TOWARDS A FRAMEWORK FOR OPERATIONAL RISK IN THE BANKING SECTOR

by

Mphekeleli Hoohlo

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Risk Theory (Finance)

Approved:

Eric Schaling, Ph.D.
Supervisor

Thanti Mthanti, Ph.D.
Co-supervisor

Odongo Kodongo, Ph.D.
Panel Member (WBS)

Thabang Mokoaleli-Mokoteli, Ph.D.
Panel Member (WBS)

Christopher Malikane, Ph.D.
Panel Member (SEBS)

Paul Alagidede, Ph.D.
Panel Member (WBS) and Academic Director
of Wits Graduate School of Business

UNIVERSITY OF THE WITWATERSRAND
Johannesburg, Gauteng

2019

ABSTRACT

Towards a Framework for Operational Risk
in the Banking Sector

by

Mphekeleli Hoohlo
University of the Witwatersrand, 2019

Major Professor: Eric Schaling, Ph.D.
Supervisor: Thanti Mthanti, Ph.D.
Department: Law, Commerce & Management

There have been a series of destructive events that have threatened the stability of the financial system due to (OpRisk). In most, if not all of these cases, human error is at the center of the chain of events that lead or may lead to (OpRisk) losses. There are many attitudes that can potentially infect organisational processes, the most persistent of these attitudes stem from human failings that are exploitable Barberis and Thaler (2003), thus forming a basis for the theoretical foundation of `OpRisk`.

Shefrin (2016) notes that people would rather incur greater risks to hold on to things they already have, than the risks they would taken to get into that position in the first place, thereby risking a banks' survival, rather than expose their trading losses by consciously deceiving senior management to hide unethical operational practices. In this paper the application of machine learning techniques on the observed data demonstrates how these issues can be resolved given their flexibility to different types of empirical data.

(116 pages)

PUBLIC ABSTRACT

Towards a Framework for Operational Risk

in the Banking Sector

Mphekeleli Hoohlo

The purpose of this research is to provide clarity; based on theory and empirical evidence, on how to tackle the specific problems in the *operational risk* (OpRisk) literature, which have earned a place in modern day recource in in risk and finance, due to how significantly its importance has increased over the last few decades. During this period, until present day, there have been and continues to be series of destructive events that have threatened the stability of financial systems due to OpRisk. In most, if not all of these cases, human error is at the center of the chain of events that lead or may lead to (OpRisk) losses. There are many attitudes that can potentially infect organisational processes, the most persistent of these attitudes stem from human failings that are exploitable Barberis and Thaler (2003), thus forming a basis for the theoretical foundation of `OpRisk`.

Shefrin (2016) notes that people would rather incur greater risks to hold on to things they already have, than the risks they would taken to get into that position in the first place, thereby risking a banks' survival, rather than expose their trading losses by consciously deceiving senior management to hide unethical operational practices. In this paper the application of machine learning techniques on the observed data demonstrates how these issues can be resolved given their flexibility to different types of empirical

DEDICATION

This work—the dissertation and all work associated with it—is dedicated to …. I will always be grateful to and for . I also dedicate this work to my child, who patiently loved a father who was often busy working, even when at home. Finally, I dedicate this work to my parents and siblings, who kept me level-headed throughout the process, providing wise and thoughtful advice.

This work is truly evidence of the love I am surrounded by.

Mphekeleli Hoohlo

ACKNOWLEDGEMENTS

Several individuals deserve recognition for their help in this work. First, I am greatly indebted to ….. for giving me the opportunity to find where my research passions are—especially quantitatively. …. provided guidance in a productive and capable researcher. I also owe a great deal ….. when my application was delayed. He has also proven to be an important mentor. Relatedly, I am grateful to providing opportunities to develop as a researcher in an important field.

My friends and colleagues in the program also deserve serious recognition for their help and support. Specifically, as I pursued my goals in the program. …. provided timely and important help throughout the process. I appreciate .

I also want to acknowledge my committee, in providing feedback, guidance, and support throughout my graduate student career. The department's and college's administrative staff and heads, although working in the background, were a huge help in getting me to where I am today. Finally, Wits University as a school has proven to be a fantastic place to grow academically and to mature into a researcher.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Purpose of the study

The purpose of this research is to apply a generalised linear model (GLM) suitable for exposure-based operational risk (EBOR) treatments within the operational risk management framework (ORMF), effectively replacing historical loss severity curves obtained from historical loss counts, by forward-looking measures using event frequencies based on actual operational risk (OpRisk) exposures. Preliminary work on EBOR models was undertaken by (Einemann, Fritscher, and Kalkbrener, 2018). Secondly, this study provides a comprehensive computational comparison of various data-intensive techniques amongst each other, and versus *classical* statistical estimation methods for classification and regression performances.

Our understanding of existing ORMF to date is limited to the assumption that financial institutions (FI's) are risk-neutral. Thirdly, in lieu of the aforementioned, this study finally seeks to invalidate the risk-neutral assumption, by means of various unsupervised learning techniques, by proposing that FI's are more risk-averse; this can be measured by analysing subtle patterns between data features and trends in the allocated risk capital estimates. In theory, a risk manager who experiences persistent/excessive losses due to particular risk events, would over-compensate cover for these particular risk types, and this would show in reduced losses in these types over time.

Fundamentals of ORMF's

Most banks' estimates for their risk are divided into credit risk (50%), market risk (15%) and OpRisk (35%). Cruz (2002) postulated that OpRisk, which focuses on the human side of risk management is difficult to manage with the reduced ability to measure it. The process of OpRisk, that is, the how manifests in conscious and/or unconscious states of the risk manager/s (Hemrit and Arab, 2012), and encompasses approaches and theories that focus on how one will choose when faced with a decision, based on how comfortable they are with the situation and the variables that are present.

Definition of operational risk

Operational risk (OpRisk) is defined as: *The risk of loss resulting from inadequate or failed internal processes, people and systems, and from external events. This definition includes legal risk, but excludes strategic and reputational risk.*(Risk, 2001).

A major managerial concern for businesses is an inability to identify and account for their susceptibility to OpRisk events following a number of very costly and highly publicized operational losses, in particular, it became popular following a fraudulent trading incident which was responsible for a catastrophic loss that lead to the collapse of Barings Bank (the UK's oldest bank) in 1995.

The term OpRisk began to be used after the afore-mentioned and similar types of OpRisk events became more common. A (rogue) trader (Nick Leeson), who risked the banks' survival rather than expose his trading losses, by consciously deceiving senior management to hide his unethical acts, was found to have been responsible for unethical trading practices when he created illegal trades in his account, then used his position in the front and back offices of the bank to hide his trading losses. Worse still, he incurred a greater risk to the bank by lying in order to give a false impression of his profits. Shefrin (2016) notes that people would rather incur greater risks to hold on to things they already have, than the risks they would taken to get into that position in the first place.

It was later discovered that he was placing illegal bets in the Asian-markets, and kept these contracts out of sight from senior management to cover up his illegal activity. When his fraudulent behaviour was discovered (after an earthquake hit at Kobe in Japan, that collapsed the Osaka Securities Exchange) he succumbed to unrecoverable losses due to trading positions he had accumulated, which resulted in a loss of around £1.3 billion to the bank, thus resulting in it's collapse. In most, if not all of these cases, human error is at the center of the chain of events that lead or may lead to OpRisk losses.

Since then, there have been a series of destructive events that have threatened the stability of the financial system due to OpRisk. Large fines have been imposed on the culprits and regulatory scrutiny has been heightened as a result of a number of operational events, e.g. the January 2016 "Dark Pool" trading penalties suffered by Barclays ($70mn) and Credit Suisse ($85mn), imposed by the United States (US) based securities exchange commision (SEC). These OpRisk loss events were due to fraudulent trading activity consisting of rogue traders dealing in illegally placed high frequency trades for private clients where prices were hidden.

In South Africa (SA), there is an upcoming case of price fixing and market allocation in trading foreign exchange (FX) currency pairs, reffered to the SA based competition tribunal for prosecution. Absa bank, Standard bank & Investec may be liable to payment of an admistrative penalty equal to 10% of their annual turnover in 2016, following accusations by the local based competition commission in February 2017, of rogue traders manipulating the price of the rand through buying and selling US dollars in exchange for the rand at fixed prices. According to the competition commission, it has been alleged that currency traders have been colluding or manipulating the price of the rand through these buy and sell orders to change supply of the currency.

This has compromised the quality and accuracy of risk management's advisory service and pedigree, and aroused huge interest as the value of the rand has implications on South African's. Furthermore, this kind of behaviour can lead to catastrophic operational losses, as with the case for the Barings event, resulting is a mismatch between business' expectations and the value the risk management practice was able to deliver, which is prevalent across FI's and remains unchanged. There are many attitudes that can potentially infect organisational processes, the most persistent of these attitudes stem from human failings that are exploitable (Barberis and Thaler, 2003); i.e. humans' propensity to be deceitful during periods of distress, thus forming a basis for a theoretical foundation of OpRisk management.

Basel Committee's quantitative operational risk management framework

The Bank for International Settlements (BIS) is an organisation consisting of a group of central bank governors and heads of supervision of central banks around the world who represent an authority on good risk management in banking. More specifically, the BIS oversee the duties of the Basel Committee on Banking Supervision (BCBS)/Basel Commitee. The role of the BCBS is to set out guidelines on international financial regulation to cover risks in the banking sector. There have been three banking accords from the BCBS under the supervision of the BIS in dealing with financial regulation, viz., Basel I, Basel II & Basel III. These accords describe an overview of capital requirements for financial institutions (FI's) in order to create a level playing field, by making regulations uniform throughout the world.

The Capital Adequacy Accord (Basel I)

Basel I was established in 1988. Basel I meant that FI's were required to assign capital for credit risk to protect against credit default. In 1996, an amendment to Basel I imposed additional requirements to cover exposure due to market risk as well as credit risks. Basel I effectively minimised rules that favoured local FI's over potential foreign competitors, by opening up global competition so that these banks could buffer against international solvency. In 2001, the Risk (2001) consultative package provided an overview of the proposed framework for regulatory capital (RC) charge for OpRisk. A fiancial institution (FI) has an OpRisk component, which constitutes a substantial risk component other than credit and market risk. There are two types of OpRisk's viz., potential high severity risk where the probability of an extreme loss is very small but costly, and high frequency/low severity risk where frequency plays a major role in the OpRisk capital charge calculation.

New Capital Adequacy Accord (Basel II)

The framework for Basel II was implemented in June 2006. The rationale for Basel II is to introduces risk sensitivity through more restrictive capital charge measures and flexibility with specific emphasis on OpRisk. The structure of the new accord is built upon a three-pillar framework: Pillar I stipulates minimum capital requirements for the calcualtion of regulatory capital for credit risk, market risk and OpRisk in order to retain capital to ward against these risks. Pillar II imposes a supervisory review process through which additional requirements can be imposed, such as the bank's internal capital assessements, or to act on needed adequate capital support or best practice for mitigating their risks. Pillar III relates to market discipline, i.e. transparency requirements which require banks to publicly provide risk disclosures to keep them in line by enabling investors to form an accurate view of their capital adequacy, in order to reward or punish them on the basis of their risk profile.

