.05=.; Residual standard error: 0.4866 on 217 degrees of freedom. Multiple *R*-squared: 0.725, Adjusted *R*-squared: 0.720, *F*-statistic: 142.8 on 4 and 217 *DF*, *p*-value: < 2.2e-16.

Other country-specific variables affected average annual losses per bank like the control of corruption perception index and GDP growth. Although it is not significant, the CPI has the correct sign: The larger the value of the index (the less corrupt a country is), the lower the size of average losses due to internal fraud.

When a country to which a bank belongs grows faster, it induces more internal fraud losses in the bank. Both results linked to country-specific variables supported the simulation model for internal losses. Results were compatible with Povel, Singh, and Winton (2007) and Stewart (2016) who posited that fraud is more likely to occur in good macroeconomic contexts.

Another relevant result is that a period of financial crisis implies more internal fraud losses. According to Hess (2011), the 2007-2009 Global Financial Crisis increased the risks for loss severity for two business lines in banking: Trading and sales as well as retail brokerage. This is another indication that the internal fraud simulation model produces loss features that are compatible with existing evidence. This result is also important for

answering the first research question (MRQ1) about the effect of the financial crisis on losses.

Once the regression equation was estimated, the next step was to scale the external losses as noted by Dahen and Dionne (2010). The regression equation is

$$\log(loss_{it}) = a_0 + a_1 \log(eb_{it}) + a_2 \mathrm{CPI}_{it} + a_3 GROWTH_{it} + a_4 \mathrm{CRISIS}_{it} \qquad (28)$$

Where eb_{it} is the employee per branch ratio; CPI_{it} is the Corruption Perception Index in the country to which Bank i belongs; $GROWTH_{it}$ is the GDP growth rate of the country to which Bank i belongs; and $\mathrm{CRISIS}_{\mathrm{it}}$ is the dummy for the crisis period, which is common for all banks and countries. Let $a_0 = \log(Common)$ be the common component and let $\log(idiosync_{it}) = a_1 \log(eb_{it}) + a_2 \mathrm{CPI}_{it} + a_3 GROWTH_{it} + a_4 \mathrm{CRISIS}_{it}$ the idiosyncratic part, which varies across banks and over time.

To compare any two banks, the analysis considered that the common component is

$$Common = \frac{loss_{it}}{idiosync_{it}} = \frac{loss_{jt}}{idiosync_{jt}} \qquad (29)$$

This proportionality between any two losses at different banks allowed scaling losses at any bank j to be comparable to losses at bank i. The scaling follows directly from Equation 29

$$loss_{it} = \left(\frac{idiosync_{it}}{idiosync_{jt}}\right) loss_{jt}, \qquad (30)$$

which means that losses at bank j are multiplied by the factor $\frac{idiosync_{it}}{idiosync_{jt}}$ to be comparable to losses at bank i. Upon scaling all banks' losses against losses at bank i (AUS.CBA), it is possible to use all loss data to perform the LDA to calculate the operational risk capital.

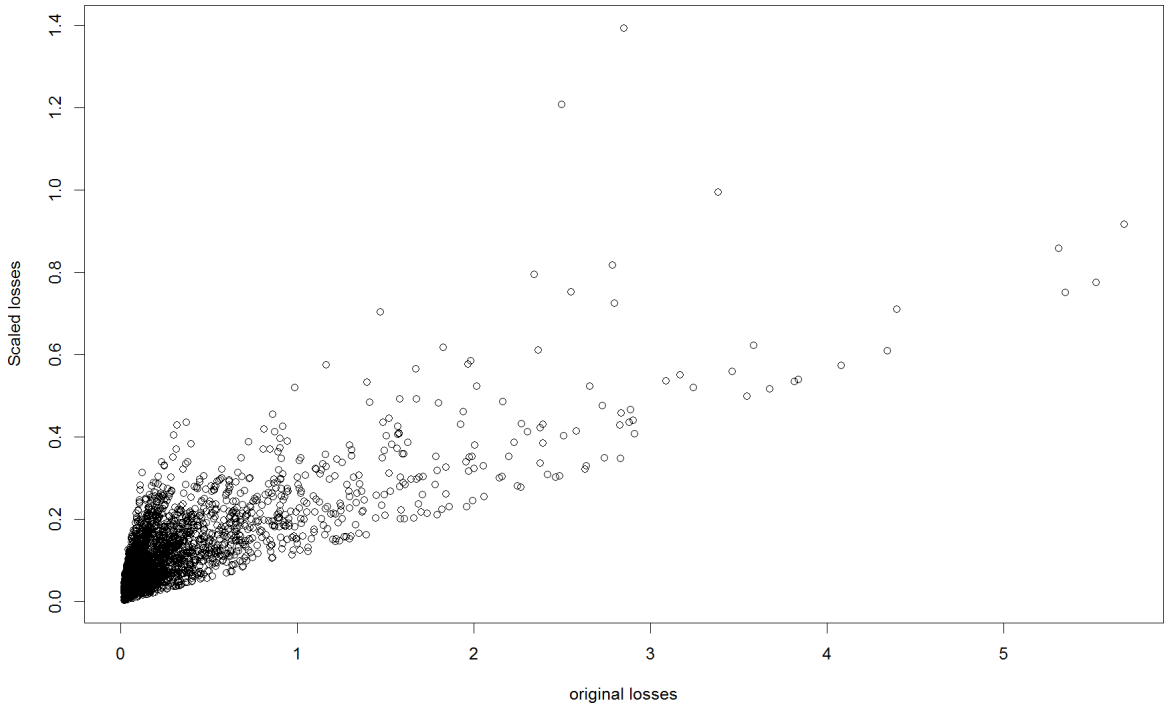


Figure 21. Bank AUS.CBA: Original losses (horizontal axis) vs. scaled losses (vertical axis).

Figure 21 provides an overview of the scaling procedure; original losses at all banks can go up to a little less than €6 million. Scaled losses that are comparable to bank AUS.CBA have a range of between €0 and €1.4 million.

Regarding the frequency of losses, there are two alternative procedures. The first alternative is to scale the frequency of operational losses in all banks to be comparable to the frequency at bank AUS.CBA. A similar regression procedure and scaling factors could be applied with this option. The second alternative is to follow the example of Dahen and Dione (2010) to fit Truncated Negative Binomial or Truncated Poisson regressions and determine the parameters of the frequency distributions conditional on certain types of covariates. Because the second alternative bears more resemblance to the Covariate-based technique, the first alternative of scaling the frequency with a new regression was applied.

The best frequency regression considers the level of operational risk controls and GDP growth as covariates following the line of Povel et al. (2007) and Stewart (2016):

$$\log(n_{it}) = b_0 + b_1 \log(c_{it}) + b_2 GROWTH_{it} \qquad (31)$$

The estimated parameters are shown in Table 15. The interpretation is simple. More levels of control reduce the frequency of losses, and a higher macroeconomic activity in the country where banks headquarters are located increases the risk of event outbreaks.

Table 15

Dependent Variable: Log of Number of Loss Events

Regressors	Estimates	Std. error	*t*-value	*p*-value	
Intercept	0.98	0.25	3.88	0.00	***
Log(controls)	-1.54	0.33	-4.61	0.00	***
GDP growth	0.13	0.02	5.19	0.02	***

Note: Significance codes: 0=***, 0.001=**, 0.01=*, 0.05=*; Residual standard error: 0.99 on 219 degrees of freedom. Multiple *R*-squared: 0.18, Adjusted *R*-squared: 0.17, *F*-statistic: 24.3 on 2 and 219 *DF*, *p*-value: < 3.04e-10.

Both regression results are the outcome of a process of searching for the best fit using a number of covariates. Tables 14 and 15 summarize these regressions and then show the best results according to the Akaike information criterion (AIC), which is a standard method to select among models.

The Scaling technique shows that operational risk controls reduce the possibility of loss events outbreaks, but the occurrence of loss events do not per se determine the severity of losses. Economic growth affects both, the severity of losses as well as the frequency of losses, a result compatible with Povel et al. (2007) and Stewart (2016). The number of employees per branch affects the severity of losses but does not affect the frequency of losses. As shown in Chapter 3, the number of employees per branch approximates the number of inappropriate employee interactions within a firm that give rise to fraud risk. Last, financial crisis does increase the size of the severity of losses but does not affect the frequency of losses. Of course, the effects of controls and the number of employees per branch on the severity and frequency of losses are embedded already, directly or indirectly, in the model. The regression results are just a confirmation of what is assumed in the first

place, but the new result in the regressions is the importance of macro-environmental variables that were not used to build the model nor to calibrate it. The effect of economic growth on the severity and frequency of losses is a clean result in the analysis and confirms previous findings noted by Povel et al. (2007) and Stewart (2016). The effect of the Global Financial Crisis on the frequency of losses is also an important result that confirms the findings of Hess (2011).

When applying the LDA in the scaling technique, the frequency data of banks can be scaled in the same fashion as loss severities are scaled. Table 16 shows an example result of the operational risk levels with the scaling technique under possible specification alternatives of the frequency regression equation and the probability density of loss severities.

Table 16

Operational Risk Capital Levels in AUS.CBA bank at Different Percentiles (Millions of Euros)

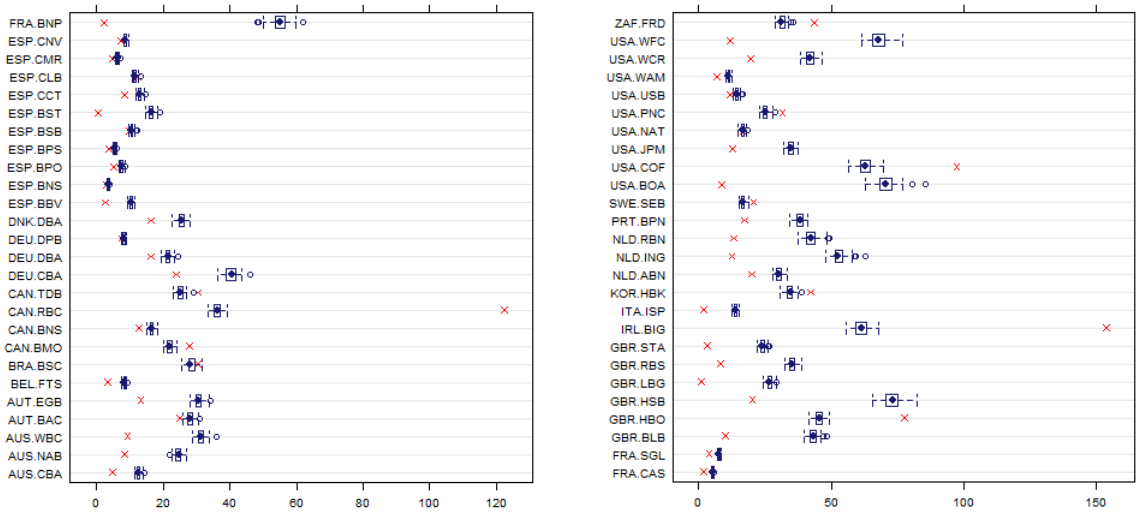
Percentile	True	Scaling 1	Scaling 2	Scaling 3
99.00	3.77	10.23	9.02	8.81
99.50	4.08	11.66	10.30	9.66
99.90	4.87	15.19	12.87	11.60
99.99	5.45	17.80	15.30	15.11

Note: Scaling 1 = Frequency regression equation contains more covariates and gamma severity density in LDA, Scaling 2 = Frequency regression equation only contains control levels and GDP growth. Gamma severity density in LDA, Scaling 3 = Frequency regression equation only contains control levels and GDP growth. Weibull severity density in LDA.

Table 16 shows that under different alternatives, the operational risk capital levels are more than twice the true capital levels estimated in the previous section for AUS.CBA. This is an early indication that the scaling technique may be exaggerating the losses for this specific bank. The possible reason of this result is that the scaling technique was based on mean losses and mean severities and does not take account of extreme behavior. In addition, the linear regression model used in this technique ignored extreme cases. In other words, the scaling technique does not properly consider extreme losses. The technique could be helpful

for studying mean losses but not extreme losses. This fact is important because operational risk capital levels need to take into account helpful models of extreme behavior. It is also important to note that, as documented in Shi et al. (2000), if there is too much heteroscedasticity in the data, the scaling regression may not be extracting the correct relationships.

Figure 22 depicts the VaR levels at the 99.9 percentile for all the dataset. The figure shows the extent of VaR variability among banks and the dispersion within each bank for 100 VaR calculations. It also shows the true operational VaR levels for all banks. It is straightforward to note that the true VaR levels are usually outside the interquartile ranges. Almost all the median values of the Scaling technique estimation of VaR at 99.9 percentile over-predict the true value. This is an indication that the simple Scaling technique implemented in this study does poorly in approximating to the true operational risk in each bank.



Note: The solid dots are median values, the boxes represent the interquartile range and the crosses stand for the true VaR level at 99.99 percentile for each bank.

Figure 22: Scaling technique VaR levels at 99.9 percentile for all the banks in the ORX.

Bayesian technique. This technique relies on the estimation of the parameters that define the probability distribution of loss event counts per year as well as the parameters that govern the severity density of loss events.

The estimations were performed with the help of a specialized R package called Laplacesdemons (Statisticat, 2015) with the distributions defined in the R package GAMLSS (Stasinopoulos & Rigby, 2007). In the GAMLSS environment, distributions may have up to four parameters. The study worked with distributions defined for one (mean) or two (mean and scale) parameters. A number of distributions to fit the parameters of frequency and severity distributions were tried from the distribution families defined in the GAMLSS package.