Basel III

Basel III establishes tougher capital standards through more restrictive capital definitions, higher RWA's, additional capital buffers, and higher requirements for minimum capital ratios (Dorval, 2013). Through Basel III, the BCBS is introducing a number of fundamental reforms grouped under three main headings (Committee and others, 2010): 1] A future of more capital through incremental trading book risk (credit items in trading book treated in the same way as if they were in banking book), 2] More liquidity through the introduction of a global liquidity risk standard (Basel III will push banks toward holding greater levels of liquid instruments, such as government bonds and more liquid corporate instruments), and 3] Lower risk under the new requirements of the capital base, i.e., establish more standardized risk-adjusted capital requirements.

Regarding the sequence Basel I and Basel II: Regulation begins as a qualitative recommendation which requires banks to have an assets-to-capital multiple of at least 20, then focuses on ratios in which both on-balance sheet and off-balance sheet items are used to calculate the bank's total risk-weighted assets (RWA's)[1], then on tail risk. In other words, auditors' discretion is replaced by market perception of capital, meaning there is a market risk capital charge for all items in the trading business line, then exciting new static risk management approaches which involve calculating a 99.9 percentile left tail confidence interval to measure OpRisk value-at-risk (VaR) and convert it into a RC charge.

The future regulatory environment requires OpRisk professionals, who are not only intelligent, creative and motivated but also have the courage to uphold the

[1] Also reffered to as risk-weighted amount, it is a measure of the bank's total credit exposure

OpRisk advisory service standards. Businesses that want to successfuly manage risk, would be well advised to utilize new theoretical and empirical techniques, such that large and small scale experiments play an important role in risk analysis and regulatory research.

Modern OpRisk measurement frameworks (ORMF's)

Basel II describes three methods of calculating capital charge for OpRisk RC viz., the standardised approach (SA), the basic indicator approach (BIA) and the internal measurement approach (IMA). The basic indicator approach (BIA) sets the OpRisk RC equal to a percentage (15%) of the annual gross income of the firm as a whole to determine the annual capital charge. The SA is similar to the BIA except the firm is split into eight business lines and assigned a different percentage of a three year average gross income per business line, the summation of which is the capital charge (Hoohlo, 2015). In the IMA, the bank uses it's own internal models to calculate OpRisk loss.

Advanced Measurement Approach (AMA)

The advanced measurement approach (AMA) is an IMA method which applies estimation techniques of OpRisk capital charge derived from a bank's internal risk measurement system Cruz (2002). Basel II proposed measurement of OpRisk to define capital requirements against unexpected bank losses whereas the unexpected loss (UL) is the quantile for the level α minus the mean. According to the AMA, which is thought to outperform the simpler SA approach and the BIA, RC requirements are defined according to the UL limit in one year and the loss distribution at a 99.9% confidence level ($\alpha = 0.01\%$) aggegate loss distribution[2] used as a

[2] The aggregate loss distribution is obtained by convoluting a loss event frequency distribution and a loss severity distribution by means of the random sums method.

measure of RC. The BCBS proposes to define RC as $RC = UL$. This involves simulations based on historical data to establish frequency and severity distributions for losses. In this case the RC is a VaR measure.

The Basel III capital adequacy rules permit model-based calculation methods for capital, including the AMA for OpRisk capital. Under Basel III, standardised methods for OpRisk capital have been overhauled, however for a while there was no prospect of an overhaul of the AMA. Given the relative infancy of the field of OpRisk measurement, banks are mostly free to choose among various AMA principle-based frameworks to a significant degree of flexibility (Risk, 2016). A bank that undertakes an AMA should be able to influence their capital requirements through modeling techniques resulting in lowered pressure on OpRisk capital levels, which in turn has a positive impact on the bank.

A FI's ability to determine the framework used for its regulatory OpRisk RC calculation, evolves from how advanced the FI is along the spectrum of available approaches used to determine capital charge (Risk, 2001). BCBS recognizes that a variety of potentially credible approaches to quantify OpRisk are currently being developed by the industry, and that these R&D activities should be incentivised. Increasing levels of sophistication of OpRisk measurement methodologies should generally be rewarded with a reduction in the regulatory OpRisk capital requirement.

The standardised measurement approach (SMA)

The flexibility of internal models was expected to narrow over time as more accurate OpRisk measurement was obtained and stable measures of RC were reached, ultimately leading to the emergence of best practice. Instead, internal models produced wildly differing results of OpRisk RC capital from bank to bank, contrary to the expectations of the BCBS. In March 2016, the BCBS published for consultation a standardised measurement approach (SMA) for OpRisk RC; that proposes to abandon the freedom of internal modelling (thus ending the AMA) approaches for OpRisk RC, in exchange for being able to use a simple formula to facilitate comparability across the industry.

Under the SMA, RC will be determined using a simple method comprising of two components: A stylised systemic risk model (business indicator component), and an idiosyncratic risk model (loss component), which are combined via an internal loss multiplier (ILM), whose function is to link capital to a FI's operational loss experience to determine SMA capital.

The SMA formula is thought to be consistent with regulators' intent for simplification and increased comparability across most banks. However, there is a feeling from some in the banking industry that the SMA is disadvantaged as it is not the same as measuring OpRisk. Mignola, Ugoccioni, and Cope (2016) and Peters, Shevchenko, Hassani, and Chapelle (2016) identified that the SMA does not respond appropriately to changes in the risk profile of a bank i.e., it is unstable viz., two banks of the same risk profile and size can exibit OpRisk RC differences exceeding 100%, and risk insensitive; that SMA capital results generally appear to be more variable across banks than AMA results, where banks had the option of fitting the loss data to statistical distributions.

Argument

Over the last twenty years, hard-won incremental steps to develop a measure for the size of OpRisk exposure along with the emergence of promising technologies presents a unique opportunity for bankers and treasurers - traditionally risk-averse players - to develop a novel type of way of looking at decision making under risk/uncertainty. New technologies have been introduced which make use of up to date technical solutions (such as homo heuristics developed by Gigerenzer and

Brighton (2009), who mainatain their methods solve practical finance problems by simple rules of thumb, or Kahneman (2003)'s intuitive judgements and deliberate decision making), argued to more likely represent the true embedded OpRisk in financial organisations as these methods are designed to fit normal behavioral patterns in their formulation, which is consistent with how decisions are made under risk/uncertainty.

What are the important steps toward completing the post crisis reforms during the current year? Should the risk management fraternity follow the chartered[3] path followed in the Risk (2016) consultative document, scrapping away twenty years of internal measurement approaches (such as the AMA), or should the focus of financial regulators shift toward improving on what they see fit within current existing AMA frameworks. The question is should OpRisk managements' focus be on stimulating active discussions on practical approaches to quantify, model and manage OpRisk for better risk management and improved controls, or abandon the adoption of innovative measurement approaches, such as the AMA, in exchange for being able to use a simple formula across the whole industry?

Context of the study

Regulatory reforms are designed and fines imposed to protect against operational errors and other conduct costs connected with wrongdoing and employee misconduct. Despite the introduction and use of these seemingly robust strategies, regulations, processes and practices relating to managing risk in FI's, bank losses continue to occur at a rather distressing frequency. A cyclical pattern of OpRisk loss events still persists; as evidenced in the recent price fixing and collusion cases,

[3] Meaning as of the publication [@risk2016supporting] the methods brought forth in the consultative document have not been approved for the public, the ideas within an experimental (leased) phase for the exclusive use of BCBS and certain FI's

defeating the explicit objectives of risk management frameworks. This demonstrates a scourge of reflexivity prevailing in financial markets emphasising that, there are theories that seem to work for a time only to outlive their use and become insufficient for the complexities that arise in reality.

Why ***OpRisk?***

A forceful narrative in management theory is that an organisation running effective maintenance procedures combined with optimal team and individual performers i.e., the right balance of skills in the labour force and adequate technological advancements, means systems and services can be used to more efficiently produce material gains, enhance organisational effectiveness, meet business objectives and increase investment activity. Conversely, the risk of the loss of business certainty associated with lowered organisational competitiveness and inadequate systems technology that underpins operations and services is a key source leading to a potential breakdown in investment services activity (Hoohlo, 2015). In fact OpRisk control could set banks apart in competition. This serves as an incentive to support regulation, particularly Basel III recovery and resolution processes.

Consider the case of a regulator in a financial system, who assumes that he/she is consiously and accurately analysing an observed subject, trusting the validity and relying on the visual information that their sense of sight reveals. In the absence of visual confirmation they are hindered from extracting and/or analysing information about the system and their efforts to regulate could potentialy fail. The organisational methods and functioning of current information systems in this industry sector obscure the full extent of OpRisk challenges from the eyes of the risk practitioner.

When an attack such as an operational error occurs at a speed that the OpRisk agent (an individual legal entity or a group) is unable to react quickly enough, due to limitations of their processing speed, and they are not able to process all the information in the given time span, they could lose control/fail to comply with regulatory standards. The latter case is more often than not the most accurate reflection of current risk management practices. The agent represents one end of the spectrum of a risk management strategy, which mitigates risk and enforces regulation, dependent on the information recieved. The other end of the spectrum is one which does not react at all to changes in the system environment.

Current conventional financial systems where information processing is slow and have a tendency to rely on manual, uncertain, unpredictable and unrealistic controls, obscure risk management reporting and produce undesirable market conditions. The OpRisk management function should be able to assist the firms' ability to mitigate risks by acquiring and/or refining risk management solutions which deliver reliable and consistent benefits of improved control and management of the risks inherent in banking operations (Dorval, 2013). This proposal attempts to fill the gap in the current system where there is a risk management information lag or an obstruction from the eyes of the risk practitioner.

Analysis and interpretation issues with behavioral finance theory

Behavioral management theory is very much concerned with social factors such as motivation, support and employee relations. A critical component of behavioral finance is building models which better reflect actual behavior. Studies have revealed that these social factors are not easy to incorporate into finance models or to understand in the traditional framework.

The traditional finance paradigm seeks to understand financial markets using models in which agents are "rational". According to Barberis and Thaler (2003), this means that agents update their beliefs on the onset of new information, and that given their beliefs, they make choices that are normatively acceptable, and that most people do this most of the time. Neoclassical theory has grown to become the primary take on modern-day economics formed to solve problems for decision making under uncertainty/risk. Expected Utility Theory (EUT) has dominated the analysis and has been generally accepted as the normative model of rational choice, and widely applied as a descriptive model of economic choice (Kahneman and Tversky, 2013).

Expected utility theory

Expected utility theory[4] (EUT): We see a fundamental relation for expected utility (Expectation) of a contract X, that yields outcome x_i with probability p_i, where $X = (x_1, p_1; ...; x_n, p_n)$ and $p_1 + p_2 + \ldots + p_n = 1$ given by:

$$U(x_1, p_1; \ldots; x_n, p_n) = p_1 u(x_1) + \ldots + p_n u(x_n) \tag{1.1}$$

corroborated by Morgenstern and Von Neumann (1953); Friedman and Savage (1948); Kahneman and Tversky (2013) & others.

A common thread running through the rational viz., the neoclassical take of modern-day economics vs the non-neoclassical schools of thought are findings of behavioral economics which tend to refute the notion that individuals behave rationally. Many argue that individuals are fundamentally irrational because they do not behave rationally giving rise to a literature and debates as to which heuristics and sociological and institutional priors are rational (Altman, 2008).

In the real world there is a point of transition between the traditional (neoclassical) approach to decision making, based on data and data anaysis (logic and rational), by adding new parameters and arguments that are outside rational conventional thinking but are also valid. For example, that neoclassical theory makes use of the assumption that all parties will behave rationally overlooks the fact that

[4] Expected utility theory provides a model of rationality based on choice.

human nature is vulnerable to other forces, which causes people to make irrational choices.

An essential ingredient of any model trying to understand trading behavior is an assumption about investor preferences (Barberis and Thaler, 2003), or how investors evaluate risky gambles. Investors systematically deviate from rationality when making financial decisions, yet as acknowledged by Kuhnen and Knutson (2005), the mechanisms responsible for these deviations have not been fully identified. Some errors in judgement suggest distinct mental operations promote different types of financial choices that may lead to investing mistakes. Deviations from the optimal investment strategy of a rational risk neutral agent are viewed as risk-seeking mistakes and risk-aversion mistakes (Kuhnen and Knutson, 2005).

Theoretical investigations for the quantification of moderm ORMF

Kuhnen and Knutson (2005) explain that these risk-seeking choices (such as gambling at a casino) and risk-averse choices (such as buying insurance) may be driven by distinct neural[5] phenomena, which when activated can lead to a shift in risk preferences. Kuhnen and Knutson (2005) found that certain areas of the brain precede risk-seeking mistakes or risky choices and other areas precede risk-aversion mistakes or riskless choices. A risk-aversion mistake is one where a gamble on a prospect of a gain is taken by a risk-averse agent in the face of the chance of a prospective loss. The fear of losing prohibits one's urge to gamble, but people engage in gambling activity anyway. Barberis and Thaler (2003) show that people regularly deviate from the traditional finance paradigm evidenced by the extensive experimental results compiled by cognitive psycologists on how people make decisions given their beliefs.

[5] As recent evidence from human brain imaging has shown [@kuhnen2005neural] linking neural states to risk-related behaviours [@paulus2003increased].

Kahneman and Tversky (2013) maintains, preferences between prospects which violate rational behaviour demonstrate that outcomes which are obtained with certainty are overweighted relative to uncertain outcomes. This will contribute to a risk-averse preference for a sure gain over a larger gain that is merely probable or a risk-seeking preference for a loss that is merely probable over a smaller loss that it certain. As a psycological principle, overweighting of certainty favours risk-aversion in the domain of gains and risk-seeking in the domain of losses.

The present discussion replicates the common behavioral pattern of risk aversion, where people weigh losses more than equivalent gains. Furthermore, neuroeconomic research shows that this pattern of behavior is directly tied to the brain's greater sensitivity to potential losses than gains (Tom, Fox, Trepel, and Poldrack, 2007). This provides a target for investigating a more comprehensive theory of individual decision-making rather than the rational actor model and thus yield new insights relevant to economic theory[6] (Kuhnen and Knutson, 2005).

If people are reasonably accurate in predicting their choices, the presence of systematic violations of risk neutral behavior provides presumptive evidence against this i.e., people systematically violate EUT when choosing among risky gambles. This seeks to improve and adapt to reality and advance different interpretations of economic behaviour; viz., to propose a more adequately descriptive model, that can represent the basis for an alternative to the way the traditional model is built for decisions taken under uncertainty. This has led some influential commentators to call for an entirely new economic paradigm to displace conventional neoclassical theory with a psycologically more realistic preference specification (List, 2004).

[6] Representing ability of FI's financial market models to characterise the repeated decision-making process that applies to loss aversion

A new class of ORMF models approach

A substantial body of evidence shows that decision makers systematically violate EUT when choosing between risky prospects. Indeed, people would rather satisfy their needs than maximise their utility, contravening the normative model of rational choice (i.e., EUT) which has dominated the analysis of decision making under risk. In recent work (Barberis and Thaler, 2003) in behavioral finance, it has been argued that some of the lessons learnt from violations of EUT are central to understanding a number of financial phenomena. In response to this, there has been several theories put forward advocating for the basis of a slightly different intepretation which describes how individuals actually make decisions under uncertainty/risk. Of all the non-EUT's, we focus on Prospect Theory (PT) as this framework has had most success matching most empirical facts[7].

Kahneman and Tversky (2013) list the key elements of PT, which are 1] a value function, and 2] a non-linear transformation of the probability scale, that factors in risk aversion of the participants. According to Kahneman and Tversky (2013), the probability scale overweights small probabilities and underweights high probabilities. This feature is known as loss/risk aversion: This means that people have a greater sensitivity to losses (around 2.5 times more times) than gains, and are especially sensitive to small losses unless accompanied by small gains[8]. Loss aversion is a strong differentiator when it comes to explaining exceptions to the general risk patterns that characterize prospect theory.

[7] OpRisk loss events in FI's are largely due to human failings that are exploitable e.g., fraudulent trading activity, and PT is based on the same behavioural element of how people make financial decisions about prospects

[8] Diminishing marginal utility for gains but opposite for losses.

Prospect theory

By relaxation of the expectation principle in equation **??**, the over-all value $\bigvee$ of the regular prospect $(x, p; y, q)$: In such a prospect, one receives x with probability p, y with probability q, and nothing with probability $1 - p - q$, is expressed in terms of two scales, $\pi(\cdot)$, and $\nu(\cdot)$, where $\pi(\cdot)$ is a decision weight and $\nu(\cdot)$ a number reflecting the subjective value of the outcome. Then $\bigvee$ is assigned the value:

$$\bigvee = \pi(p)\nu(x) + \pi(q)\nu(y) \qquad \text{iff} \qquad p + q \leq 1 \tag{1.2}$$

The scale, π, associates with each probability p a decision weight which reflects the impact of p on the over-all value of the prospect. The second scale, ν, assigns to each outcome x a number $\nu(x)$, which measures the value of deviations from a reference point i.e., gains or losses. π is not a probability measure and $\pi(p) + \pi(1 - p) < 1$. Through PT we add new parameters and arguments to improve the mathematical modelling method for decisions taken under risk/uncertainty, such that the value of each outcome is multiplied by a decision weight, not by an additive probability.

PT looks for common attitudes in people (in FI's) with regard to their behaviour toward taking financial risks or gambles that cannot be captured by EUT. In light of this view, people are not fully invested in either of the percieved outcomes x and y, Which tells us that $p + q \leq 1$. In lieu of this, an FI using (internal) historical OpRisk loss data to model future events; say a historical case of fraud at the FI occurs and is incorporated in the model, the probability of making the same error in future is provided for in the model versus risk events that haven't happened. The modelled future should over-provide for the loss events that have already occured, which fits normal patterns around individuals psycological make up and is consistent with risk-averse behavior. The idea at the basis of PT is that a better modeling method can be obtained which leads to a closer approximation of the over-all-value of OpRisk losses.

Modeling

In this study, an important new algorithm for ORMFs and is laid out coupled with data intensive estimation techniques; viz. Generalised Additive Models for locatin Scale & Shape (GAMLSS), Generalized Linear Models (GLDs), Artificial Neural Networks (ANNs), Random Forest (RF) & Decision Trees (DTs), which have capabilities to tease out the deep hierarchies in the features of covariates irrespective of the challenges associated with the non-linear or multi-dimensional nature of the underlying problem, at the same time supporting the call from industry for a new class of EBOR models that capture forward-looking aspects. Machine Learning (ML) is used as a substitute tool for the traditional model based Autoregressive Moving Average (ARMA) used for analysing and representing stochastic processes. As opposed to the statistical tool, ML does not impose a functional relationship between variables, the functional relationship is determined by extracting the pattern of the training set and by learning from the data observed.

Using computationally intensive (using ML techniques on historical data) OpRisk measurement techniques and mixing with a theory is not a new approach for modeling, particularly in calculating OpRisk RC; as evidenced through Agostini, Talamo, and Vecchione (2010) in a study whereby the LDA model for forecasting OpRisk RC, via VaR, was implemented in conjunction with the use of advanced credibility theory (CT). The idea at the basis of their use of CT, is to advance the very recent literature that a better estimation of the OpRisk RC measurement can be obtained by integrating historical data and scenario analysis i.e., combining the historical simulations with scenario assessments through formulas that are weighted averages of the historical data entries and scenario assessments, advocating for the combined use of both experiences.

However, applying ML is an original way of looking at the approximation issue as opposed to advanced CT. The essential feature of PT are assumptions which are more compatible with basic principles of perception and judgement for decisions taken under uncertainty, whereas ML will reveal additional chance probabilities determined through the natural clusters of unknown data feature findings from which new discoveries are made.

According to Kahneman and Tversky (2013), the decision maker, who is a risk agent within the FI, constructs a representation of the losses and outcomes that are relevant to the decision, then assesses the value of each prospect and chooses according to the losses (changes in wealth), not the overall financial state of the FI. We wish to bring the prescribed model to equilibrium, by applying a method that tries to establish what accurately ascribes to decision rules that people wish to obey, in made predictions about what operational loss events might result in the future, then use empirical data to test this idea in a way that is falsifyable.

Problem statement

Main problem

The existing models of OpRisk VaR measurement frameworks assume FI's are risk neutral, and do not learn from past losses/mistakes: We address weaknesses in current OpRisk VaR measurement frameworks by assuming that FI's are more risk averse. Furthermore, introducing exposure-based operational risk modeling, we gain an understanding of how capturing past losses and exposures of forward looking aspects affect risk attitudes using machine learning techniques. As a consequence, projected future losses are estimated through a learning algorithm adapting capital estimates to changes in the risk profile, i.e. in the introduction of new products or changes in the business mix of the portfolio (e.g. mergers, trade terminatons, allocations or disinvestments), providing sufficient incentives for OpRisk management to mitigate risk.