Following the example of Shevchenko (2011), priors were elicited empirically from the data, as permitted by the external data. In all cases, the prior chosen was the Gamma density, which allows only positive values for a parameter. To perform the elicitation, the external data about the number of losses in each bank or the loss per event in each bank permitted finding means and standard deviations for each bank. The cross section of means and standard deviations per bank are the data used to fit Gamma densities. In other words, the estimated Gamma densities describe the behavior of the parameters of location (mean) and scale (standard deviation) of the losses. These densities are called empirical or objective priors densities because they are elicited directly from data by estimating simple maximum likelihood estimations.

The prior densities were then combined with the likelihood of the data pertaining to each bank, namely, the likelihood functions of corresponding severities and frequencies. Table 17 shows the distinct distributional assumptions about the likelihood functions for both the frequency and severity.

Table 17

Distributions Used to Model Likelihood Functions of Operational Loss Data

Likelihood	Parameters
Frequency distributions	
Truncated Poisson	1
Truncated negative binomial	2
Severity densities	
Truncated Gamma	2
Truncated inverse Gaussian	2
Truncated Gumbel	2
Truncated Weibull	2
Truncated log normal	2
Truncated generalized Pareto	2

Note: Frequency distributions are left truncated (from zero); severity densities are left truncated from 0.02 (€20,000).

Figure 23 shows the estimated shapes of the posterior densities for the case of the annual frequency of loss events in internal fraud and retail banking for bank AUS.CBA. The Negative Binomial is better than the Poisson model because the Poisson model has higher standard deviation (equal to the mean), which does not fit frequency data well. In all the banks in the dataset, the Negative Binomial distribution was chosen to apply the Bayesian Technique. Given that the prior distributions of parameters are defined by Gamma distributions. The frequency model is defined as the Gamma-Negative Binomial model.

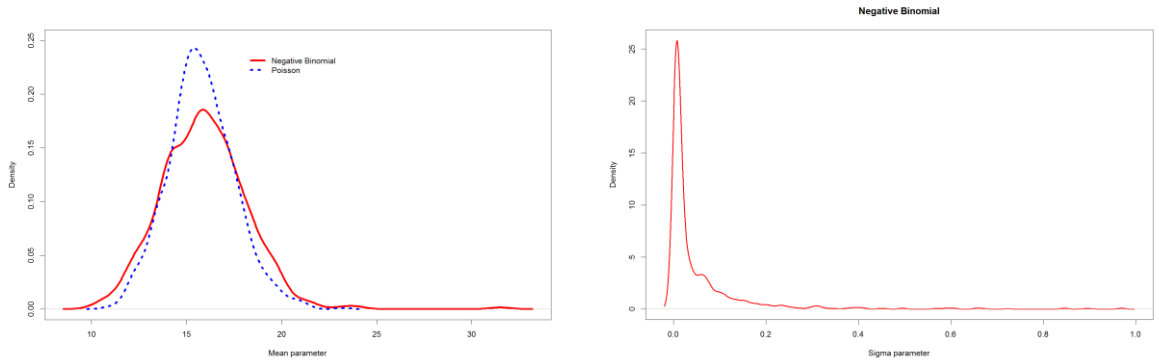


Figure 23. Posterior densities of mean and sigma parameters that govern frequency behavior.

Figure 24 shows the estimated posterior densities for the parameters that control the severity distributions in bank AUS.CBA. The four different types of densities show some

disparities. The gamma and Weibull are somewhat closer. Upon performing Bayes factor analysis to choose from these densities, the Weibull turns out to be the best. For example, it has the highest posterior probability. Bayes factor analysis is the preferred method for model choice under the Bayesian paradigm, in contrast to information criteria used in classical regression analysis. In all the banks, the Weibull was chosen to perform the Bayesian technique. Given that the prior distributions of severities are defined by Gamma distributions. The severity model is called as the Gamma-Weibull model.

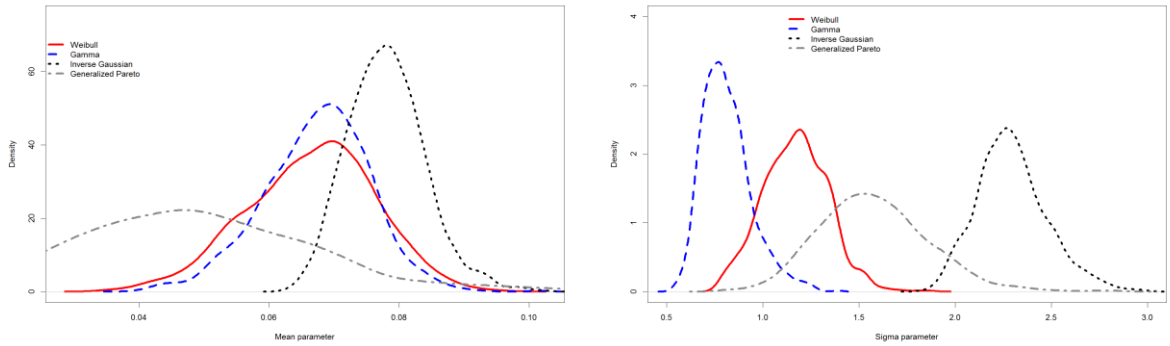


Figure 24. Posteriors for parameters governing severity.

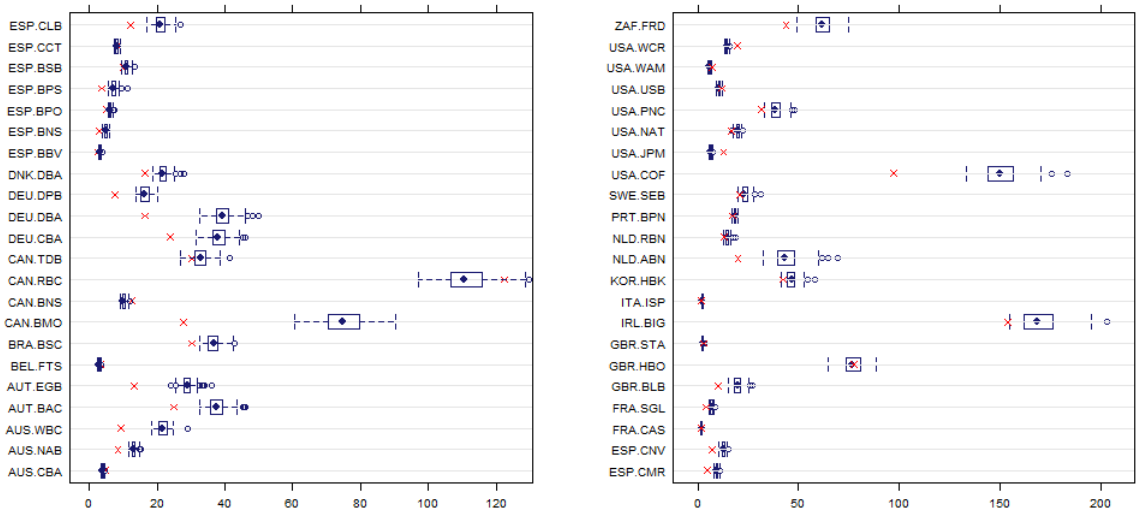
With these results, the LDA used the Gamma-Negative Binominal model to set the number of losses and the Gamma-Weibull model to set the severity of each loss. The Bayesian LDA was performed for all banks using the above combination of models for frequency and severity via simulation from the posterior predictive distributions as shown before in Chapter 3.

One important feature of Bayesian modeling is that the entire shape of the distributions of parameters, losses, and frequencies are modeled at once. The technique does not rely on mean behavior. Therefore, it is suited to capture possible extreme responses.

Figure 25 summarizes the estimation of operational risk levels for each bank under the Bayesian technique considering only the 99,9 percentile. For each bank, 100 operational risk levels at different percentiles were estimated. The solid blue dots are the median operational risk level, the boxes represent the interquartile range, the whiskers proxy for extreme values

and circles represent isolated extreme points. In the Figure, the crosses represent the true operational risk capital. As opposed to the Scaling technique results in Figure 22, the Bayesian technique results are closer to the true operational risk capital levels. In this case, more dots are inside the whisker ranges.

It is important to note that Figure 25 only shows 44 banks and not the original 52 banks considered in the sample. This is because eight banks have a very small number of losses and therefore it is not possible to estimate severity or frequency likelihoods based on the internal data and hence, it is not possible to obtain posterior distribution of parameters and perform the LDA.



Note: The solid dots are median values, the boxes represent the interquartile range (25 percentile to 75 percentile)

Figure 25: Bayesian technique VaR levels at 99.9 percentile for all the banks in the ORX.

Covariate-based technique. Following from Chapter 3, this technique models the parameters of severity and frequency distributions as functions of variables specific and common to each bank. These types of generalized regressions were performed with the GAMLSS R package as done in Ganegoda and Evans (2013). In particular, the regressions comprised the mean and the scale parameters using a number of distributional assumptions and specifications about the behavior of the mean or the scale of the distributions as functions of the covariates. The possible set of covariates used is the same as the Scaling technique. In

contrast to the simple Scaling technique, the advantage of the generalized regression in location and scale allows modeling the heteroscedasticity problem in the data, and hence, the Covariate-based technique with the generalized model has an important advantage over the simple Scaling technique with linear regression.

Tables 18 and 19 show the results for the best performing model for the severity of loss events in retail banking associated with internal fraud. This model is the truncated Weibull with mean and scale parameter. Given that losses smaller than €20,000 are not considered in the data, all loss data is truncated from below and therefore, truncated distribution functions needed to be estimated (Greene, 2012). Results in Table 18 show how the mean is affected by covariates. In particular, when a country to which a bank belongs to grows, the average size of losses increases. This result is similar to findings in the Scaling technique regressions and compatible with findings reported in Povel et al. (2007) and Stewart (2016). This positive impact has to do with the opportunistic behavior of workers stressed by Blacker and McConnell (2015). Arguably, in general macroeconomic boom periods more fraud opportunities arise.

In addition, when a country is perceived as more corrupt (lower value of CPI), the average losses are higher. The link between corruption perception in a given country and the size of losses due to internal fraud are stressed in the analytical framework outlined in Chapter 3. In the section about the model, the study highlighted the Cressey's triangle and pointed out that the level of bad interactions inside and outside banks bring about a rationalization for fraud. The corruption perception of a country is a proxy for outside bad interactions. In this sense, if criminal or corrupted behavior of citizens is broadly accepted in a society, then workers find more rationale for stealing from banks.

The regressions also considers idiosyncratic variables. For example, higher levels of operational risk controls reduce the severity of losses, higher employees per branch increase

the severity of losses and higher assets per employee reduce the severity of losses. These three effects are embedded in the loss model calibration or specifications. Therefore, these results are expected.

Nevertheless, the remarkable finding is that neither the GDP growth nor the CPI were used to calibrate the loss model or as causal variables in the model specification in the previous chapter, yet they show a significant association with fraud losses. Furthermore, these associations conform well with existing theory of how internal fraud losses occur.

Table 18

Estimates of the Regression in the Mean (Truncated Weibull)

	Estimate	Std. error	*t*-value	*p*-value	
Intercept	-2.80	0.098	-28.43	0.0000	***
GDP growth	0.01	0.006	2.09	0.0369	*
CPI	-0.05	0.012	-4.06	0.0000	***
Control	-0.40	0.184	-2.17	0.0304	*
Employees per branch	0.13	0.003	51.57	0.000	***
Assets per employee	-0.01	0.005	-1.95	0.0510	.

Note: Significance codes: 0 = ***, 0.001='**', 0.01='*', 0.05='.', 0.1=' '

As shown, there is a fundamental difference between this regression result and that obtained under the Scaling technique. The Scaling technique was a model only of the mean losses or frequencies under the standard linear regression assumption. The generalized linear model of mean, shape, and scale assumes general distributional assumptions, and not only the mean is modeled, but also, the scale and shape of the response variable. In other words, the GAMLSS procedure in Rigby, Stasinopoulos, Heller, and Voudouris (2014) allows non-linear behavior and controls for heteroscedasticity, which is more suited to extreme losses.

Table 19 shows the results for the regression that explains the scale parameter of the truncated Weibull distribution. This means that the GAMLSS set up allows modeling not only the mean of the severity of fraud losses but also its variance. The results suggest that a

higher corruption perception (lower index) implies higher fraud loss volatility, but more employees per branch diminish the fraud loss volatility. These results are not straightforward to justify in terms of theory but provides an interesting starting point for further research. For now, this is beyond the scope of the present research.

The regression in the scale parameter also includes smoothed GDP growth, controls, and assets per employee. The only smoothed variable with some individual significance is assets per employee.

Table 19

Estimates of the Regression in the Scale Parameter (Truncated Weibull)

	Estimate	Std. error	*t*-value	*p*-value	
Intercept	0.733	0.085	8.612	0.0000	***
CPI	-0.068	0.011	-6.321	0.0000	***
Employees per branch	-0.012	0.002	-5.843	0.000	***
GDP growth (a)	0.002	0.005	0.452	0.6510	
Control (a)	-0.015	0.153	-0.095	0.9242	
Assets per employee (a)	0.009	0.005	1.848	0.0646	.

Note: (a) Smoothing is performed with *p* splines and significance codes are 0 = ***, 0.001='**', 0.01='*', 0.05='.', 0.1=' '.

Tables 18 and 19 report the baseline regression. The number of models that can be explored is vast. Models in the GAMLSS setup may differ in the underlying distribution of the error terms. In the standard OLS setup the only distribution modelled is the Normal. In the GAMLSS there are families of distributions from which to choose. Once a distribution is chosen, models still differ because they may have different covariates.

The approach taken in this research considered first choosing the distributions given a benchmark set of covariates. Table 20 shows the Akaike information criteria (AIC) statistics for the estimated models and specifications for the underlying distributions of the errors. The AIC supports the truncated Weibull model with mean and scale. This is the distribution that supports the regressions shown in Tables 18 and 19.