Objectives of the study

The research objectives are three-fold:

Exposure-based OpRisk (EBOR) models

To quantify OpRisk losses by introducing generalised additive models for location, scale and shape (GAMLSS) in the framework for OpRisk management, that captures exposures to forward-looking aspects of the OpRisk loss prediction problem. EBOR treatments effectively replace historical loss severity curves obtained from historical loss counts, by looking into deep hierarchies in the features of covariates in investment banking (IB), and by forward-looking measures using event frequencies based on actual operational risk (OpRisk) exposures in the business environment and internal control risk factors (BEICF) thereof.

Modeling OpRisk depending on covariates

To investigate the performance of several supervised learning classes of data-intensive methodologies for the improved assessment of OpRisk against current *traditional* statistical estimation techniques. Three different machine learning techniques viz., DTs, RFs, and ANNs, are employed to approximate weights of input features (the risk factors) of the model. A comprehensive list of user defined input variables with associated root causes contribute to the *frequency* of OpRisk events of the underlying value-adding processes. Moreover, the *severity* of OpRisk is also borne out through loss impacts in the dataset . As a consequence of theses new mwthodologies, capital estimates should be able to adapt to changes in the risk profile of the bank, i.e. upon the addition of new products or varying the business mix of the bank providing sufficient incentives for ORMF to mitigate risk (Einemann et al., 2018).

Interpretation Issues using cluster analysis

To identify potential flaws in the mathematical framework for the loss distribution approach (LDA) model of ORM, which is based the derivation of OpRisk losses based on a risk-neutral measure $\mathbb{Q}$, by employing Cluster Analysis (CA). The study addresses weaknesses in the current *traditional* LDA model framework, by assuming managerial risk-taking attitudes are more risk averse. More precisely, CA learns the deep hierarchies of input features[9] that constitute OpRisk event *frequencies* & *severities* of losses during banking operations. In theory, a risk manager who experiences persistent/excessive losses due to particular risk events, would over-compensate cover for these particular risk types. This would show in reduced losses in those loss event types over time, subsequently determining whether risk adverse techniques over-compensate for persistent losses.

Significance of the study

This study fills a gap in that advancing OpRisk VaR measurement methods beyond simplistic and traditional techniques, new data-intensive techniques offer an important tool for ORMFs and at the same time supporting the call from industry for a new class of EBOR models that capture forward-looking aspects of ORM (Embrechts, Mizgier, and Chen, 2018). The current *traditional* approach consists of a

[9] A typical approach taken in the literature is to use an unsupervised learning algorithm to train a model of the unlabeled data and then use the results to extract interesting features from the data [@coates2012learning]

loss data collection exercise (LDCE) which suffers from inadequate technologies at times relying on spreadsheets and manual controls to pull numbers together, and therefore do not support the use of data intensive techniques for the management of financial risks. In this study, a new dataset with unique feature characteristics is developed using an automated LDCE, as defined by Committee and others (2011) for internal data. The dataset in question is at the level of individual loss events, it is fundamental as part of the study to know when they happened, and be able to identify the root causes of losses arising from which OpRisk loss events.

This study will provide guidance on combining various supervised learning techniques with extreme value theory (EVT) fitting, which is very much based on the Dynamic EVT-POT model developed by Chavez-Demoulin, Embrechts, and Hofert (2016). This can only happen due to an abundance of larger and better quality datasets and which also benefits the loss distribution approach (LDA) and other areas of OpRisk modeling. In Chavez-Demoulin et al. (2016), they consider dynamic models based on covariates and in particular concentrate on the influence of internal root causes that prove to be useful from the proposed methodology. Moreover, EBOR models are important due to wide applicability beyond capital calculation and the potential to evolve into an important tool for auditing process and early detection of potential losses, culminating in structural and operational changes in the FI, hence releasing human capital to focus on dilemmas that require human judgement.

Organisation of the study

This study is made up of seven chapters. The introductory chapter is to the purpose, overview, research problem & objectives, and the significance of the study. The introductory chapter is succeded by a general literaty review chapter (two) followed by three stand alone chapters each focusing on the three research objectives regarding the issues in OpRisk capital requirement estimation.

Chapter one begins with an account of significance and a commentary on the nature and scope of the practical problem. It then provides a background of current issues when dealing with OpRisk measurement, the research problem and research questions thereof. Chapter two gives an overview of the literature concerning the LDA, an AMA technique used in the generation of `OpVaR`. It concludes by proposing the a research methodology in which a combination of ML techniques and statistical theory underlying ORMF's would benefit measurement of capital requirements for OpVaR.

Chapter three looks at the methodological and empirical determinants of OpRisk measurement. It explores the different dataset...

CHAPTER 2

LITERATURE REVIEW

Introduction

A look into literary sources for OpRisk indicates (Acharyya, 2012) that there is insufficient academic literature that looks to characterize its theoretical roots, as it is a relatively new discipline, choosing instead to focus on proposing a solution to the quantification of OpRisk. This chapter seeks to provide an overview of some of the antecedents of OpRisk measurement and management in the banking industry. As such, this chapter provides a discussion on why OpRisk is not trivial to quantify and attempts to understand its properties in the context of risk aversion with the thinking of practitioners and academics in this field.

According to Cruz (2002), FI's wish to measure the impact of operational events upon profit and loss (P&L), these events depict the idea of explaining the *volatility of earnings* due to OpRisk data points which are directly observed and recorded. By seeking to incorporate data intensive statistical approaches to help understand the data, the framework analyses response variables that are decidedly non-normal (including categorical outcomes and discrete counts) which can shed further light on the understanding of firm-level OpRisk RC. Lastly, a synopsis of gaps in the literature is presented.

The theoretical foundation of OpRisk

Hemrit and Arab (2012) argue that common and systematic operational errors in hypothetical situations poses presumtive evidence that OpRisk events, assuming that the subjects have no reason to disguise their preferences, are created sub-consciously. This study purports, supported by experimental evidence, behavioural finance theories should take some of this behaviour into account in trying to explain, in the context of a model, how investors maximise a specific utility/value function.

Furthermore its argued by integrating OpRisk management into behavioral finance theory,[1], that it may be possible to improve our understanding of firm level RC by refining the resulting OpRisk models to account for these behavioral traits - implying that people's economic preferences described in the model, have an economic incentive to improve the OpRisk RC measure.

Wiseman and Catanach Jr (1997) suggest that managerial risk-taking attitudes are influenced by the decision (performance) context in which they are taken. In essence, managerial risk-taking attitude is considered as a proxy for measuring OpRisk (Acharyya, 2012). In so doing, Wiseman and Catanach Jr (1997) investigate more comprehensive economic theories, viz. prospect theory and the behavioural theory of the firm, that prove relevant to complex organizations who present a more fitting measure for OpRisk.

In a theoretical paper, Wiseman and Catanach Jr (1997) discussed several organizational and behavioural theories, such as PT, which influence managerial risk-taking attitudes. Their findings demonstrate that behavioural views, such as PT and the behavioural theory of the firm explain risk seeking and risk averse behaviour in the context of OpRisk even after agency based influences are controlled

[1] In behavioral finance, we investigate whether certain financial phenomena are the result of less than fully rational thinking [@barberis2003survey]

for. Furthermore, they challenge arguments that behavioral influences are masking underlying root causes due to agency effects. Instead they argue for mixing behavioral models with agency based views to obtain more complete explanations of risk preferences and risk taking behavior (Wiseman and Catanach Jr, 1997).

Despite the reality that OpRisk does not lend itself to scientific analysis in the way that market risk and credit risk do, someone must do the analysis, value the RC measurement and hope the market reflects this. Besides, financial markets are not objectively scientific, a large percentage of successful people have been lucky in their forecasts, it is not an area which lends itself to scientific analysis.

Overview of operational risk management

It is important to note how OpRisk manifests itself: King (2001) has established the causes and sources of operational loss events as observed phenomena associated with operational errors and are wide ranging. By definition, the occurence of a loss event is due to P&L volatitlity from a payment, settlement or a negative court ruling within the capital horizon over a time period (of usually one year) (Einemann et al., 2018). As such, P&L volatitlity is not only related to the way firms finance their business, but also in the way they *operate*.

In operating practice, one assumes that on observing or on following instructions we are consciously analysing and accurately executing our tasks based on the information. However, the occurence of operational loss events indicates that there are sub-concious faults in information processing, which we are not consciously aware of. These operational loss events are almost always initiated at the dealing phase of a trading process, which more often than not implicates front office (FO) personnel to bear the responsibility for the losses e.g., during the trading process in cases where OpRisk events occur as a result of a mismatch between the trade booked (booking in trade feed) and the details agreed by the trader. The middle office (MO) and back offices (BO) conduct the OpRisk management, who undertake a broad view of P&L attribution carried out from deal origination to settlement within the perspective of strategic management, and detects the interrelationships between OpRisk factors with others to conceptualise the potential overall consequences (Acharyya, 2012) e.g., in the afore-mentioned example, human error (a subconscious phenomenon) is usually quoted as the source of error, and the trade is fixed by "amending" or manually changing the trade details.

Furthermore, Acharyya (2012) recognised that organizations may hold OpRisk due to external causes, such as failure of third parties or vendors (either intentionally or unintentionally), in maintaining promises or contracts. The criticism in the literature is that no amount of capital is realistically reliable for the determination of RC as a buffer to OpRisk, particularly the effectiveness of the approach of capital adequacy from external events, as there is effectively no control over them.

The loss collection data exercise (LCDE)

The main challenge in OpRisk modeling is in poor loss data quantities, and low data quality. There are usually very few data points and are often characterised by high frequency low severity (HFLS) and low frequency high severity (LFHS) losses. It is common knowledge that HFLS losses at the lower end of the spectrum tend to be ignored and are therefore less likely to be reported, whereas low frequency high severity losses (LFHS) are well guarded, and therefore not very likely to be made public.

In this study, a new dataset with unique feature characteristics is developed using the official loss data collection exercise (LDCE), as defined by Committee and others (2011) for internal data. The dataset in question is at the level of individual loss events, it is fundamental as part of the study to know when they happened, and be able to identify the root causes of losses arising from which OpRisk loss events.

The LCDE is carried out drawing statistics directly from the trade generation and settlement system, which consists of a tractable set of documented trade detail extracted at the most granular level, i.e. on a trade-by-trade basis [as per number of events (frequencies) and associated losses (severities)], and then aggregated daily. The dataset is split into proportions and trained, validated and tested. The aforementioned LDCE, is an improved reflection of the risk factors by singling out the value-adding processes associated with individual losses, on a trade-by-trade level.

Current operational risk measurement modeling framework

Historical severity curves obtained from historical loss counts have been widely considered to be the most reliable models when used in OpRisk loss estimation. However they have not been successfull when used as measures capturing forward-looking aspects of the OpRisk loss prediction problem.

In this paper, we develop data intensive analysis techniques which yield a more realistic estimation for underlying risk factors, through linking risk factors to covariates based on internal control vulnerabilities (ICV's). ICV's are selected as measures of trading risk exposure, business environment and internal control factors (BEICF's) i.e., trade characteristics and causal factors. For each loss event, information such as unique trade identifier, trader identification, loss event capture personnel, trade status and instrument type, loss event description, loss amount, market variables, trading desk and business line, beginning and ending date and time of the event, and settlement time are given.

AMA's allow banks to use their internally generated risk estimates Under

Basel II; a first attempt internal measurement approach (IMA) capital charge calculation for OpRisk (i.e.) is similar to the Basel II model for credit risk, where a loss event is a default in the credit risk jargon. There are generally seven event type categories (Risk, 2001) and eight business lines. Potential losses are decomposed into several ($7 \times 8 = 56$) sub-risks using event types and business line combinations: e.g., execution, delivery & process management is one such category defined the risk that operational losses/problems would take place in the banks transactions, given as:

$$
\begin{aligned}
\mathcal{C}^{IMA}_{OpRisk} &= \sum_{i=1}^{8}\sum_{k=1}^{7}\gamma_{ik}\epsilon_{ik} \qquad (2.1)\\
where \quad \epsilon_{ik} &: \quad \text{expected loss for business line } i\text{, risk type } k\\
\gamma_{ik} &: \quad \text{scaling factor}
\end{aligned}
$$

The business line/ event type (BL/ET) matrix

The 3-dimensional diagram, Figure **??** depicts the formation of the BL/ET matrix: Duration (time $T + \tau$) is represented along the depth ordinate.

Loss Distribution Approach (LDA)

The Loss Distribution Approach (LDA) is an AMA method whose main objective is to provide realistic estimates to calculate VaR for OpRisk RC in the banking sector and it's business units based on loss distributions that accurately reflect the frequency and severity loss distributions of the underlying data. Having calculated separately the frequency and severity distributions, we need to combine them into one aggregate loss distribution that allows us to produce a value for the OpRisk VaR.

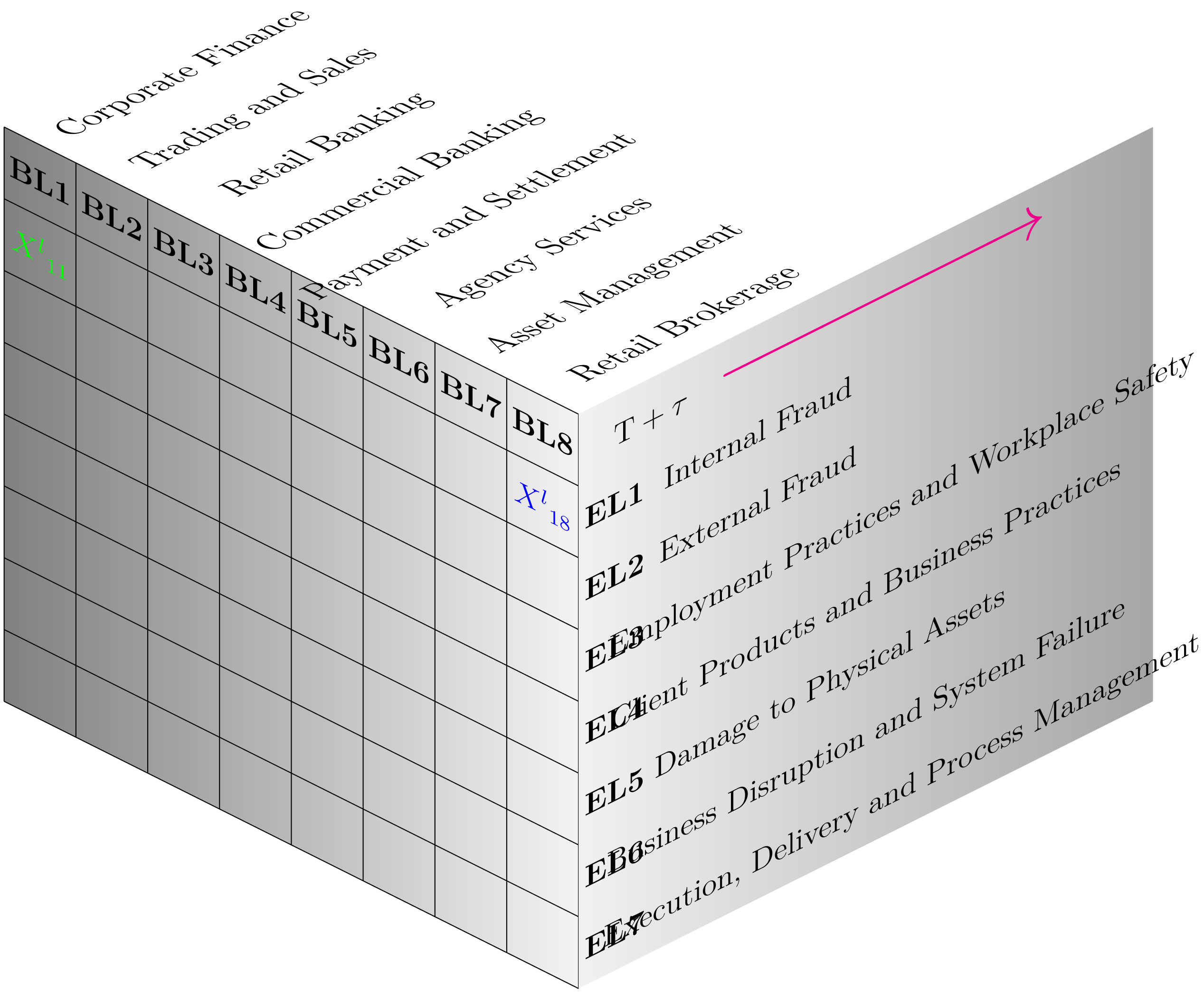


Figure 2.1: The 3-Dimensional grid of the BL/ET matrix for 7 event types and 8 business lines

We begin by defining some concepts:

- In line with Basel II, and according to @frachot2001loss, we consider a matrix consisting of business lines BL and (operational) event types ET. The bank estimates, for each business line/event type (BL/ET) cell, the probability functions of the single event impact and the event frequency for the next three months. More precisely, in each cell of the BL/ET matrix separate distributions for loss frequency and severity are modeled and aggregated to a loss distribution at the group level. The aggregated operational losses can be seen as a sum S of a random number N of individual operational losses $(X_1, \ldots, X_N)$. This sum can be represented by:

$$S = X_1, \ldots, X_N, \quad N = 1, 2, \ldots \tag{2.2}$$

- Three month daily statistics are taken of the time series of internal processing errors (frequency data) and their associated severities and used in each cell of the BL/ET matrix. Frequency refers to the number of events that occur within the specified time period (daily buckets) T and $T + \tau$ and severity refers to the P&L impact resulting from the frequency of events. The time (1 day bucket) period is chosen in order to ensure that the number of data points is sufficient for statistical analysis.

Computing the frequency distribution

- Let $\mathbf{N}_{ij}$ be variable in random selection, representing **the number of times of process risk event failures** between times T & $T + \tau$. Suppose subscript i refers to the BL which ranges from $1, \ldots, k$ and subscript j to ET ($j = 1$ for process risk). We have taken a random sample implying that the observations N_{ij}, where $i, j = (1, 1)\,, \ \ldots, (k, 1)$ are independent and identically distributed (i.i.d).

- The random variable N_{i1}[2] has distribution function[3] The random variable has distribution function (d.f.) $\mathbf{P}_{i1}(n/\theta_0)$, where θ_0 is an unknown parameter of the estimated distribution. The unknown parameter θ_0 may be a scalar or a vector quantity $\boldsymbol{\theta_0}$, for example, The Poisson distribution depends on one parameter called λ whereas the univariate normal distribution depends on two parameters, μ and σ^2, the mean and variance. These parameters are to be estimated in some way. We use the Maximum Likelihood Estimate (m.l.e) which is that value of θ that makes the observed data "most probable"or "most likely".
- The d.f. $\mathbf{P}_{i1}(n/\theta_0)$, is the probability that N_{i1} takes a value less than or equal to n, where n is a small sample from the entire population of observed frequencies, i.e.

$$\mathbf{P}_{ij}(n) = Pr\,(N_{ij} \leq n) \quad i, j = (1, 1), \ldots, (k, 1) \tag{2.3}$$

- The probability density function (p.d.f) : A density function is a non–negative function $p(n)$ whose integral, extended over the entire x axis, is equal to 1 for a given continuous random variable X. i.e. it is the area under the probability density curve, of the discrete random variable N_{i1} takes discrete values of n with finite probabilities. In the discrete case the term for p.d.f. is the probability function (p.f.) also called the probability mass function, i.e. N_{i1} is given by the probability that the variable takes the value n, i.e.

$$p_{ij}(n) = Pr\,(N_{ij} = n)\,, \quad i, j = (1, 1), \ldots, (k, 1) \tag{2.4}$$

- The r.h.s of equation (2.3) is the summation of the r.h.s of equation (5.1), we derive a relation for the **loss frequency distribution** in terms of the (p.f):

[2] N_{ij} where subscript $j = 1$ since we are only dealing with 1 event type i.e. process risk

[3] The term distribution function is monotonic increasing function of n which tends to 0 as $n \longrightarrow -\infty$, and to 1 as $n \longrightarrow \infty$

$$\mathbf{P}_{ij}(n) = \sum_{k=1}^{n_k} p_{ij}(n) \quad i, j = (1, 1), \ldots, (k, 1) \tag{2.5}$$

Computing the severity distribution

- Suppose X_{ij} is a random variable representing **the amount of one loss event** in a cell of the BL/ET matrix. Define next period's loss in each cell (i, j), where i is the number of business line cells, $L^{T+1}{}_{i,j}$: Operational loss for loss type $j = 1$ (process risk). One models the amount of the total operational loss of type j at a given time T & $T + 1$, over the future (say 3 months), as:

$$L^{T+1} = \sum_{i=1}^{k} L_{i1}^{T+1} = \sum_{i=1}^{2} \sum_{l=1}^{N_{i1}{}^{T+1}} X^{l}{}_{i1} \quad l = 1, 2, \ldots, N_{i1} \tag{2.6}$$

- Let $N_1, N_2, ..., N_m$ (where m in the number of combinations in the BL/ET matrix) be random variables that represent the loss frequencies. It is usually assumed that the random variables X_{i1} are independently distributed and independent of the number of events N_m. A fixed number of a particular loss type would be denoted by $X^1{}_{i1}$, i.e the random variable $X^l{}_{i1}$, represents random samples of the severity distribution [@aue2006lda].