Other covariates apart from those considered in Tables 18 and 19 were considered. Table F-1 in Appendix F shows a set of models that are estimated. The key findings about the effect of GDP growth rates and the CPI are robust across models. When the dummy for the financial crisis is included, the GDP growth rate becomes an insignificant regressor. Therefore, we have a model that only considers the crisis dummy instead of GDP growth but the model is inferior in terms of the AIC.

Table 20

Severity Model Selection

Distributions	AIC
Truncated Weibull (mean and scale with smoothing in regression)	-6,581.01
Truncated generalized Pareto (mean and scale with smoothing in regression)	-6,503.15
Truncated Weibull (mean and scale regression)	-6,362.03
Truncated generalized Gamma (mean and scale regression)	-6,361.38
Truncated Weibull (only mean regression)	-6,298.77
Truncated generalized Gamma (only mean regression)	-6,296.86

The same general procedure was used to estimate the frequency regressions. Variants of the Poisson, Negative Binomial Type I, and Negative Binomial Type II (see Rigby et al. 2014) were estimated with GAMLSS. The chosen model, according to the AIC (see Table 21) was the Negative Binomial regression with regressions in the mean and scale.

Table 21

Frequency Model Selection

Distributions	AIC
Negative binomial (mean and scale with smoothing in regression)	1,611.76
Negative binomial II (mean and scale with smoothing in regression)	1,623.96
Negative binomial (mean and scale regression)	1,631.98
Negative binomial II (mean and scale regression)	1,635.26
Negative binomial (only mean regression)	1,637.36
Negative binomial II (only mean regression)	1,658.48

Regression results are shown in Table 22. GDP growth affects the number of annual loss events positively; more controls and more retail assets per employee reduce the number of losses. In the case of frequency, the CPI is not a significant regressor, neither a group of regressors such as government effectiveness, regulatory quality, rule of Law, control of corruption; all taken from the World Bank Worldwide Governance Indicators. See Table F-2 in Appendix F for the details of the alternative regressions.

Across all regressions performed, GDP growth stands robustly significant with a parameter value of 0.13. This means that there is a strong evidence that the opportunistic behavior described in Cressey's triangle also affects the number of fraud events. Hence, GDP growth affects both the number of events and the severity of those events. This is in line with the theory of opportunistic behavior of fraudsters when good aggregate economic times arrive.

In addition, the regression in the scale or the variance of the number of losses shown in Table 22 confirms that more employees per branch increase the variance in the number of annual losses and more operational risk controls reduce it.

Table 22

Regression Results in the Frequency Model

	Estimate	Std. error	*t*-value	*p*-value	
Regression in the mean					
Intercept	4.33	0.298	14.528	0.000	***
GDP growth	0.13	0.019	7.020	0.000	***
Controls	-3.50	0.575	-6.096	0.000	***
Assets per employee	-0.03	0.016	-1.969	0.050	.
Regression in the scale					
Intercept	-0.14	0.507	-0.272	0.786	
Employees per branch	0.05	0.015	3.019	0.003	**
Controls (a)	-2.18	1.057	-2.064	0.040	*

Note: Smoothing is performed with *p* splines and significance codes are 0 = ***, 0.001='**', 0.01='*', 0.05='.', 0.1=' '.

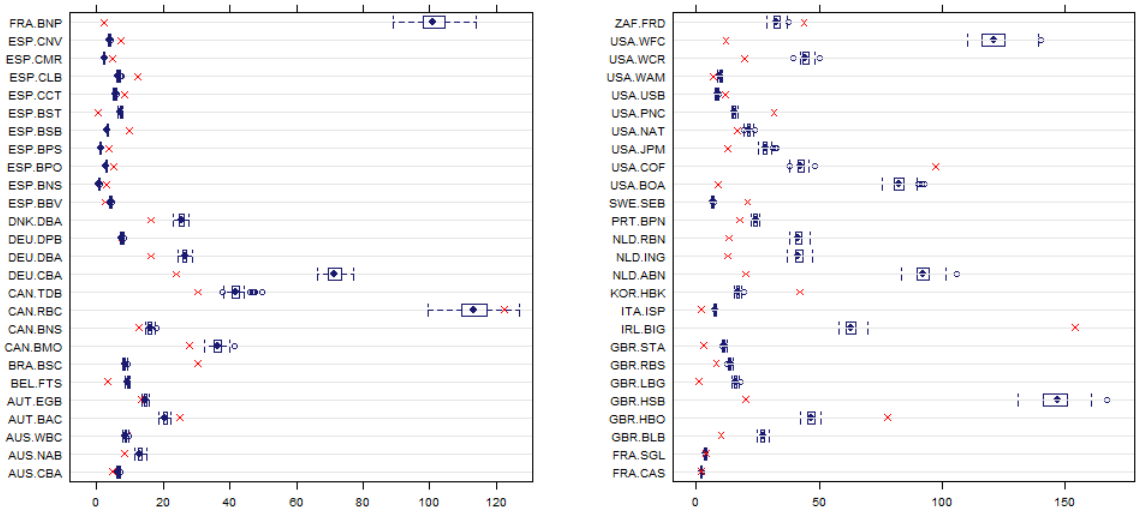
Discussion of Covariate-based regression results. These generalized linear regression results are new because they uncover one important determinant of fraud losses not previously documented in the literature: The corruption perception index of a country. The results imply that higher corruption perception indices at the national level have a direct effect on the size of losses due to internal fraud events. These results are compatible with the theory outlined in Chapter 3 about the rationale for committing fraud. According to Cressey's triangle described in Blacker and McConnell, the level of bad interactions inside and outside the firms shape fraudster's rationality. The corruption perception at the corporate or country level is a proxy variable for the level of interactions outside the bank. The regression results also bear some resemblance to ideas in organizational behavior such as Ashforth and Anand (2003) who noted that values and beliefs evolve to rationalize fraud in ways that neutralize the stigma of corruption. Societal-level pressures as described in Zahra, Priem, and Rasheed (2007) drive these values and beliefs.

The findings show that it is the yearly average size of losses that are affected by national corruption perceptions, not the frequency of losses within a year. More specifically, it is the extensive, not the intensive, margin of losses that can be explained by corruption perceptions. These results are also compatible with a recent global survey by KPMG about the profile of 75 fraudsters. According to the survey (KPMG, 2016), 66% of fraudsters mentioned that their main motive for committing fraud was personal financial gain or greed but an important 13% also said that the motives were also culturally driven. This last element is likely to be related to the overall corruption perception environment.

LDA within the Covariate-based technique. The LDA under the Covariate-based technique was applied with the truncated Weibull model for severity and the Negative Binominal model for the number of loss events per year. Both distributions were conditioned by the particular value of the covariates each bank had at the end of 2010. The LDA helped

forecast the distribution of total losses in 2011; therefore, the value of the covariates in 2010 are good approximation of the state of each bank when they perform such forecasts.

Figure 26 shows the estimations in all the 52 banks in the database, notwithstanding the fact that they may have few or no internal losses at all. At first sight, the results are closer to those obtained by Bayesian technique.



Note: The solid dots are median values; the boxes represent the interquartile range (25 percentile to 75 percentile)

Figure 26: Covariate-based technique VaR levels at 99.9 percentile for all the banks in the ORX

Comparison of Techniques

Based on the arguments presented in Chapter 3, a simulation-based comparison was performed. The application of the LDA for each technique and for each bank relies on 10,000 simulations each. This huge number of simulations is necessary to extract accurate extreme percentiles because then it is possible to extract the shape of the distribution at the tails more accurately. This process delivered operational risk capital levels for each technique, but in order to make proper comparisons, the procedure repeated the same process a number of times to be able to make comparisons among operational risk capital levels.

The simulation-based comparison relied on replicating all operational risk capital calculations 100 times for each of the banks. With this number of calculations, key statistics

were gathered. First, for each bank, squared errors against their operational risk capital level for a given percentile were calculated. Therefore, there were 100 squared errors for each bank and for each percentile of interest. The calculations considered the 50, 99, 99.5, 99.9 and 99.99 percentiles. Figures 27, 28 and 29 show the distributions of the squared errors for 44 banks. The closer to cero are these distributions the better. The graphs compare both the Bayesian and the Covariate-based technique calculations of the squared errors. Graphs show box and whisker plots as a way to summarizing the distributions of the squared errors. The boxes comprise the interquartile range and the whiskers plus or minus 1.5 times the interquartile range. A technique is strongly superior to another if the range of the whiskers is completely below the range of the other distribution. A technique is weakly superior if most part of its whiskers range is below the other distribution though there is some overlap between the two distributions. There is no winner if boxes overlap or one box contains the other box.

Table 23

Summary of the mean squared error comparison between the Bayesian and Covariate-based techniques at the 99.9 percentile

Bank	**Winner**	**Degree**	**Bank**	**Winner**	**Degree**	**Bank**	**Winner**	**Degree**
AUS.CBA	Bayesian	weak	ESP.BBV	Bayesian	strong	ITA.ISP	Bayesian	strong
AUS.NAB	-	-	ESP.BNS	-	-	KOR.HBK	Bayesian	strong
AUS.WBC	Covariate	strong	ESP.BPO	Bayesian	strong	NLD.ABN	Bayesian	strong
AUT.BAC	Covariate	strong	ESP.BPS	Covariate	weak	NLD.RBN	Bayesian	strong
AUT.EGB	Covariate	strong	ESP.BSB	Bayesian	strong	PRT.BPN	Bayesian	strong
BEL.FTS	Bayesian	strong	ESP.CCT	Bayesian	strong	SWE.SEB	Bayesian	strong
BRA.BSC	Bayesian	strong	ESP.CLB	Covariate	weak	USA.COF	-	-
CAN.BMO	Covariate	strong	ESP.CMR	Covariate	weak	USA.JPM	Bayesian	strong
CAN.BNS	Bayesian	weak	ESP.CNV	Covariate	strong	USA.NAT	Bayesian	weak
CAN.RBC	-	-	FRA.CAS	Bayesian	weak	USA.PNC	Bayesian	strong
CAN.TDB	Bayesian	strong	FRA.SGL	Covariate	strong	USA.USB	Bayesian	strong
DEU.CBA	Bayesian	strong	GBR.BLB	Bayesian	strong	USA.WAM	Bayesian	strong
DEU.DBA	Covariate	strong	GBR.HBO	Bayesian	strong	USA.WCR	Bayesian	strong
DEU.DPB	Covariate	strong	GBR.STA	Bayesian	strong	ZAF.FRD	Covariate	weak
DNK.DBA	Bayesian	weak	IRL.BIG	Bayesian	strong	-	-	-

Table 23 summarizes the comparison. In the 44 banks considered, the Bayesian technique strongly dominates the Covariate-based technique in 23 cases and weakly dominates in five more times. The Covariate-based strongly dominates the Bayesian technique in eight cases and does it weakly in four more cases. In four banks there is a tie.

Therefore, the simulation study methodology performed in the thesis, akin to stablished statistical research (Matzkin, 2003; Chen & Pouzo, 2012; Sarafidis, 2016; Wang and Zhao, 2016) allows to conclude that the Bayesian technique is clearly better than the Covariate-based technique at the 99.9 percentile in 64 percent of the cases. It remains to see what type of idiosyncratic factors turn a technique appropriate in each bank but this interesting avenue is not the scope of the present research.

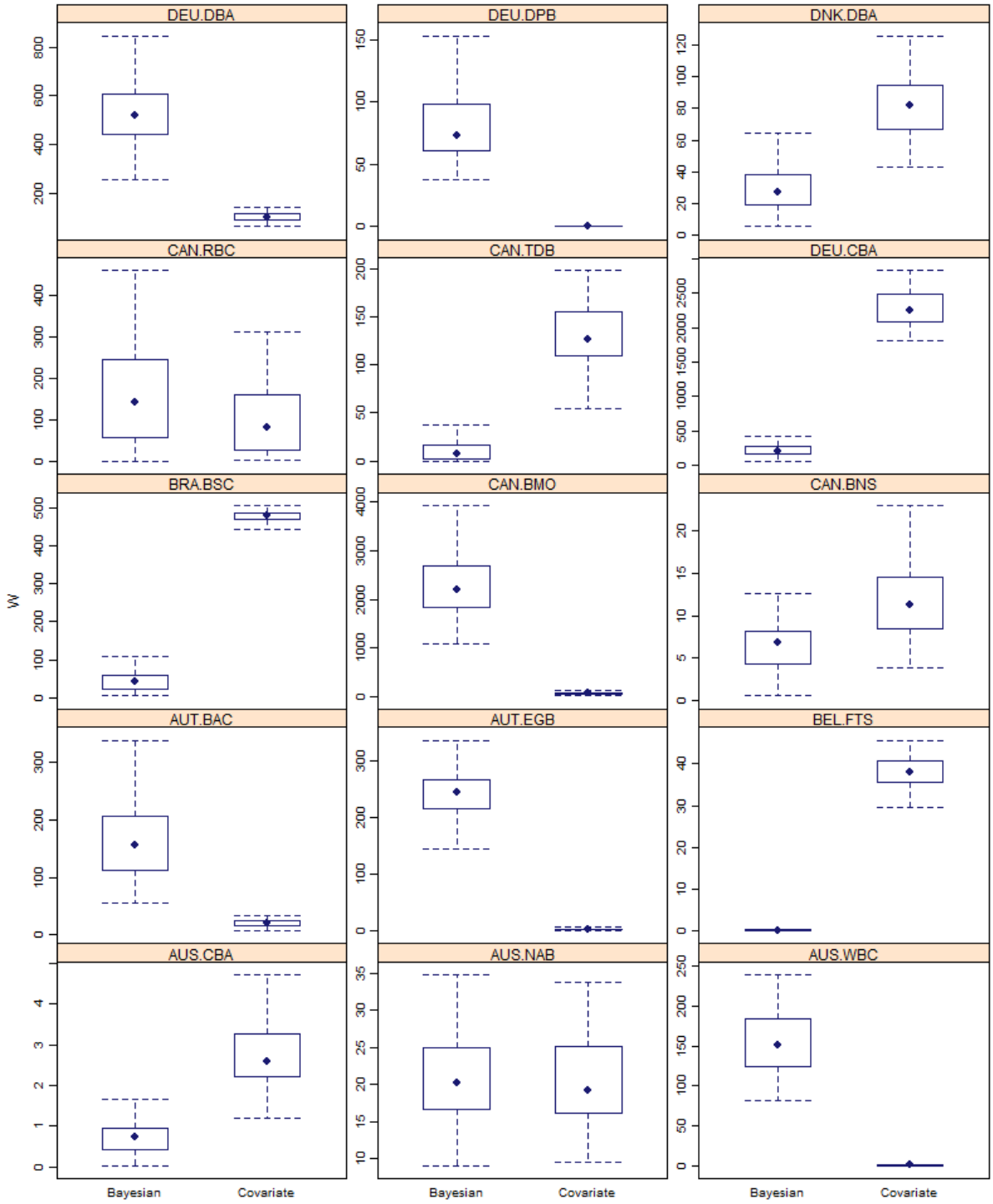


Figure 27: Distributions of squared errors of operational risk capitals (banks 1-15) Obtained by the Bayesian or Covariate-based technique against the true operational risk capital in each bank at the 99.9 percentile.