  The **loss severity distribution** is denoted by $\mathbf{F}_{i1}$. Since loss severity variate X is continuous (i.e. can take on any real value), we define a level of precision h such that the probability of X being within $\pm h$ of a given number x tends to zero. The loss severity, X_{i1} has a (d.f.) $\mathbf{F}_{i1}(x/\theta_1)$, where θ_1 is an unknown parameter and x is a small sample from the entire population of loss severity.
- We define probability density in the continuous case as follows:

$$\begin{aligned} f_X(x) &= \lim_{h\to 0} \frac{Pr[x < X \leq x+h]}{h} \\ &= \lim_{h\to 0} \frac{F_X(x+h) - F_X(x)}{h} \\ &= \frac{dF_X(x)}{dx} \end{aligned} \tag{2.7}$$

operate with $\int\ dx$ on both sides of 2.7

$$\mathbf{F}_{X_{ij}}(x) = \int_{k=1}^{\infty} f_{X_{ij}}(x)dx \quad i,j = (1,1),\ldots,(k,1) \tag{2.8}$$

where $f_{X_{ij}}(x)$ is the probability density function (p.d.f.). Once again, the subscript X identifies the random variable for severity (P&L impact) of one loss event while the argument x is an arbitrary sample of the severity events.

Formal Results

Having calculated both the frequency and severity process we need now to combine them in one aggregate loss distribution that allows us to predict an amount for the operational losses to a degree of confidence. There is no simple way of aggregating the frequency and severity distribution. Numerical approximation techniques (computer algorithms) successfully bridge the divide between theory and implementation for the problems of mathematical analysis.

The aggregated losses at time t are given by $\vartheta(t) = \sum_{n=1}^{N(t)} X_n$ (where X represents individual operational losses). Frequency and severity distributions are estimated, e.g., the poisson distribution is a representation of a discrete variable commonly used to model operational event frequency (counts), and a selection from continuous distributions which can be linear (e.g. gamma distribution) or non-linear (e.g. lognormal distribution) for operational loss severity amounts. The compound loss distribution $\mathbf{G}(t)$ can now be derived. Taking the aggregated losses we obtain:

$$\mathbf{G}_{\vartheta(t)}(x) = Pr[\vartheta(t) \leq x] = Pr\left(\sum_{n=1}^{N(t)} X_n \leq x\right) \tag{2.9}$$

For most choices of $N(t)$ and X_n, the derivation of an explicit formula for $\mathbf{G}_{\vartheta(t)}(x)$ is, in most cases impossible. $\mathbf{G}(t)$ can only be obtained numerically using the Monte Carlo method, Panjer's recursive approach, and the inverse of the characteristic function [Frachot, Georges, and Roncalli (2001); Aue and Kalkbrener (2006); Panjer (2006); & others].

- We now introduce the aggregate loss variable at time t given by $\vartheta(t)$. This new variable represents **the loss for business line** i **and event type** j. The aggregate loss is defined by $\vartheta(t) = \sum_{n=1}^{N(t)} X_n$ (where X represents individual operational losses). Once frequency and severity distributions are estimated, the compound loss distribution $\mathbf{G}(t)$ can be derived. Taking the aggregated losses we obtain:

$$\mathbf{G}_{\vartheta(t)}(x) = Pr[\vartheta(t) \leq x] = Pr\left(\sum_{n=1}^{N(t)} X_n \leq x\right) \tag{2.10}$$

- The derivation of an explicit formula for $\mathbf{G}_{\vartheta(t)}(x)$ is, in most cases impossible. Again we implicitly assume that the processes $\{N(t)\}$ and $\{X_n\}$ are independent and identically distributed (i.i.d). Deriving the analytical expression for $\mathbf{G}_{\vartheta(t)}(x)$, we see a fundamental relation corroborated by @frachot2001loss, @cruz2002modeling, @embrechts2013modelling, & others:

$$\mathbf{G}_{\vartheta(t)}(x) = \left\{ \begin{array}{cc} \sum_{n,k=0,1}^{\infty} p_k(n)\mathbf{F}_X^{k\star}(x) & x > 0 \\ p_k(0) & x = 0 \end{array} \right\} \tag{2.11}$$

where $\star$ is the *convolution* operator on d.f.'s, $\mathbf{F}^{k\star}$ is the k-fold convolution of $\mathbf{F}$ with itself. The convolution of two functions $f(x)$ and $g(x)$ is the function

$$\int_0^x f(t)g(x-t)dt \tag{2.12}$$

, i.e. $\mathbf{F}_X^{k\star}(x) = Pr(X_1 + \ldots + X_k \leq x)$, the d.f. of the sum of k independent random variables with the same distribution as X.

- The aggregate loss distribution $\mathbf{G}_{\vartheta(t)}(x)$ cannot be represented in analytic form, hence approximations, expansions, recursions of numerical algorithms are proposed to overcome this problem. For purposes of our study, an approximation method will do. One such method consists of taking a set $\langle\vartheta_1, \ldots, \vartheta_s\rangle$, otherwise known as the ideal generated by elements $\vartheta_1, \ldots, \vartheta_s$ which are s simulated values of the random variable ϑ_{ij} for $s = 1, \ldots, S$ [@fraleigh2003first].

  This method is popularly known as Monte Carlo simulation coined by physicists in the 1940's, it derives its name and afore–mentioned popularity to its similarities to games of chance. The way it works in layman's terms is; in place of simulating scenario's based on a base case, any possible scenario through the use of a probability distribution (not just a fixed value) is used to simulate a model many times. In the LDA separate distributions of frequency and severity are derived from loss data then combined by Monte Carlo simulation.

Dependence Effects (Copulae)

The standard assumption in the LDA is that frequency and severity distributions in a cell are independent and the severity samples are i.i.d. According to Basel II, dependence effects in OpRisk are not considered. Economic capital allocation however, could benefit if it were determined in a way that recognises the risk-reducing impact of correlation effects between the risks of the BL/ET combinations. Concluding remarks from a study by Urbina and Guillén (2014) allude that failure to account for correlation may lead to risk management practices that are unfair, as evidenced in an example using data from the banking sector.

One of the main issues we are confronted with in OpRisk measurement is the aggregation of individual risks (in each BL/ET element). A powerful concept to aggregate the risks – the *copula* function – has been introduced in finance by Embrechts, McNeil, and Straumann (2002). Copulas have been used extensively in finance theory lately and are sometimes held accountable for recent global financial failures, e.g. the global credit crunch of 2008 - 2009. They are nevertheless still applicable and in use for OpRisk as operational risk models follow a different stochastic process to other areas of risk, e.g. operational VaR is subject to more jumps than market VaR and is thought to be discrete whereby market VaR is continuous.

Copulas are functions which conveniently incorporate correlation into a function that combines each of the frequency (marginal) distributions to produce a single bivariate cumulative distribution function. Our model is used to determine the aggregate (bivariate) distribution of a number of correlated random variables through the use a Clayton copula. Dependence matters due to the effect of the addition of risk measures over different risk classes (cells in the BL/ET matrix).

More precisely, the frequency distributions of the individual cells of the BL/ET matrix are correlated through a Clayton copula in order to replicate observed correlations in the observed data. Let m be the number of cells, $\mathbf{G_1}, \mathbf{G_2}, ..., \mathbf{G_m}$ the distribution functions of the frequency distributions in the individual cells and $\mathbf{C}$ the so–called copula. Abe Sklar proved in 1959 through his theorem (Sklar's Theorem) that for any joint distribution $\mathbf{G}$ the copula $\mathbf{C}$ is unique. $\mathbf{C}$ is a distribution function on $[0,1]^m$ with uniform marginals. We refer to a recent article by Chavez-Demoulin, Embrechts, and Nešlehová (2006) for further information: It is sufficient to note that $\mathbf{C}$ is unique if the marginal distributions are continuous.

$$\mathbf{G}(x_1, \ldots, x_m) = \mathbf{C}\left(\mathbf{G_1}(x_1), \ldots, \mathbf{G_m}(x_m)\right) \tag{2.13}$$

Conversely, for any copula $\mathbf{C}$ and any distribution functions $\mathbf{G_1}, \mathbf{G_2}, ..., \mathbf{G_m}$, the functions $\mathbf{C}\left(\mathbf{G_1}(x_1), \ldots, \mathbf{G_m}(x_m)\right)$ is a joint distribution function with marginals $\mathbf{G_1}(x_1), \ldots, \mathbf{G_m}(x_m)$. Moreover, combining given marginals with a chosen copula through Equation 2.13 always yields a multivariate distribution with those marginals. The copula function has then a great influence on the aggregation of risk.

LDA model shortcomings

After most complex banks adopted the LDA for accounting for RC, significant biases and delimitations in loss data remain when trying to attribute capital requirements to OpRisk losses (Frachot et al., 2001). OpRisk is related to the internal processes of the FI, hence the quality and quantity of internal data (optimally combined with external data) are of greater concern as the available data could be rare and/or of poor quality. Such expositions are unsatisfactory if OpRisk, as Cruz (2002) professes, represents the next frontier in reducing the riskiness associated with earnings.

Opdyke (2014) advanced studies intending on eliminating bias apparently due to heavy tailed distributions to further provide insight on new techniques to deal with the issues that arise in LDA modeling, keeping practitioners and academics at breadth with latest research in OpRisk VaR theory. Recent work in LDA modeling has been found wanting (Badescu, Lan, Lin, and Tang, 2015), due to the very complex characteristics of data sets in OpRisk VaR modeling, and even when studies used quality data and adequate historical data points, as pointed out in a recent paper by Hoohlo (2015), there is a qualitative aspect in OpRisk modeling that is often ignored, but whose validity should not be overlooked.

Opdyke (2014), Agostini et al. (2010), Jongh, De Wet, Raubenheimer, and Venter (2015), Galloppo and Previati (2014), and others explicate how greater accuracy, precision and robustness uphold a valid and reliable estimate for OpRisk capital as defined by Basel II/III. Transforming this basic knowledge into "risk culture" or firm-wide knowledge for the effective management of OpRisk, serves as a starting point for a control function providing attribution and accounting support within a framework, methodology and theory for understanding OpRisk measurement. FI's are beginning to implement sophisticated risk management systems similar to those for market and credit risk, linking theories which govern how these risk types are controlled to theories that govern financial losses resulting from OpRisk events.

Jongh et al. (2015) and Galloppo and Previati (2014) sought to address the shortcomings of Frachot et al. (2001) by finding possible ways to improve the problems of bias and data delimitation in operational risk management. They follow the recent literature in finding a statistical-based model for integrating internal data and external data as well as scenario assessments in on endeavor to improve on accuracy of the capital estimate.

A new class of models capturing forward-looking aspects

Agostini et al. (2010) also argued that banks should adopt an integrated model by combining a forward-looking component (scenario analysis) to the historical operational VaR, further adding to the literature through their integration model which is based on the idea of estimating the parameters of the historical and subjective distributions and then combining them by using the advanced CT.

The idea at the basis of CT is that a better estimation of the OpRisk measure can be obtained by combining the two sources of information: The historical loss data and expert's judgements, advocating for the combined use of both experiences. Agostini et al. (2010) seek to explain through a weight called the credibility, the amount of credence given to two components (historical and subjective) determined by statistical uncertainty of information sources, as opposed to a weighted average approach chosen on the basis of qualitative judgements.

Thus generating a more predictable and forward looking capital estimate. He deemed the integration method as advantageous as it is self contained and independent of any arbitrary choice in the weight of the historical or subjective components of the model.

Applicability of EBOR methodology for capturing forward-looking aspects of ORM

Einemann et al. (2018), in a theoretical paper, construct a mathematical framework for an EBOR model to quantify OpRisk for a portfolio of pending litigations. Their work unearths an invaluable contribution to the literature, discussing a strategy on how to integrate EBOR and LDA models by building hybrid frameworks which facilitate the migration of OpRisk types from a classical to an exposure-based treatment through a quantitative framework, capturing forward looking aspects of BEICF's (Einemann et al., 2018).

The fundamental premise of the tricky nature behind ORMF, is to provide an exposure-based treatment of OpRisk losses which caters to modeling capital estimates for forward-looking aspects of ORM due to the lag in the loss data. By the very nature of OpRisk, there is usually a significant lag between the moment the OpRisk event is conceived to the moment the event is observed and accounted.i.e., there is a gap in time between the moment the risk is conceived and the realised losses. This timing paradox often results in questionable capital estimates, especially for those near misses, pending and realised losses that need to be captured in the model.

Definition of exposure

Exposure is residual risk, or the risk that remains after risk treatments have been applied. In the ORMF context, it is defined as:

Definition 2.0.0.1 *The* ***exposure*** *of risk type i, d_i is the time interval, expressed in units of time, from the initial moment when the event happened, until the occurrence of a risk correction.*

Definition of rate

The **rate**, R is defined as:

Definition 2.0.0.2 *the* ***rate*** *is the mean count per unit exposure*

i.e.,

$$\begin{aligned} R &= \frac{\mu}{\tau} \qquad \text{where} \qquad R = \text{rate}, \quad \tau = \text{exposure}, d_i \quad \text{and} \\ \mu &= \text{mean count over an exposure duration of} \quad T + \tau \end{aligned}$$

Intepretation

In turn, with reference to **??**, the fundamental premise behind the LDA is that each firm's OpRisk losses are a reflection of it's underlying Oprisk exposure. In particular, the assumption behind the use of the poisson model to estimate the frequency of losses, is that both the the intensity (or rate) of occurrence and the opportunity (or exposure) for counting are constant for all available observations.

The measure of exposure we need to use depends specifically on projecting the number of Oprisk event types (frequency of losses) and is different to the measure if the target variable were the severity of the losses. We need historical exposure for experience rating because we need to be able to compare the loss experience of different years on a like-for-like basis and to adjust it to current exposure levels(Parodi, 2014).

When observed counts all have the same exposure, modeling the mean count μ as a function of explanatory variables $x_1, \ldots, x_p$ is the same as modeling the rate R.

Benefits and Limitations

These approaches in 2, were found to have significant advantages over conventional LDA methods, proposing that an optimal mix of the two modeling elements could more accurately predict OpRisk VaR over traditional methods. Particularly Agostini et al. (2010), whose integration model represents a benchmark in OpRisk measurement by including a component in the AMA model that is not obtained by a direct average of historical and subjective VaR.

Instead, the basic idea of the integration methodology in 2 is to estimate the parameters of the frequency and severity distributions based on the historical losses and correct them; via a statistical theory, to include information coming from the scenario analysis. The method has the advantage of being completely self contained and independent of any arbitrary choice in the weight of the historical or subjective component of the model, made by the analyst. The components weights are derived in an objective and robust way, based on the statistical uncertainty of information sources, rather than through risk managers choices based on qualitative motivations.

However, they could not explain the prerequisite coherence between the historical and subjective distribution function needed in order for the model to work; particularly when a number of papers (Chau, 2014), propose using mixtures of (heavy tailed) distributions commonly used in the setting of OpRisk capital estimation (Opdyke, 2014).

In 2, their model (Einemann et al., 2018) is particularly well-suited to the specific risk type dealt with in their paper i.e., the portfolio of litigation events, due to better usage of existing information and more plausible model behavior over the litigation life cycle, but is bound to under-perform for many other OpRisk event types, since these EBOR models are typically designed to quantify specific aspects of OpRisk - litigation risk have rather concentrated risk profiles. However, EBOR models are important due to wide applicability beyond capital calculation and its potential to evolve into an important tool for auditing process and early detection of potential losses.

Gap in the Literature

There is cognitive pressure which seeks to remove information which we are largely unaware of, because they are undetectable to human senses that no one could ever see them. We seek to remove this pressure, effectively lowering uncertainty and allowing us to position ourselves to develop a defense against our cognitive biases. It is through patterns in that information that we are largely unaware of that predictions could arise; or that, OpRisk management incorporates rather than dismiss the many alternatives that were not imagined, the possibility of market inefficiencies or finding value in unusual places.

Conclusion

A substantial body of evidence suggests that loss aversion, the tendency to be more sensitive to losses than to gains plays an important role in determining how people evaluate risky gambles. In this paper we evidence that human choice behavoir can substantially deviate from neoclassical norms.

PT takes into account the loss avoidance agents and common attitudes toward risk or chance that cannot be captured by EUT; which is not testing for that inherent bias, so as to expect the probability of making the same operational error in future to be overcompensated for i.e., If an institution suffers from an OpRisk event and survives, it's highly unlikely to suffer the same loss in the future because they will over-provide for particular operational loss due to their natural risk aversion. This is a testable proposition which fits normal behavioral patterns and is consistent with risk averse behaviour.

CHAPTER 3
EXPOSURE-BASED OPERATIONAL RISK ANALYSIS

Introduction

The fundamental premise in the nature behind ORMFs, is to provide an exposure-based treatment of OpRisk losses which caters to modeling capital estimates for forward-looking aspects of ORM. This proves tricky due to the lag between the time the loss event occurs and the actual realised loss, i.e. by the very nature of OpRisk, there is a significant lag between the moment the OpRisk event is conceived to the moment the event is observed and accounted for. There is a gap in time between τ the moment the risk is conceived and the time $\tau + 1$ when impact of the loss is realised. This timing paradox often results in questionable capital estimates, especially for those near misses, pending and realised losses that need to be captured in the model

EBOR methodology for capturing forward-looking aspects of ORM

Einemann et al. (2018), in a theoretical paper, construct a mathematical framework for an EBOR model to quantify OpRisk for a portfolio of pending litigations. Their work unearths an invaluable contribution to the literature, discussing a strategy on how to integrate EBOR and LDA models by building hybrid frameworks which facilitate the migration of OpRisk types from a *classical* to an exposure-based treatment through a quantitative framework, capturing forward looking aspects of BEICF's (Einemann et al., 2018), a key source of the OpRisk data.

The measure of exposure we need to use depends specifically on projecting the number of OpRisk event types (frequency of losses) as the target variable in the model and is different to the measure if the target variable were the severity of the losses. As per definition 2, the lag represents exposure; we need historical exposure for experience rating because we need to be able to compare the loss experience of different years on a like-for-like basis and to adjust it to current exposure levels (Parodi, 2014).

With reference to the current section, Section 3, the fundamental premise behind the LDA is that each firm's OpRisk losses are a reflection of it's underlying Oprisk exposure. In particular, the assumption behind the use of the Poisson distribution in the model to estimate the frequency of losses, is that both the the intensity (or rate) of occurrence and the opportunity (or exposure) for counting are constant for all available observations.

Limitations of the EBOR model

In their model (Einemann et al., 2018), the definition of exposure, Definition 2, is particularly well-suited to the specific risk type dealt with in their paper i.e., the portfolio of litigation events, due to better usage of existing information and more plausible model behavior over the litigation life cycle. However, it is bound to under-perform for many other OpRisk event types since these EBOR models are typically designed to quantify specific aspects of OpRisk i.e., litigation risk have rather concentrated risk profiles. Furthermore, EBOR models are important due to wide applicability beyond capital calculation and its potential to evolve into an important tool for auditing process and early detection of potential losses.

Generalised Linear Models (GLM's)

Operational riskiness in FIs grows as trading transactions grow in complexity, i.e. the more complex and numerous trading activity builds the higher the rate at which new cases of OpRisk events occur. Therefore, it is likely that the rate of operational hazard may be increasing exponentially over time. The scientifically interesting question is whether the data provides any evidence that the increase in the underlying operational hazard generation is slowing.

The aforementioned postulate provides a plausible model to start investigating this question. The model assumes that if μ_i is the (rate) number of expected new OpRisk events per time interval $[\tau, \tau + 1]$ since the beginning of the data, then μ_i increases according to:

$$\mu_i = \gamma \text{exp}(\delta\tau_i)$$

where γ and δ are unknown parameters. Taking a log link turns the model into Generalised Linear Model (GLM) form so that:

$$\log(\mu_i) = \log(\gamma) + \delta\tau_i = \beta_0 + \tau_i\beta_1 \tag{3.1}$$

Where the LHS is the observed number of new cases over time τ and $\tau + 1$, and the RHS is a linear in the parameters $\beta_0 = \log(\gamma)$ and $\beta_1 = \delta$.

We define a binary variable LossIndicator, which takes on the value 1 for realised losses, and the value 0 for pending losses, and/or near misses. The choice of the poisson distribution shows the number of relevant Oprisk event's incidence in a specified time interval τ and $\tau + 1$. A quadratic term $(\beta_2 t_i^2)$ could be added to the model so that its residual values increase monotonically with time as with the fitted (modelled) values, which usefully approximates other situations which may influence the counts adapted to the poisson case. Amending the RHS of Equation 3.1 with the quadratic term so other situations other than the unrestricted spread of OpRisk hazards are represented yields:

$$\mu = d_i \text{exp}(\beta_0 + \beta_1 \tau_i + \beta_2 {\tau_i}^2) \tag{3.2}$$

GLM for count data

The LHS of a GLM formula is the model's random component i.e., observations of the number of OpRisk transactions over the trading transaction period in an FI's portfolio; given by the independent random variables $y_1, y_2, \ldots, y_n$, not i.i.d (Wood, 2017, Covrig et al. (2015)). Y takes a (exponential) family argument, depending on parameters $ln\lambda$, where λ which represents the average frequency of the OpRisk transactions. It is worth distinguishing between the response data y_i which is an observation of Y.