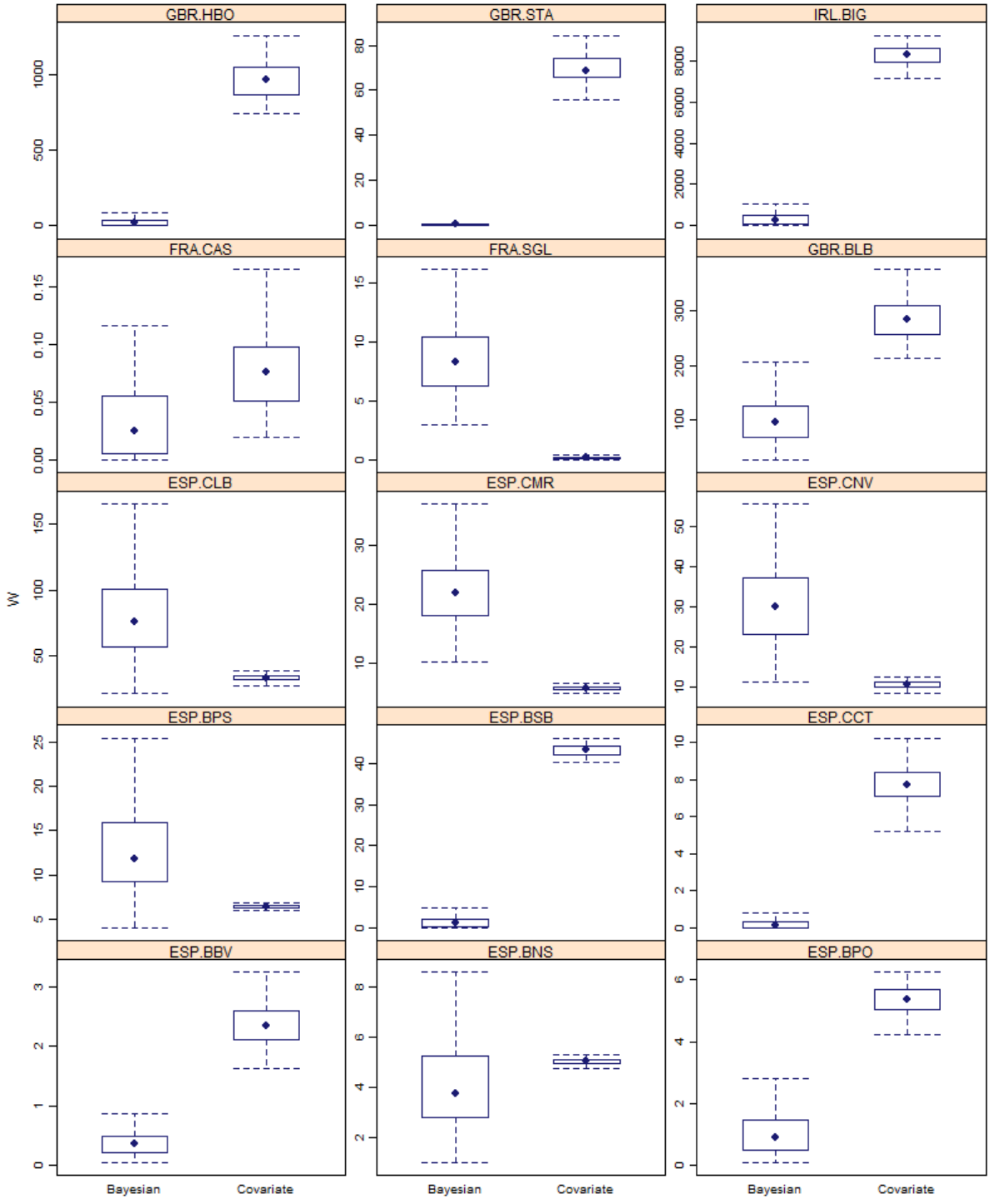


Figure 28: Distributions of squared errors of operational risk capitals (banks 16-30) Obtained by the Bayesian or Covariate-based technique against the true operational risk capital in each bank at the 99.9 percentile.

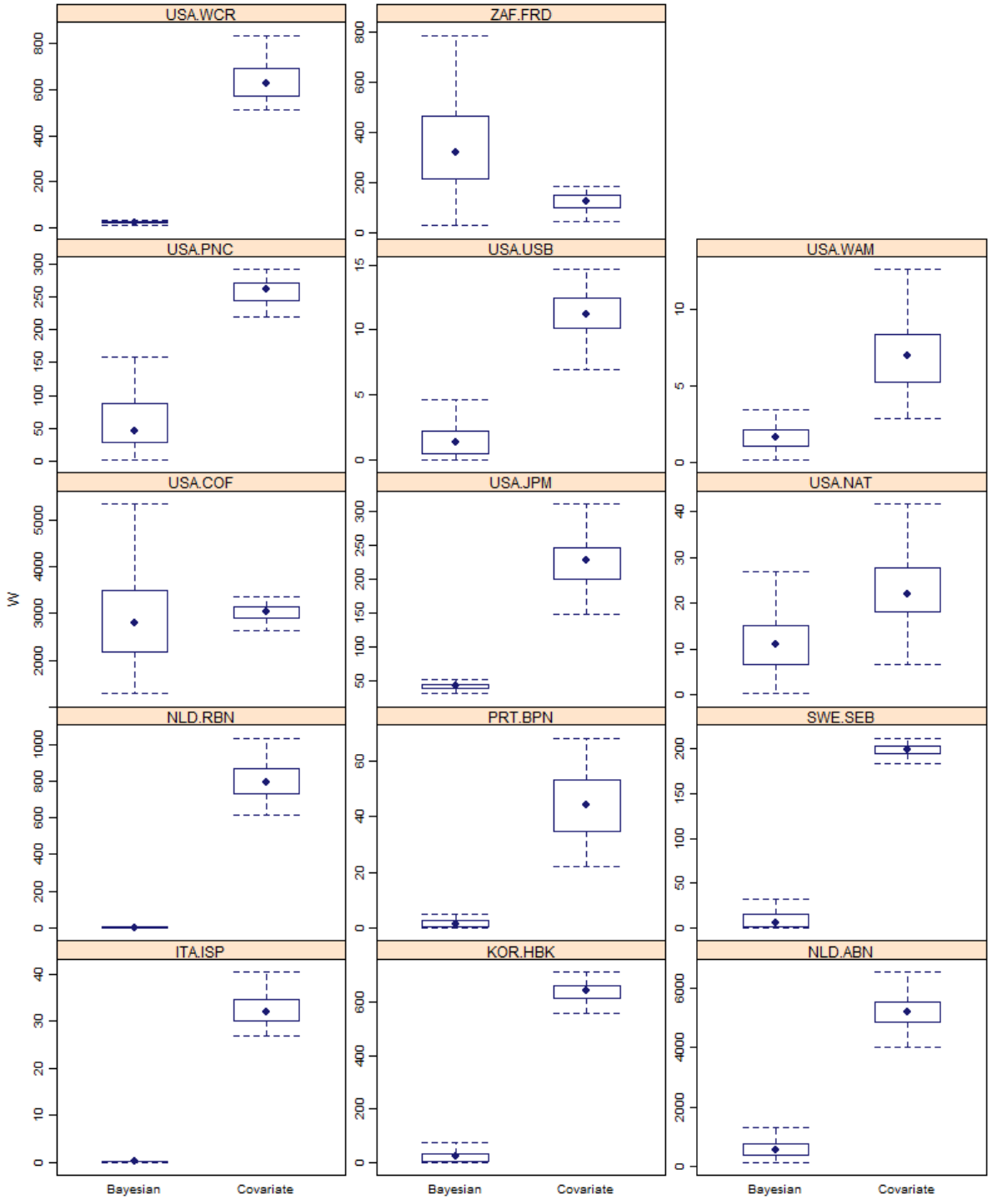


Figure 29: Distributions of squared errors of operational risk capitals (banks 31-44) Obtained by the Bayesian or Covariate-based technique against the true operational risk capital in each bank at the 99.9 percentile.

Findings

The quantitative results described in this chapter relate to the research questions laid out in Chapter 3. By construction, the model generated internal fraud losses that were close to reality, as shown in Figure 16, and that were associated with operational risk control levels in banks and with the observed number of employees per branch. The true empirical test of the operational risk model presented is to know whether those losses were correlated with macro environmental variables. In particular, MRQ1 is focused how the losses are related to the 2007-2009 Global Financial Crisis. The answer is that there is indeed some evidence that the Global Financial Crisis increased the severity of internal fraud losses. The simple ordinary least squares regression used in the scaling technique captured the aforementioned effect, but the generalized linear regression used in the covariate-based technique suggested that the Global Financial Crisis was not important, not for the severity nor the frequency of internal fraud losses.

As regards to the MRQ2, given the same conditions as MRQ1, namely, how are internal fraud losses related to the perception about corruption in the country where the main headquarters of a bank is located, in general, the answer was that there is some evidence that the corruption perception index affects internal fraud severities. This is the result of the generalized linear regression model used in the covariate-based technique. However, the simple liner regression model used in the implementation of the scaling technique did not capture the effect of the corruption perception over internal fraud losses. The size of macroeconomic activity, given by GDP growth, was present in both the linear and the generalized linear regression models, always with a positive sign. This suggested that economic growth is a strong predictor of internal fraud losses.

Nevertheless, it is critical to note that the relationships stemming from the covariate-based regression results were more reliable. The covariate-based technique relied on the

generalized linear regression in location and scale while the scaling technique relied on the simple linear regression model. The generalized regression is a more flexible representation of the data and can control for heteroscedasticity. Therefore, there is indeed evidence that the corruption perception index affects internal fraud severities.

MRQ3 was focused on the selection of the best internal-external data integration technique: Is there any technique that can be considered best practice to estimate a correct operational risk capital across all levels of risk tolerance? Here the answer is simple. There is no a best technique that delivers superior outcomes across all levels of risk appetite, but it is possible to rule out the simple scaling technique as an approach to perform internal-external data integration. The resulting operational risk capital levels in the scaling technique are far larger than are those obtained from the other two techniques whereas the Bayesian technique dominates the Covariate-based technique. Table 23 shows that the Bayesian technique performs better than the Covariate-base technique in 64 percent of the banks in the sample.

Therefore, results outlined here show that all three research hypothesis outlined in chapter 1 are rejected. Namely, $H_0$1, which stated that there is no change in the pattern of operational losses before and after the Global Financial Crisis, $H_0$2, which states that neither the frequency nor the severity of internal fraud operational losses are correlated with the corruption perception index of the country where the main headquarters of the bank is located is also rejected. Last, $H_0$3: One of the three techniques is systematically better than the others across possible risk tolerance values.

Summary

The results provided in this chapter were two-fold. First, the model set up in Chapter 3 was calibrated and simulated to generate operational losses due to internal fraud in retail banking. The calibration used data extracted from all banks in the ORX consortium that run a retail-banking segment as well as textual data from banks' annual reports publicly posted on

their Web pages. An important part of the process was to collect data for the 52 banks under analysis. All data were publicly available and could be used for replication purposes.

The second part of the results deals with the implementation of three data aggregation techniques to deal with the problem of operational risk capital estimation. A fundamental element of operational risk management is the estimation of capital level associated with the risk exposure. The problem, as described in previous chapters, is that banks cannot rely on their own data to perform operational risk capital estimations; instead, they must use data from other banks. The data integration techniques allowed for the pooling of internal and external data appropriately to perform risk management. The implementation of the three data integration techniques relied on the existing literature and was conducted with the simulated data generated in the first part of this chapter.

Both the scaling and the covariate-based technique rely on regressions that take into account idiosyncratic and common data to banks. Macroeconomic and political data were introduced into the analysis, for example the GDP growth of the country, where a bank's headquarters are located, or the corruption perception index. One remarkable result in these regressions is that GDP growth and the corruption perception index are associated with the outbreak of losses and their severity. In general, when a bank is located in a country where there is high macroeconomic growth, the average size of operational losses increases. In addition, when a country is perceived as more corruption prone, average loss amounts and the number of losses per year tend to increase. These results are new in the operational risk literature for internal fraud and support the internal fraud model developed.

Once the data integration techniques were implemented, a comparison exercise was conducted through a simulation study in accordance to standard practice in the statistical literature (Matzkin, 2003; Chen & Pouzo, 2012; Sarafidis, 2016; Wang and Zhao, 2016). The aim of the study is to shed light about the best-performing technique. The results favored both

the Bayesian and the covariate-based technique. Taken at the most used quantile (99,99%), the Bayesian technique performs better than the Covariate-based technique in 64 percent of the cases. These results led to the rejection of H_03 because no technique was shown to be superior always.

Chapter 5: Conclusions and Recommendations

In operational risk management, one of the AMA Basel II requirements is that any operational loss measurement system must include internal and relevant external data. This study included the evaluation of three prominent data integration techniques: The Scaling, Bayesian, and the Covariate-based techniques. Each technique led to different quantitative results for the operational risk capital required for a banking institution. Financial institutions that apply AMA for operational risk calculations need to know whether there are some circumstances in which one technique would perform better than do the others.

The purpose of this research was to apply a simulation study to determine which of three techniques performed best in reflecting the true operational loss distribution of the financial institution as required by regulators around the world. Performance was measured by the comparison of estimates of operational risk capital associated with each technique. The estimation of operational risk capital was based on a specific extreme quantile of the cumulative density function of operational losses in a given institution estimated through an LDA associated with each technique. A dynamic internal fraud model for operational losses was used to simulate the internal and external loss data necessary to perform these estimations. The purpose of the dynamic model was to capture the nature of internal fraud and the operational controls to mitigate or avoid the monetary losses caused by internal fraudsters to the retail segment of banks.

The research method applied to resolve the research problem is called a simulation study or simulation-based evaluation. This method is common in the statistics literature (Greene, 2012; Stern, 2000; Voss, 2013) and has been applied in many fields of research such as medicine, biology, psychology, physics, management, and economics. The approach included three main implementation steps: (a) Data simulation through an internal fraud model, (b) implementation of each data integration technique, and (c) a simulation study

evaluation to discover which data integration technique delivered a level of operational risk capital closer to the true operational risk implied by the data simulation model.

A limitation of the research design was that a model is never as complex as reality is. It is inevitable that aspects of reality would be excluded from a model. Given that the purpose of the study was to compare data integration techniques, it was sufficient to know that the absent features of the internal fraud model were not correlated with the features of the techniques. Furthermore, the evaluation of models or statistical estimators using the Monte Carlo simulation of an assumed data generating process is an acceptable and standard practice in academia (see, for example, McNeil et al., 2005).

This chapter is organized as follows: First some concluding remarks are made, then implications for operational risk management are drawn, and last, the chapter ends up with some recommendations.

Conclusions

This study contributes to the operational risk management literature in two important ways. The first contribution of the study is that it sheds light on a long-standing problem in operational risk management that in practice has hindered further development in the use of external data for risk management in financial institutions. Recent surveys such as BCBS (2014) and Deloitte (2015) have pointed out that the handling of external data by banks to perform risk management analysis is a long-standing problem and only slow progress has been made in recent years. The study served to tackle the internal-external data combination problem with the hope of contributing to the solution of this real business problem.

The existing literature about internal-external data aggregation in operational risk has proposed a number of techniques to perform the aggregation. In particular, there are three prominent techniques: The Scaling, the Bayesian, and the Covariate-base techniques. However, practitioners that seek to implement these techniques in real contexts do not have a

clear idea about the advantages or disadvantages of the techniques or simply do not know which technique they should use. This is because there is no study that provides hints about which technique is best.

This study is the first in the literature to compare the three techniques. The comparison is made by using a rigorous statistical procedure. The findings are that the standard Scaling technique is by far the less useful for estimating operational risk capital. The Bayesian and the Covariate-based techniques perform best. For the ORX banks during the period 2006-2010, and for the 99,9% quantile the Bayesian technique is best in 64 percent of banks while the Covariate-based technique dominates in 36 percent of the sample.

An important element of the simulation-study approach of the study is that it needs to set up a quantitative model for internal fraud in the retail business line of financial institutions. Hence, the study contributed in a second way by building and validating a quantitative model that can be used for exploratory analysis about the nature of internal fraud losses.

In contrast to models of operational loss events such as Kühn and Neu (2003, 2004), Leippold and Vanini (2005), and Bardoscia and Bellotti (2011), the study delved into internal fraud losses. The existing literature has been focused more on losses associated with information technology. The closest papers to the study are Fragnière et al. (2010) and Yang (2010), which incorporated human factors in the outbreak of losses. Those two papers do not directly tackle internal fraud losses, however.

The dynamic internal fraud model described in this study incorporates human factors such as the level of employees per branch as well as the ethical quality of workers. It also incorporates the extent of the endogenous risk controls driven by risk managers. The model is based on theories about people risk management, human resources, auditing, and organizational literature. The model has been validated with hard data. It turns out that losses

generated by the model in the heterogeneous banks across the world are, in general, associated with the GDP growth and the corruption perception of the country where banks are located. This result is captured in Povel et al. (2007) who stated that losses are higher during good macroeconomic times. This result, to the best of knowledge, is the first time that the corruption perception in a country has been associated with internal fraud losses empirically. When a country is perceived as more corrupt, retail banking in that country will feature more severe internal fraud losses.

Implications

The findings in this study apply specifically to operational risk management in financial institutions and in particular, to operational risk management associated with losses attributed to people within organizations. Human factors such as the number of employees in retail banking units or the ethical quality of workers have an effect on the frequency and severity of operational losses due to internal fraud. These results have implications for hiring policies at banks as well as the type of controls that need to be instituted to reduce internal fraud events. In essence, the findings have implications for organization leadership because sound leadership implies sound employees in an efficiency and ethical sense. Sound employees, in turn, would mean lower internal fraud loss events.

In a more technical sense, the results of the study have implications for the quantitative application of the AMA. When calculating operational risk, the Bayesian or the covariate-based technique should be chosen. If a bank were very concerned about operational risk losses, it would prefer the extreme quantiles and therefore should choose the Bayesian technique for data integration.

Recommendations

A number of recommendations are made based on the findings outlined above:

1. Regulators should standardize the criteria to integrate internal and external operational risk capital because the quantitative results could be very different depending on the technique applied. Competition in the financial system can be distorted otherwise, thus affecting the efficiency of the market.
2. In a financial institution authorized to apply an AMA, data integration is essential to perform operational risk capital estimations. The Bayesian as well as the covariate-based technique should be evaluated and implemented carefully. The scaling technique is not recommended.
3. The dynamic internal fraud model and the simulation-based evaluation of data integration techniques can be extended to other lines of business apart from retail banking.
4. An avenue for future research is to validate the model with actual data, be it the ORX dataset or other database consortiums.

References

Mihov, A., Curti, F., & Abdymomunov, A. (2015). US Banking Sector Operational Losses and the Macroeconomic Environment. Available at SSRN 2738485.

Agostini, A., Talamo, P., & Vecchione, V. (2010). Combining operational loss data with expert opinions through advanced credibility theory. *Journal of Operational Risk*, 5(1), 3–28. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Aleksy, M., Seedorf, S., & Cuske, C. (2008). A distributed simulation environment for simulation modeling in operational risk management. *Proceedings from 2008 International Conference Complex, Intelligent and Software Intensive Systems* (CISIS) (pp. 126-131). Barcelona, Spain: Polytechnic University of Catalonia. doi:10.1109/CISIS.2008.38

AlHussaini, W., & Karkoulian, S. (2015). Mitigating operational risk through knowledge management. *Journal of International Management Studies*, 15(2), 31–40. doi:10.18374/JIMS-15-2.4

Ames, M., Schuermann, T., & Scott, H. S. (2015). Bank capital for operational risk: A tale of fragility and instability. *Journal of Risk Management in Financial Institutions*, *8*(3), 227–243. doi:10.2139/ssrn.2396046

Andrews, D. W. (1993). Exactly median-unbiased estimation of first order autoregressive/unit root models. *Econometrica, 61*(1), 139–165. Retrieved from http://www.jstor.org/stable/2951781

Aranha, E., & Borba, P. (2008). Using process simulation to assess the test design effort reduction of a model-based testing approach. In Q. Wang, D. Pfahl, & D. Raffo (Eds.). *Making globally distributed software development a success story* (pp. 282–293). Berlin Heidelberg, Germany: Springer.

Aue, F., & Kalkbrener, M. (2006). LDA at work: Deutsche Bank's approach to quantifying operational risk. *Journal of Operational Risk, 1*(4), 49–93. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Ashforth, B. E., & Anand, V. (2003). The normalization of corruption in organizations. *Research in Organizational Behavior*, 25, 1–52. doi:10.1016/S0191-3085(03)25001-2

Banks, J. (1998). Principles of simulation. In J. Banks (Ed.), *Handbook of simulation* (pp. 3-30). New York, NY: John Wiley & Sons.

Bardoscia, M., & Bellotti, R. (2011). A dynamical approach to operational risk measurement. *The Journal of Operational Risk, 6*(1), 3–19. Retrieved from http://m.risk.net/digital_assets/4709/jop_v6n1a1.pdf

Basel Committee on Banking Supervision. (2006). *International convergence of capital measurement and capital standards: A revised framework.* Retrieved from http://www.bis.org/publ/bcbs107.htm

Basel Committee on Banking Supervision. (2011). *Operational risk supervisory guidelines for the advanced measurement approaches.* Retrieved from http://www.bis.org/publ/bcbs196.htm

Basel Committee on Banking Supervision. (2014). *Review of the principles for the sound management of operational risk.* Retrieved from http://www.bis.org/publ/bcbs292.pdf

Baud, N., Frachot, A., & Roncalli, T. (2002). *Internal data, external data and consortium data for operational risk measurement: How to pool data properly?* Retrieved from http://www.thierry-roncalli.com/download/oprisk-data-light-version.pdf

Baud, N., Frachot, A., & Roncalli, T. (2003, February). How to avoid over-estimating capital charge for operational risk? *Operational Risk-Risk's Newsletter*. Retrieved from http://ssrn.com/abstract=1032591

Benyon, D. (2008). Top 100 banks - A new dawn for disclosure. *OpRisk & Compliance, 9*(10), 22. Retrieved from http://db.riskwaters.com/data/operationalrisk/pdf/ORC_Top_100.pdf

Blacker, K., & McConnell, P. (2015). *People risk management: A practical approach to managing the human factors that could harm your business*. [Kindle DX Reader version]. Retrieved from http://www.amazon.com.

Bolancé, C., Guillén, M., Gustafsson, J., & Nielsen, J.P. (2012). *Quantitative operational risk models* [Kindle DX Reader version]. Retrieved from http://www.amazon.com

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of econometrics, 31*(3), 307–327. Retrieved from http://public.econ.duke.edu/~boller/Published_Papers/joe_86.pdf

Bühlmann, H., & Gisler, A. (2005). *A course in credibility theory and its applications*. Berlin Heidelberg, Germany: Springer-Verlag.

Bühlmann, H., Shevchenko, P. V., & Wüthrich, M. V. (2007). A "toy" model for operational risk quantification using credibility theory. *The Journal of Operational Risk*, *2*(1), 3–19. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24), 4279–4292. doi:10.1002/sim.2673

Carmassi, J., & Micossi, S. (2012*). Time to set banking regulation right*. Brussels, Belgium: CEPS Paperbacks. Retrieved from http://aei.pitt.edu/id/eprint/47690

Cernauskas, D., & Tarantino, A. (2009). Operational risk management with process control and business process modeling. *Journal of Operational Risk*, *4*(2), 3–17. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Chaudhury, M. (2010). A review of the key issues in operational risk capital modeling. *Journal of Operational Risk, 5*(3), 37–66. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Chen, X., & Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1), 277-321. doi: 10.3982/ECTA7888

Cheng, F., Gamarnik, D., Jengte, N., Min, W., & Ramachandran, B. (2007). Modeling operational risks in business processes. *Journal of Operational Risk*, *2*(2), 73–98. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Chernobai, A., Rachev, S., Fabozzi F. (2007). *Operational risk: A guide to Basel II capital requirements, models, and analysis*. New Jersey, NJ: Wiley Finance.

Chernobai, A., & Yildirim, Y. (2008). The dynamics of operational loss clustering. *Journal of Banking & Finance, 32(12)*, 2655–2666. doi:10.1016/j.jbankfin.2008.06.001

Chernobai, A., Jorion, P., & Yu, F. (2011). The determinants of operational risk in U.S. financial institutions. *Journal of Financial and Quantitative Analysis*, *46*(6), 1683–1725. doi:10.1017/S0022109011000500

Cooke, D. L., & Rohleder, T. R. (2005, July). A conceptual model of operational risk. *Proceedings of the 23rd International Conference of the System Dynamics Society Boston* (pp. 17–21). Retrieved from http://www.systemdynamics.org/conferences/2005/proceed/papers/COOKE134.pdf

Cope, E., & Labbi, A. (2008). Operational loss scaling by exposure indicators: Evidence from the ORX database. *Journal of Operational Risk*, *3*(4), 25–45. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Cope, E. W., Piche, M. T., & Walter, J. S. (2012). Macroenvironmental determinants of operational loss severity. *Journal of Banking & Finance, 36*(5), 1362–1380. doi:10.1016/j.jbankfin.2011.11.022

Cressey, D. R. (1953). *Other people's money: A study in the social psychology of embezzlement*. New York: Free Press.

Cruz, M. G. (2002). *Modeling measuring and hedging operational risk.* Chichester, England: Wiley.

Cummings, A., Lewellen, T., McIntire, D., Moore, A., & Trzeciak, R. (2012). *Insider threat study: Illicit cyber activity involving fraud in the U.S. financial services sector* (CMU/SEI-2012-SR-004). Retrieved July 29, 2014, from the Software Engineering Institute, Carnegie Mellon University website: http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=27971

Cuske, C., Dickopp, T., & Seedorf, S. (2005). JOntoRisk: An ontology-based platform for knowledge-based simulation modeling in financial risk management. *Proceedings of the European Simulation and Modeling Conference*. Porto, Portugal: University of Porto. Retrieved from http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID802665_code488453.pdf?abstractid=802665&mirid=1

Dahen, H., & Dionne, G. (2010). Scaling models for the severity and frequency of external operational loss data. *Journal of Banking & Finance, 34*(7), 1484–1496. doi:10.1016/j.jbankfin.2009.08.017

Delignette-Muller, M., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4), 1–34. Retrieved from http://www.jstatsoft.org/v64/i04/

Deloitte. (2015). *Global Risk Management Survey: Operating in the new normal: Increased regulation and heightened expectations* (9th ed.). Retrieved from http://www2.deloitte.com/ru/en/pages/financial-services/articles/9th-global-risk-management-survey.html

Ergashev, B., Mittnik, S., & Sekeris, E. (2013). A Bayesian approach to extreme value estimation in operational risk modeling. *The Journal of Operational Risk, 8*(4), 55–81. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Figini, S., Gao, L., & Giudici, P. (2013). Bayesian operational risk models. *DEM Working Paper Series, Vol. 47* (pp. 7–13). Retrieved from ftp://economia.unipv.it/DEM/DEMWP0047.pdf

Finke, G., Singh, M., & Rachev, S. (2010). Operational risk quantification: A risk flow approach. *The Journal of Operational Risk*, *5*(4), 65–89. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Frachot, A., & Roncalli, T. (2002). Mixing internal and external data for managing operational risk. Available at SSRN website: http://ssrn.com/abstract=1032525

Fragnière, E., Gondzio, J., & Yang, X. (2010). Operations risk management by optimally planning the qualified workforce capacity. *European Journal of Operational Research*, *202*(2), 518–527. doi:10.1016/j.ejor.2009.05.026

Ganegoda, A., & Evans, J. (2013). A scaling model for severity of operational losses using generalized additive models for location scale and shape (GAMLSS). *Annals of Actuarial Science*, *7*(01), 61–100. doi:10.1017/S1748499512000267

Goodhart, C. (2011). *The Basel Committee of Banking Supervision: A history of the early years 1974-1997* [Kindle DX Reader version]. Retrieved from http://www.amazon.com

Greene, W. (2012). *Econometric analysis* (7th ed.). New Jersey, NJ. Prentice Hall.

Guegan, D., & Hassani, B. K. (2013). Using a time series approach to correct serial correlation in operational risk capital calculation. *The Journal of Operational Risk, 8(3),* 31–56. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Guillen, M., Gustafsson, J., & Nielsen, J.P. (2008). Combining underreported internal and external data for operational risk measurement. *Journal of Operational Risk, 3*(4), 3–24. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Hassani, B., & Renaudin, A. (2013). *The cascade Bayesian approach for a controlled integration of internal data, external data and scenarios* (CES Working Paper Series Report No. 2013-09). Retrieved from http://halshs.archives-ouvertes.fr/docs/00/79/50/46/PDF/13009.pdf.

Hatzakis, E. D. M., Nair, S. K., & Pinedo, M. (2010). Operations in financial services—An overview. *Production and Operations Management*, *19*(6), 633–664. doi:10.1111/j.1937-5956.2010.01163.x

Haubenstock, M., & Harding, L. (2003). *The loss distribution approach*. In C. Alexander (Ed.). *Operational risk: Regulation, analysis and management* (pp. 171–214). London, England: FT Prentice Hall.

Hemrit, W., & Ben Arab, M. (2012). The determinants of frequency and severity of operational losses in Tunisian insurance industry. *The Journal of Risk Finance*, *13*(5), 438-475. doi:10.1108/15265941211273759

Hess, C. (2011). The impact of the financial crisis on operational risk in the financial services industry: Empirical evidence. *The Journal of Operational Risk, 6(1),* 23–35. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

ISO. (2009). Risk management-vocabulary. Guide 73. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:guide:73:ed-1:v1:en

Jansen-Vullers, M. H., & Netjes, M. (2006). Business process simulation-tool survey. In K. Jensen (Ed.), *The seventh workshop on the practical use of coloured petri nets and CPN tools* (DAIMI PB 579, pp. 77–96). Aarhus, Denmark: University of Aarhus.

Jarrow, R. A. (2008). Operational risk. *Journal of Banking & Finance*, *32*(5), 870–879. doi:10.1016/j.jbankfin.2007.06.006

Jiménez, E., J. M. Feria, & J. L. Martín. (2009). *Advanced versus non-advanced approaches for estimating operational capital at risk: Evidence from retail banking.* Paper presented at the AECA XV Congress, Valladolid, España.

Kearney, C., & Liu, S. (2014, May). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis, 33*, 171–185. doi:10.1016/j.irfa.2014.02.006

Kessler, A. M. (2007). A systemic approach to operational risk measurement in financial institutions. *The Journal of Operational Risk*, *2*(4), 27–68. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Kochan, N. (2013, October 13), Culture, oversight and fraud prevention. *Financial Risk Management News and Analysis*. Retrieved from http://www.risk.net/

KPMG. (2016). *Global profiles of the fraudster: Technology enables and weak controls fuel the fraud.* Retrieved from https://assets.kpmg.com/content/dam/kpmg/pdf/2016/05/profiles-of-the-fraudster.pdf

Kühn, R., & Neu, P. (2003). Functional correlation approach to operational risk in banking organizations. *Physica A, 322*, 650–666. doi:10.1016/S0378-4371(02)01822-8

Kühn, R., & Neu, P. (2004). Adequate capital and stress testing for operational risks. In M. Cruz (Ed.), *Operational risk modeling and analysis: Theory and practice* (pp. 273–292). London, England: Risk Books.

Lambrigger, D. D., Shevchenko P. V., & Wüthrich, M. V. (2007). The quantification of operational risk using internal data, relevant external data and expert opinions. *The Journal of Operational Risk, 2*(3), 3–27. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Lambrigger, D. D., Shevchenko, P. V., & Wüthrich, M. V. (2008). Data combination under Basel II and solvency 2: Operational risk goes Bayesian. *Bulletin Français d'Actuariat, 8*(16), 4–13. Retrieved from http://www.ressources-actuarielles.net/EXT/IA/sitebfa.nsf/e17f9d0572826b2ac12572ba0022fd23/bc591bd5f7b07218c12574cb002c3fa2?OpenDocument

Leippold, M., & Vanini, P. (2005). The quantification of operational risk. *Journal of Risk, 8*(1), 59–85. Retrieved from http://www.risk.net/type/journal/source/journal-of-risk

Lukic, D., Margaryan, A., & Littlejohn, A. (2013). Individual agency in learning from incidents. *Human Resource Development International, 16*(4), 409–425. doi:10.1080/13678868.2013.792490

Matzkin, R. L. (2003). Nonparametric estimation of nonadditive random functions. *Econometrica,* 71(5), 1339-1375. doi: 10.1111/1468-0262.00452

McNeil, J. M., Frey, R., & Embrechts, P. (2005). *Quantitative risk management: Concepts, techniques, and tools*. Princeton, NJ: Princeton University Press.

Medova, E. A., & Berg-Yuen, P. E. (2009). Banking capital and operational risks: Comparative analysis of regulatory approaches for a bank. *Journal of financial transformation, 26*(7), 27–38. Retrieved from http://www.capco.com/sites/all/files/public/JOURNAL26_Web.pdf

Mizgier, K. J., Hora, M., Wagner, S. M., & Jüttner, M. P. (2015). Managing operational disruptions through capital adequacy and process improvement. *European Journal of Operational Research*, 245(1), 320–332. doi:10.1016/j.ejor.2015.02.029

Moosa, I. (2011). Operational risk as a function of the state of the economy. *Economic Modelling*, *28*(5), 2137–2142. doi:10.1016/j.econmod.2011.05.011

Na, H. S., van den Berg, J., Couto, L., & Leipoldt, M. (2006). An econometric model to scale operational losses. *Journal of Operational Risk*, *1*(2), 11–31. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Operational Riskdata eXchange Association. (2012). *2012 ORX report on operational risk loss data*. Retrieved from http://www.orx.org/Pages/ORXData.aspx

Panjer, H. H. (2006). *Operational risk: Modeling analytics*. New Jersey, NJ: John Wiley & Sons.

Paredes, L. (2006). Operational risk capital and insurance in emerging markets. *Documento de Trabajo* (SBS 04/2006). Retrieved from http://www.sbs.gob.pe/.../0/0/.../Parte2_Rocio_Paredes.pdf

Patel, S. (2010). Quantifying operational risk [PDF document]. Retrieved from http://www.casact.org/education/reinsure/2010/handouts/CS14-PatelAppendix.pdf

Peters, G. W., & Sisson, S. A. (2006). Bayesian inference, Monte Carlo sampling and operational risk. *Journal of Operational Risk*, *1*(3), 27–50. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Pond, K. (2014). Retail banking (Third edition). Cranbrook, Kent Global Professional Publishing.

Povel, P., Singh, R., & Winton, A. (2007). Booms, busts, and fraud. *Review of Financial Studies*, *20*(4), 1219–1254. doi:10.1093/revfin/hhm012

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org

Rigby, B., Stasinopoulos, M., Heller, G., & Voudouris, V. (2014). *The distribution toolbox of GAMLSS*. Retrieved from http://www.gamlss.org/wp-content/uploads/2014/10/distributions.pdf.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods*. New York, NY: Springer.

Rootzen, H., & Tajvidi, N. (1997). Extreme value statistics and windstorm losses: A case study (1995). *Scandinavian Actuarial Journal*, *1*, 70–94. doi:10.1080/03461238.1997.10413979

Sabatini, J, & Wills, S. (2008). *Use of external data for operational risk management* [PDF document]. Retrieved from https://www.boj.or.jp/en/announcements/release_2008/data/fsc0804a4.pdf

Sabatini, J. (2009). Profile of ORX and a case study in the use of consortium loss data [PDF document]. Retrieved from http://www.abieventi.it/documenti/2973/Sabatini-JPMorgan-Chase-ORX.pdf

Sarafidis, V. (2016). Neighbourhood GMM estimation of dynamic panel data models. *Computational Statistics & Data Analysis*, 100, 526-544. doi:10.1016/j.csda.2015.11.015

Shepperd, M., & Kadoda, G. (2001). Comparing software prediction techniques using simulation. *Software Engineering, IEEE Transactions, 27*(11), 1014–1022. doi:10.1109/32.965341

Shevchenko, P. V. (2011). *Modelling operational risk using Bayesian inference*. Berlin, Germany: Springer-Verlag.

Shevchenko, P. V., & Peters, G. W. (2013). *Loss distribution approach for operational risk capital modelling under Basel II: Combining different data sources for risk estimation*. Retrieved from http:// arXiv preprint arXiv:1306.1882

Shevchenko, P. V., & Wüthrich, M. V. (2006). The structural modelling of operational risk via Bayesian inference: Combining loss data with expert opinions. *The Journal of Operational Risk, 1*(3), 3–26. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Shih, J., Samad-Khan, A. H., & Medapa, P. (2000). Is the size of an operational loss related to firm size? *Operational Risk Magazine*, *2*, 1–4. Retrieved from http://www.risk.net/operational-risk-and-regulation/feature/1508327/is-the-size-of-an-operational-loss-related-to-firm-size

Smith, R. L., & Shively, T. S. (1995). Point process approach to modeling trends in troposphereic ozone based on exceedances of a high threshold. *Atmospheric Environment*, *29*(23), 3489–3499. doi:10.1016/1352-2310(95)00030-3

Stasinopoulos, D. M., & Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS). *Journal of Statistical Software*, *23*(7), 1–46. Retrieved from http://www.jstatsoft.org/v23/i07

Statisticat, LLC. (2015). *Complete environment for Bayesian inference*. Retrieved from https://github.com/ecbrown/LaplacesDemonStern, S. (2000). Simulation-based inference in econometrics: Motivation and methods. In R. Mariano, T. Schuermann, & M. Weeks (Eds.). *Simulation-based inference in econometrics: Methods and applications* (pp. 9–37). Cambridge, England: Cambridge University Press.

Stewart, R. T. (2016). Bank fraud and the macroeconomy. *Journal of Operational Risk*, 11(1). doi: 10.21314/JOP.2016.172

Supatgiat, C., Heusler, L., & Kenyon, C. (2006). Cause-to-effect operational risk quantification, *Risk Management, 8*(1), 16–42. doi:10.1057/palgrave.rm.8250001

Teker, D. (2005). Comparative analysis of operational risk measurement techniques. *International Trade and Finance Association Conference Papers.* Retrieved from http://EconPapers.repec.org/RePEc:bep:itfapp:1020

The Banker Database. (2015). *Sample ranking—Top 5 UK Banks: Top world banks 2014.* Retrieved from http://www.thebankerdatabase.com/files/Top_5_UK_Banks.pdf http://EconPapers.repec.org/RePEc:bep:itfapp:1020

Voss, J. (2013). *An introduction to statistical computing: A simulation-based approach*. [Kindle DX Reader version]. Retrieved from http://www.amazon.com

Wahlström, J. (2013). *Operational risk modeling: Theory and practice*. (Master's thesis, Royal Institute of Technology, Stockholm, Sweden). Available from http://www.math.kth.se/matstat/seminarier/131218b.htm

Wang, C. S., & Zhao, Z. (2016). Conditional Value-at-Risk: Semiparametric estimation and inference. *Journal of Econometrics* (forthcoming). doi: 10.1016/j.jeconom.2016.07.002

Wei, R. (2007). Quantification of operational losses using firm-specific information and external database. *Journal of Operational Risk, 1*(4), 3–34. Retrieved from http://www.risk.net/type/journal/source/journal-of-operational-risk

Weiß, B., & Winkelmann, A. (2011, January 4-7). Developing a process-oriented notation for modeling operational risks—A conceptual metamodel approach to operational risk management in knowledge intensive business processes within the financial industry. In *44th Hawaii International Conference on System Sciences (HICSS)* in Kauai, HI (pp. 1–10). doi:10.1109/HICSS.2011.156

Yang, X. (2010). *Applying stochastic programming models in financial risk management*. Retrieved from the Edinburgh Research Archive-Mathematics Thesis and Dissertation Collection website: https://www.era.lib.ed.ac.uk/handle/1842/4068

Zahra, S. A., Priem, R. L., & Rasheed, A. A. (2007). Understanding the causes and effects of top management fraud. *Organizational Dynamics*, 36(2), 122–139. doi:10.1016/j.orgdyn.2007.03.002

Zeigler, B. P., Praehofer, H., & Kim, T. G. (1976). *Theory of modeling and simulation*. New York, NY: Wiley.

Appendix A: Abbreviations Used in the Thesis

Abbreviation	Meaning
AIC	Akaike Information Criterion
AMA	Advanced Measurement Approach (for operational risk capital calculation)
BCBS	Basel Committee of Banking Supervision
BIA	Basic Indicator Approach (for operational risk capital calculation)
BPS	Business Process Simulation
BPM	Business Process Management
CDF	Cumulative Density Function
CVaR	Conditional Value at Risk
GAAP	Generally Accepted Accounting Principles
GAMLSS	Generalized Additive Model for Location Scale and Shape
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
LDA	Loss Distribution Approach (for operational risk capital calculation)
OpRK	Operational Risk Capital
$OpRK_c$	Operational risk capital estimated by the scaling technique
$OpRK_B$	Operational risk capital estimated by the Bayesian technique
$OpRK_C$	Operational risk capital estimated by the covariate-based technique
$OpRK_{true}$	True operational risk capital
ORX	Operational Riskdata eXchange Association
PDF	Probability Density Function
SM	Standardized Method (for operational risk capital calculation)
SDS	System Dynamics Simulation
TVaR	Tail Value at Risk
VaR	Value at Risk

Appendix B: Country Codes and Bank Codes

No.	Bank Code	Bank Name	Country	Country Code
1	AUS.CBA	Commonwealth Bank of Australia	Australia	AUS
2	AUS.NAB	National Australia Bank	Australia	AUS
3	AUS.WBC	Westpac Banking Corporation	Australia	AUS
4	AUT.BAC	Bank Austria – Creditanstalt	Austria	AUT
5	AUT.EGB	Erste Group Bank AG	Austria	AUT
6	BEL.FTS	Fortis	Belgium	BEL
7	BRA.BSC	Banco Bradesco S/A	Brazil	BRA
8	CAN.BNS	Bank of Nova Scotia	Canada	CAN
9	CAN.BMO	Bank of Montreal	Canada	CAN
10	CAN.RBC	Royal Bank of Canada (RBC)	Canada	CAN
11	CAN.TDB	Toronto Dominion Bank Group	Canada	CAN
12	DNK.DBA	Danske Bank A/S	Denmark	DNK
13	FRA.BNP	BNP Paribas	France	FRA
14	FRA.CAS	Credit Agricole SA	France	FRA
15	FRA.SGL	Société Générale	France	FRA
16	DEU.CBA	Commerzbank AG	Germany	DEU
17	DEU.DBA	Deutsche Bank AG	Germany	DEU
18	DEU.DPB	Deutsche Postbank AG	Germany	DEU
19	IRL.BIG	Bank of Ireland Group	Ireland	IRL
20	ITA.ISP	Intesa SanPaolo	Italy	ITA
21	NLD.ABN	ABN AMRO	Netherlands	NLD
22	NLD.ING	ING Group	Netherlands	NLD
23	NLD.RBN	Rabobank Nederland	Netherlands	NLD
24	PRT.BPN	Banco Portugues de Negocios	Portugal	PRT
25	ZAF.FRD	First Rand	South Africa	ZAF
26	KOR.HBK	Hana Bank	South Korea	KOR
27	ESP.BSB	Banc Sabadell	Spain	ESP
28	ESP.BBV	Banco Bilbao Vizcaya Argentaria	Spain	ESP
29	ESP.BPS	Banco Pastor	Spain	ESP
30	ESP.BPO	Banco Popular	Spain	ESP

31	ESP.BST	Banco Santander	Spain	ESP
32	ESP.BNS	Banesto	Spain	ESP
33	ESP.CCT	Caixa Catalunya	Spain	ESP
34	ESP.CNV	Caixanova	Spain	ESP
35	ESP.CLB	Caja Laboral	Spain	ESP
36	ESP.CMR	Cajamar	Spain	ESP
37	SWE.SEB	Skandinaviska Enskilda Banken	Sweden	SWE
38	GBR.STA	Standard Chartered Bank	UK	GBR
39	GBR.BLB	Barclays Bank	UK	GBR
40	GRB.HBO	HBOS PLC	UK	GBR
41	GRB.HSB	HSBC Holdings plc	UK	GBR
42	GRB.LBG	Lloyds Banking Group	UK	GBR
43	GRB.RBS	Royal Bank of Scotland Group	UK	GBR
44	USA.BOA	Bank of America	USA	USA
45	USA.COF	Capital One	USA	USA
46	USA.JPM	JPMorgan Chase & Co.	USA	USA
47	USA.NAT	National City	USA	USA
48	USA.PNC	PNC Bank	USA	USA
49	USA.USB	US Bancorp	USA	USA
50	USA.WCR	Wachovia Corporation	USA	USA
51	USA.WAM	Washington Mutual	USA	USA
52	USA.WFC	Wells Fargo & Co	USA	USA

Appendix C: Data Sources

Table C-1

Data Sources about Macro Variables

Key	Variable name	Data source
country_name	Country Name	
country_code	Country Code	
year	Year 2006-2010	
gdp_growth	GDP growth (annual %)	World Development Indicators
gover_effective	Government Effectiveness	The Worldwide Governance Indicators
reg_quality	Regulatory Quality	The Worldwide Governance Indicators
rule_law	Rule of Law	The Worldwide Governance Indicators
cont_corrup	Control of Corruption	The Worldwide Governance Indicators
enforce_act	Enforcement actions taken over the past 5 years (2006-2010)	Bank Regulation and Supervision
sp_rem_bd	Remuneration of the board directors as part of the supervisory process of risk-taking	Bank Regulation and Supervision
sp_rem_sbm	Remuneration of senior bank management as part of the supervisory process of risk-taking	Bank Regulation and Supervision
sp_rem_bs	Remuneration of other bank staff as part of the supervisory process of risk-taking	Bank Regulation and Supervision
reg_act	Authority of the supervisory agency to take regulatory action when it considers that the remuneration or compensation is excessive	Bank Regulation and Supervision
cpi	Corruption Perceptions Index (CPI) score (2006-2010)	Transparency International

Appendix D: Data

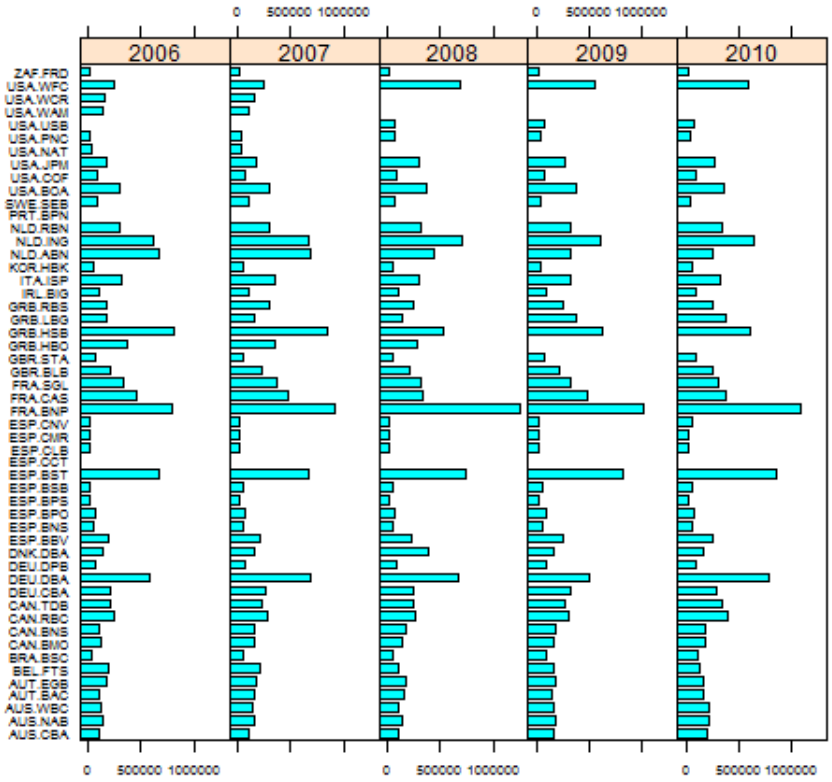


Figure D1. Banks by retail assets.

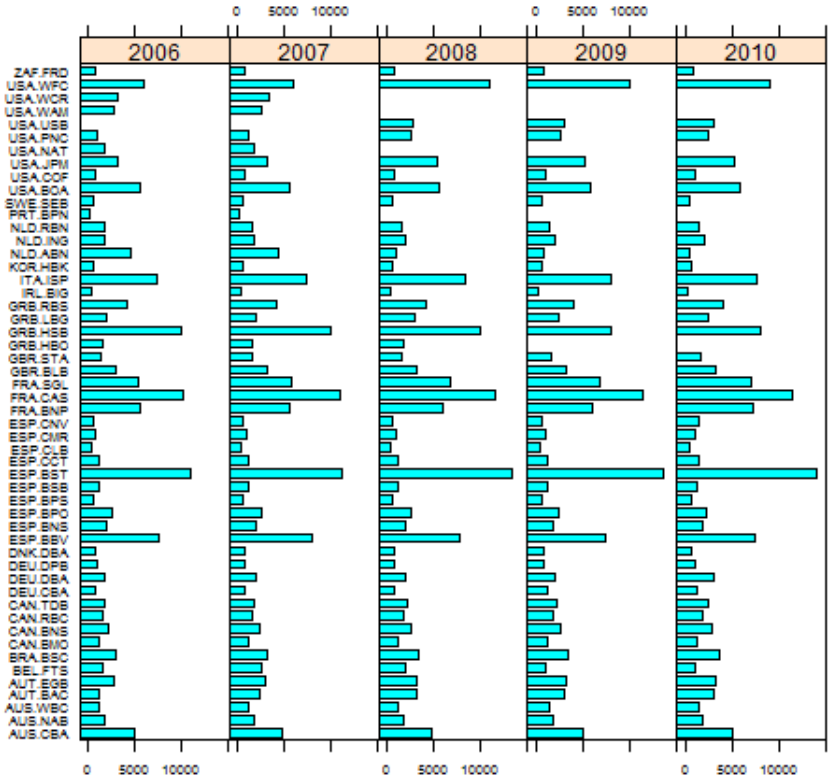


Figure D2. Banks by number of branches.

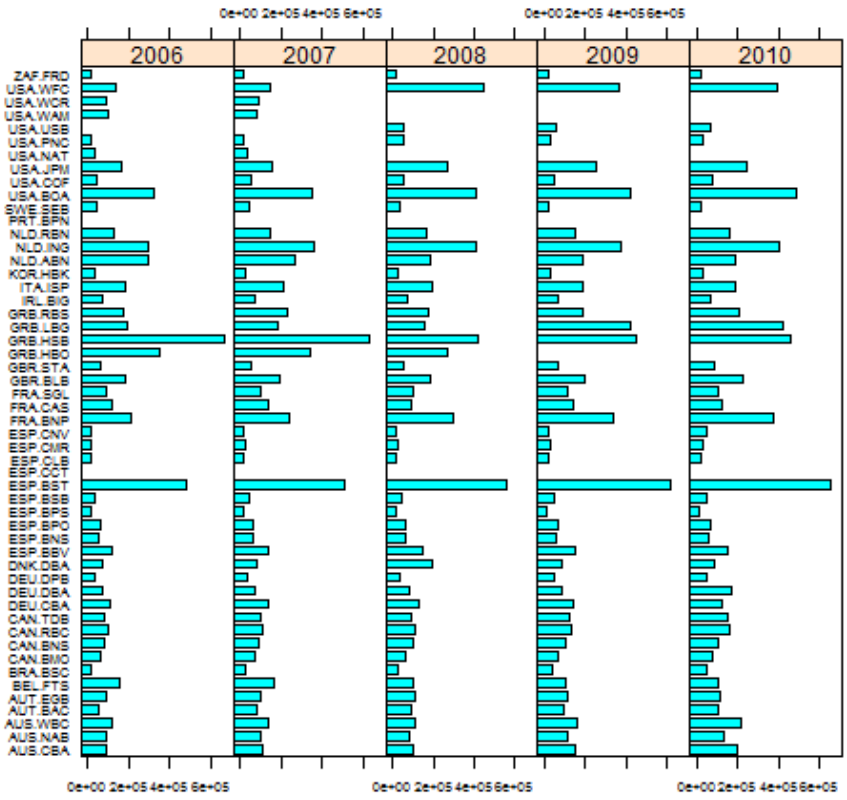


Figure D3. Banks by retail loans.

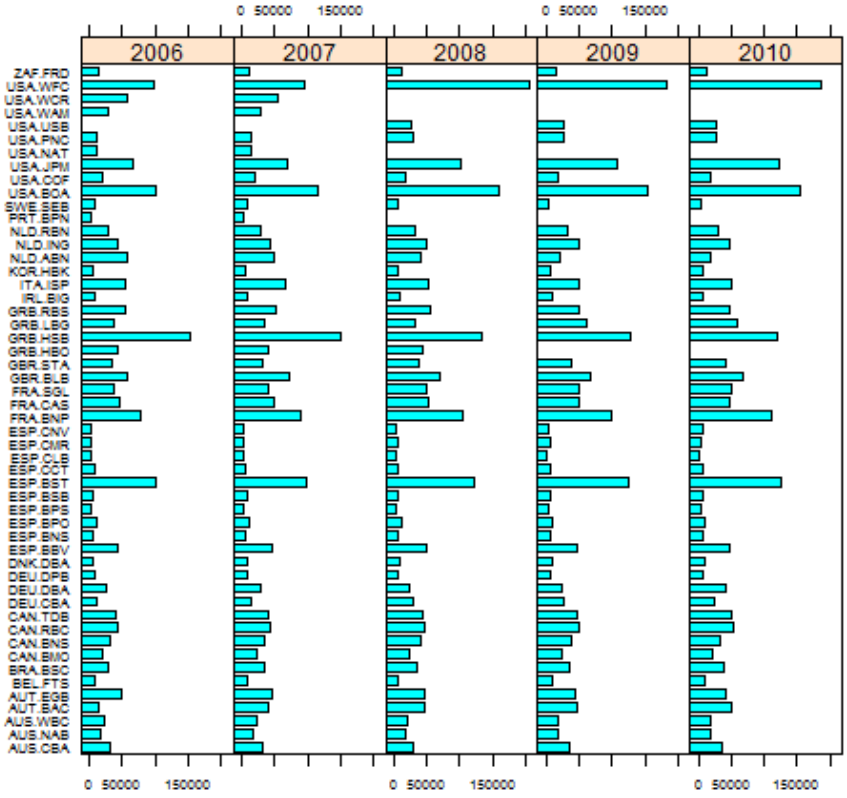


Figure D4. Banks by retail staff.

Appendix E: Distribution of idiosyncratic parameters

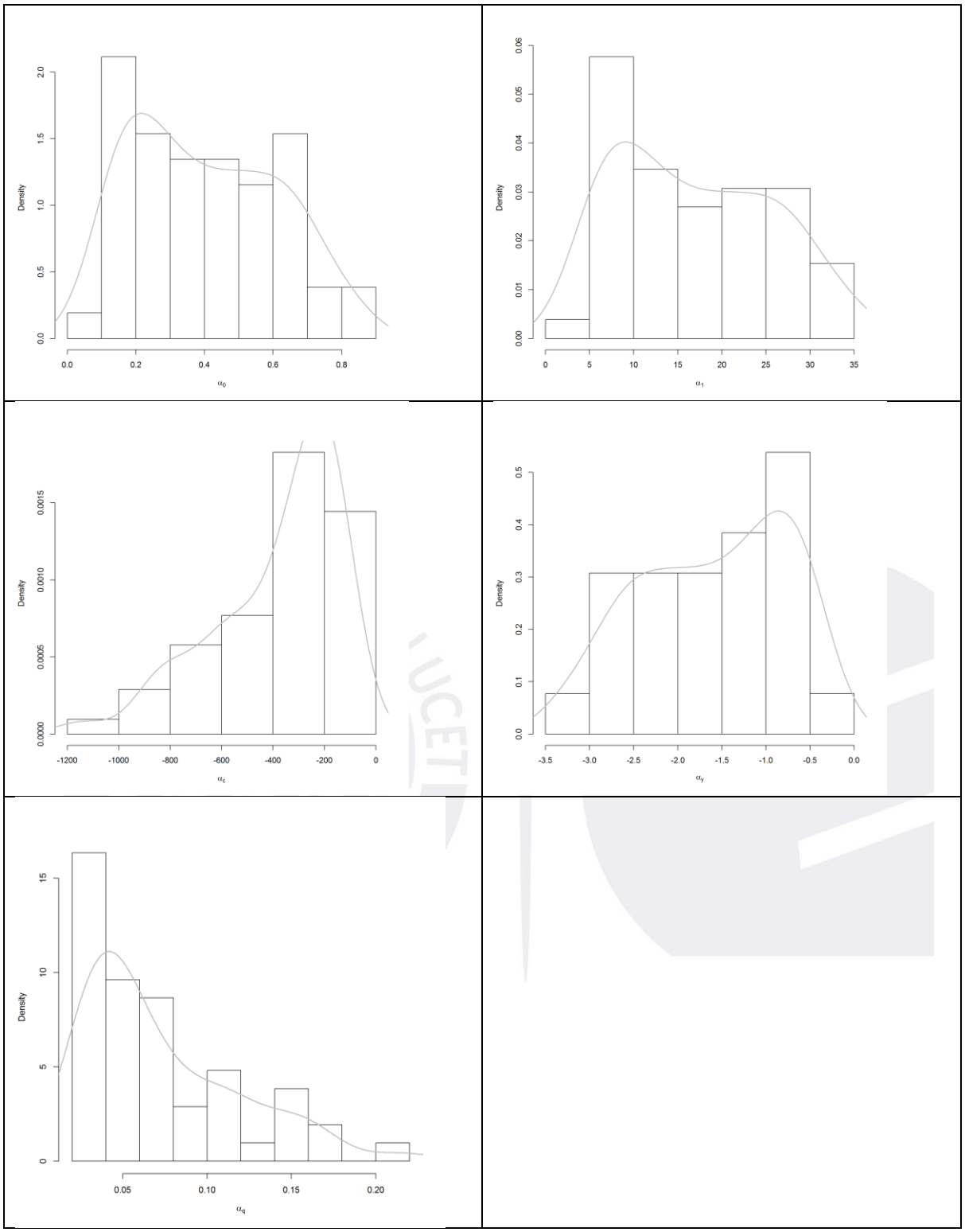


Figure E1. Distribution of parameters of the ramp function that defines the outbreak of operational losses in each bank.

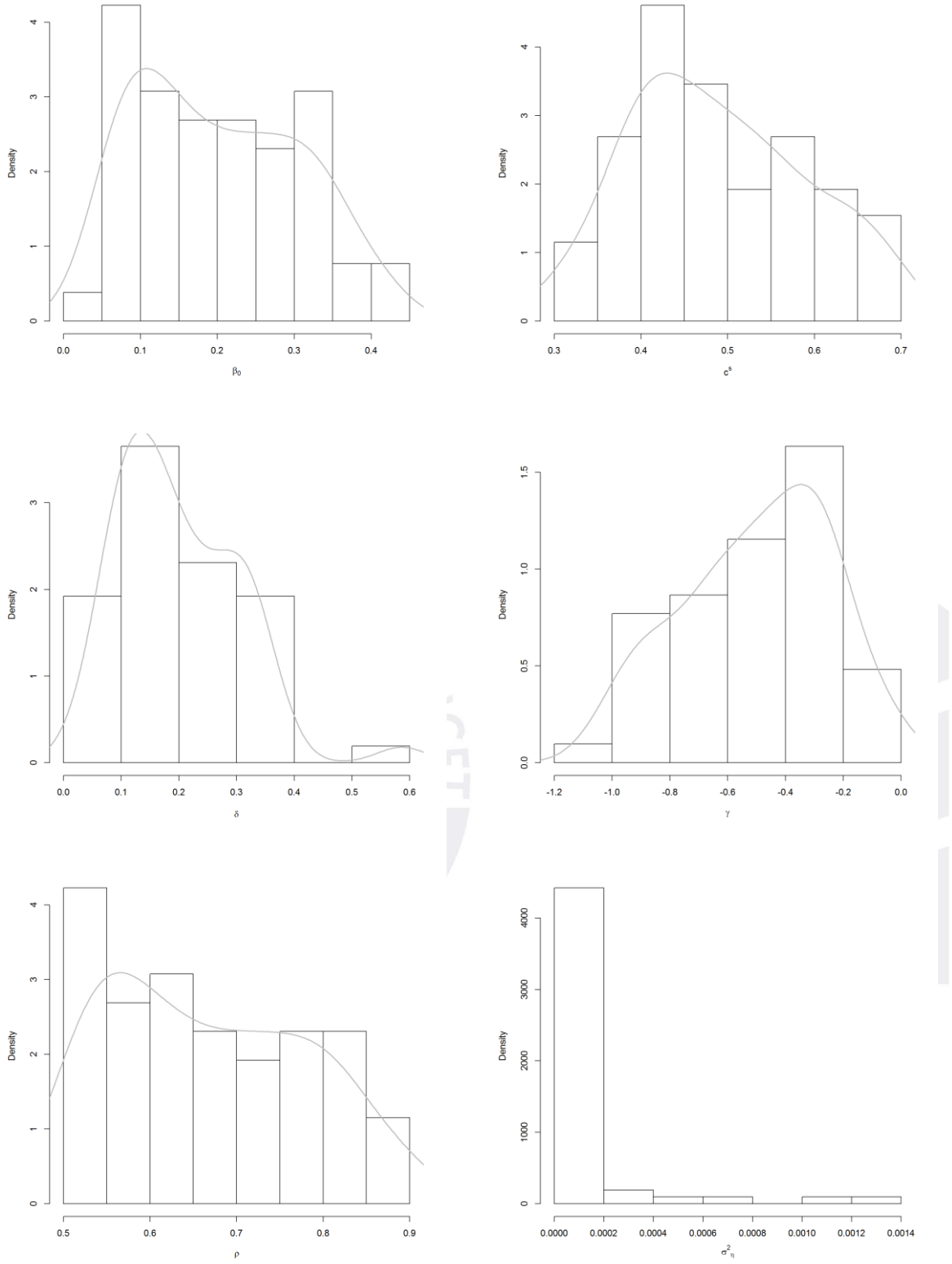


Figure E2. Distribution of other idiosyncratic parameters in the operational loss model.

Appendix F: Covariate-based technique regressions

Table F1

Severity regressions with GAMLSS

Mean regression	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Intercept	-2.80	-2.78	-2.43	-2.65	-2.47	-2.59
	(***)	(***)	(***)	(***)	(***)	(***)
GDP growth	0.01		0.01	0.01	0.01	0.01
	(*)		(*)	(*)	(*)	(*)
Crisis		-0.08				
		(**)				
CPI	-0.05	-0.05	-0.17	-0.12	-0.16	-0.10
	(***)	(***)	(***)	(***)	(***)	(.)
Government Effectiveness			0.33			
			(***)			
Regulatory Quality				0.21		
				(*)		
Rule of Law					0.29	
					(***)	
Control of Corruption						0.10
Control	-0.40	-0.63	-0.26	-0.44	-0.42	-0.42
	(*)	(***)		(*)	(*)	(*)
Employees per branch	0.13	0.14	0.13	0.14	0.14	0.13
	(***)	(***)	(***)	(***)	(***)	(***)
Assets per employee	-0.01	-0.01	-0.01	-0.00	-0.00	-0.01
	(.)	(*)	(*)			(***)
Scale regression						
Intercept	0.73	0.56	1.12	0.83	1.12	0.92
	(***)	(***)	(***)	(***)	(***)	(***)
GDP growth	0.01	-0.02				
	(*)					
CPI	-0.07	-0.02	-0.20	-0.15	-0.21	-0.11
	(***)	(.)	(***)	(***)	(***)	(***)
Employees per branch	-0.01	-0.01	-0.02	-0.02	-0.02	-0.01
	(***)	(***)	(***)	(***)	(***)	(***)
Diagnostics						
Global Deviance	-6657.35	-6633.58	-6677.04	-6718.08	-6732.36	-6658.17
AIC	-6581.01	-6553.02	-6598.11	-6610.79	-6624.83	-6577.47
SBC	-6344.62	-6303.55	-6353.68	-6278.53	-6291.84	-6327.55

Notes:

1) All models include additional repressors that are defined in terms of smoothed terms. They are not reported here because they are used as additional controls. Smoothing is performed with p-splines.
2) Significance codes are 0 = ***, 0.001='**', 0.01='*', 0.05='.', 0.1=' '.

Table F2

Frequency regressions with GAMLSS

Mean regression	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Intercept	4.33	4.32	4.28	4.23	4.22	4.24	4.40
	(***)	(***)	(***)	(***)	(***)	(***)	(***)
GDP growth	0.13	0.13	0.13	0.13	0.14	0.14	0.12
	(***)	(***)	(***)	(***)	(***)	(***)	(***)
Dummy for crisis		0.04					
CPI			0.02				
Gov. effectiveness				0.12			
Regulatory Quality					0.13		
Rule of Law						0.17	
Control of Corruption							0.00
Control	-3.50	-3.51	-3.68	-3.70	-3.69	-3.83	-3.69
	(***)	(***)	(***)	(***)	(***)	(***)	(***)
Assets per employee	-0.03	-0.03	-0.03	-0.03	-0.03	-0.04	-0.04
	(.)	(*)	(*)	(*)	(*)	(*)	(*)
Scale regression							
Intercept	-0.14	-0.14	-0.11	-0.10	-0.07	-0.07	-0.47
Employees per branch	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	(**)	(**)	(**)	(**)	(**)	(**)	(**)
Diagnostics							
Global Deviance	1581.63	1581.58	1570.44	1581.11	1570.75	1562.64	1535.25
AIC	1611.76	1613.72	1606.82	1612.27	1609.40	1604.36	1584.57
SBC	1663.01	1668.40	1668.74	1665.30	1675.15	1675.34	1668.48

Notes:

1) All models include additional repressors that are defined in terms of smoothed terms. They are not reported here because they are used as additional controls. Smoothing is performed with p-splines.
2) Significance codes are 0 = ***, 0.001='**', 0.01='*', 0.05='.', 0.1=' '.