A poisson regression operational hazard model

The target variable (LossIndicator) which shows the number of relevant Oprisk event's incidence in a specified time interval τ and $\tau + 1$, justifying the choice of poisson distribution as a reasonable model for is count data. It's probability mass function (pdf) is:

$$\begin{aligned} Y \quad &\sim \quad \text{poisson}(\lambda), \quad f(y;\lambda) = \frac{\lambda^y}{y!} \text{e}\dot{\text{x}}\text{p} {-y} \\ \text{where} \quad & y \in \mathbb{N}, \text{and} \quad \lambda > 0 \end{aligned} \tag{3.3}$$

the expectation and variance $E[Y] = \text{VaR}[Y] = \lambda^1$, are both equal to parameter λ simultaneously. The RHS of Equation 3.3 is the model's systematic component, and it specifies the linear predictor. It builds on equation 3.2 with $p + 1$ parameters, $\beta = (\beta_0 \ldots, \beta_p)^t$, and p explanatory variables:

$$\nu = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \qquad \text{where} \quad i = 1, \ldots, n \tag{3.4}$$

If sample variables $Y_i \sim \text{Poisson}(\lambda_i)$, then $\mu = E[Y_i] = \lambda_i$; the link function between the random and systematic components, viz. a tranformation by the model by some function $g()$, which does not change features essential to to fitting, but rather a scaling in magnitude so that:

$$\begin{aligned} \nu_i &= g(\lambda_i) = \text{ln}\lambda_i, \qquad \text{that is} \\ \nu &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \end{aligned} \tag{3.5}$$

so the mean frequency or otherwise the rate R, will be predicted by the model...

$$\begin{aligned} \lambda_i &= d_i \text{exp}(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) \quad \text{or} \\ \lambda_i &= d_i \cdot e^{\beta_0} \cdot e^{\beta_1 x_{i1}} \cdot e^{\beta_2 x_{i2}} \ldots e^{\beta_p x_{ip}} \end{aligned} \tag{3.6}$$

Where d_i represents the risk exposure for transaction i. Taking logs on both sides of equation 3.6, the regression model for the estimation of loss frequency is:

$$\text{ln}\lambda_i = \text{ln}d_i + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \tag{3.7}$$

[1] If you were to guess an independent Y_i from a random sample, the best guess is given by this expression

Table 3.1: The generalized linear model link functions with their associated units of interpretation.

Link Function	Target variable Effect
Identity	Original Continuous Unit
Log	Original Continuous Unit
Logit	Risk
Probit	Risk
Poisson	Count
Gamma	Count
Negative Binomial	Count

where lnd_i is the natural log of risk exposure, called the "offset variable".

Interpretation

Table 3.1 presents the various units produced for the various GLM links.

The poisson distribution is restrictive when applied to approximate counts, due to the assumption made about it that the mean and variance of the number of events are equal. However, in models for count data where means are low so that the number of zeros and ones in the data is exessive are well adapted to the poisson case (Wood, 2017). These cases are characteristic of scenarios in OpRisk other than those modeling situations when the unchecked spreading of negligent behaviour may result in an operational hazard. For example, the negative binomial and/or quasipoisson regression models ascribe to data that exhibits *overdispersion*, wherein the variance is much larger than the mean for basic count data, therefore they have been eliminated in this paper.

Research Objective 1

To introduce the generalised additive model for location, scale and shape (GAMLSS) framework for OpRisk management, that captures exposures to forward-looking aspects in the OpRisk loss prediction problem, due to deep hierarchies in the features of covariates in the investment banking (IB) business environment, and internal control risk factors (BEICF) thereof.

Exploratory data analysis

The main source of the analysis dataset is primary data, a collection of internal OpRisk losses for the period between 1 January 2013 and 31st March 2013 at an investment bank in SA. The method of data generation and collection is at the level of the individual trade deal, wherein deal information is drawn directly from the trade generation and and settlement system (TGSS) and edit detail from attribution reports generated in middle office profit & loss (MOPL). The raw source consists of two separate datasets on a trade-by-trade basis of daily frequencies (number of events) and associated loss severities.

The raw frequency data consists of 58,953 observations of 15 variables, within the dataset there are 50,437 unique trades. The raw severity data consists of 6,766 observations of 20 variables; within the severity dataset there are 2,537 unique trades. The intersection between the frequency and severity datasets consists of 2,330 individual transactions which represent realised losses, pending and/or near misses. This dataset is comprised of 3-month risk correction detail, in the interval between 01 January 2013 and 31 March 2013.

Two new variables are derived from the data; a target variable (LossIndicator) is a binary variable whereupon, a 1 signifies a realised loss, and 0 for those pending losses, or near misses. The *exposure* variable is computed by deducting the time between the trade amendment (UpdateTime) and the time when the trade was booked (TradeTime). It is a measure that is meant to be rougly proportional to the risk of the transaction or a group of transactions. The idea is that if the exposure

Table 3.2: The contents of the traded transactions of the associated risk correction events.

		Storage
Covariate	Levels	Type
Trade		numeric
UpdateTime		numeric
UpdatedDay		numeric
UpdatedTime		numeric
TradeTime		numeric
TradedDay		numeric
TradedTime		numeric
Desk	10	categorical
CapturedBy	5	categorical
TradeStatus	4	categorical
TraderId	7	categorical
Instrument	23	categorical
Reason	19	categorical
Loss		numeric
EventTypeCategoryLevel	5	categorical
BusinessLineLevel	8	categorical
LossIndicator	2	binary
exposure		numeric

(e.g. the duration of a trade, the number of allocation(trade splits), etc.) doubles whilst everything else (e.g. the rate, nominal of the splits, and others) remains the same, then the risk also doubles.

In R, the GLM function works with two types of covariates/explanatory variables: numeric (continuous) and categorical (factor) variables as depicted in table 3.2. Multi-level categorical variables are recoded by building dummy variables corresponding to each level. This is achieved through an implemented algorithm in R, through a transformation as recommended for the estimation of the GLM, particularly in the estimation of the poisson regression model for count data.

The model revolves around the fact that for each categorical variable (covariate), previously transformed into a dummy variable, one must specify a reference category from which the corresponding observations under the same covariate are estimated and assigned a weight against in the model (Covrig et al., 2015). By default in the GLM, the first level of the categorical variable is taken as the reference level. As best practice, De Jong, Heller, and others (2008), Frees and Sun (2010), Denuit, Maréchal, Pitrebois, and Walhin (2007), Cameron and Trivedi (2013) and others recommend that for each categorical variable one should specify the modal class as the reference level; as this variable corresponds to the level with the highes order of predictability, excluding the dummy variable corrresponding to (weight coefficient = 0) the biggest absolute frequency.

Description of the dataset

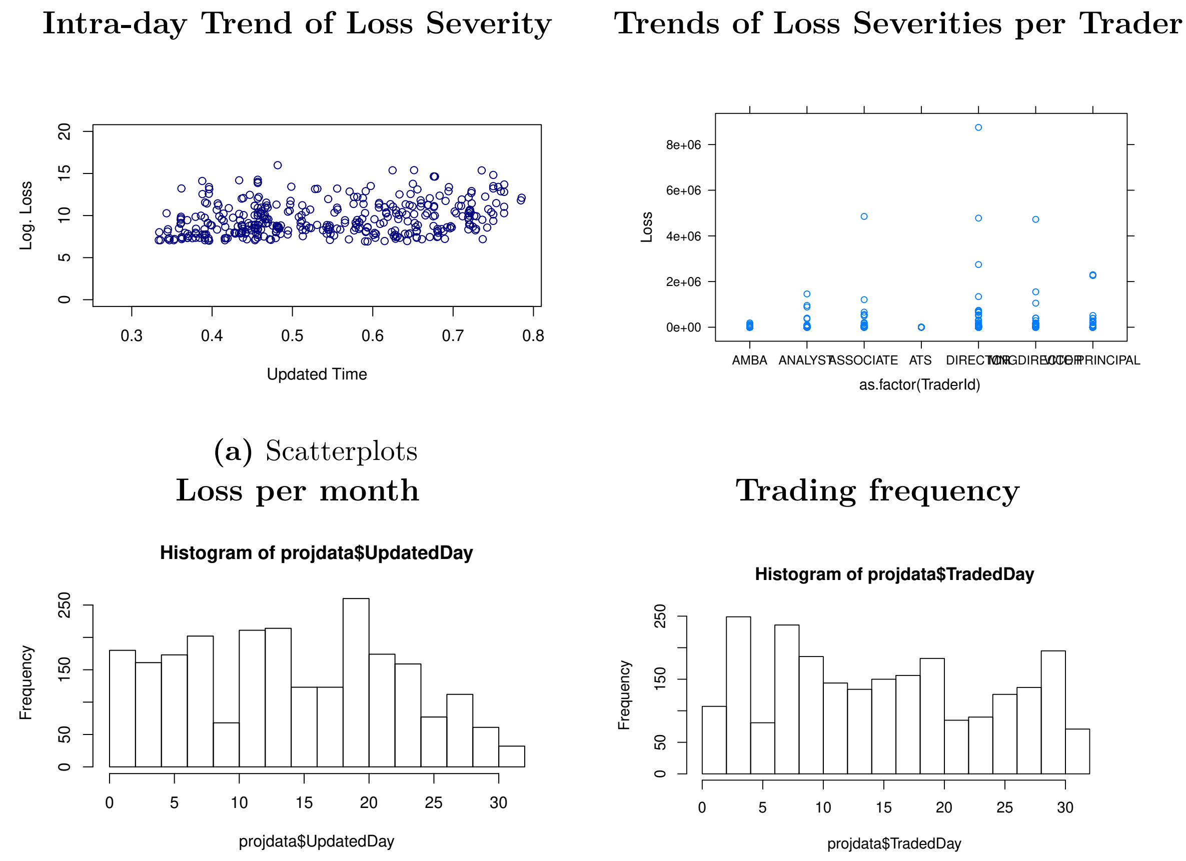
In this section, section 3, the dataset called *OpRiskDataSet_exposure*, provides data on the increase in the numbers of operational events over a three month period, beginning 01 January 2013 to end of 20 March 2013. For each transaction, there is information about: trading risk exposure, trading characteristics, causal factor characteristics and their cost.

Characteristics of exposure

The exposure of risk of type i, d_i shows the daily duration, from when the trade was booked to the moment the operational risk event was observed and ended. This measure is defined this way when specifically applied to projecting the number of loss events (frequencies) and can be plotted as follows depicted in graphs depicted in Figure 3.2.

The variable follows a logistic trend on $[0, 1]$, implying an FIs operational risk portfolio rises like a sigmoid function throughout the period of observation, typically starting from 0, which then observes a plateau in growth. The average exposure is 389.99 or about 1 year.

Grid plots 3.2 portray the logistic function, together with a simple compari-



(a) Scatterplots

(b) Histograms

Figure 3.1: (a) Scatterplots of intra-day trend analysis for logs of severities of operational events and trends incident activity for identifying the role of the trader originating the incidents. (b) As for (a) but in the form of histograms showing the frequency distrbution of the number daily operational indicents and the number of trades over a monthly period.

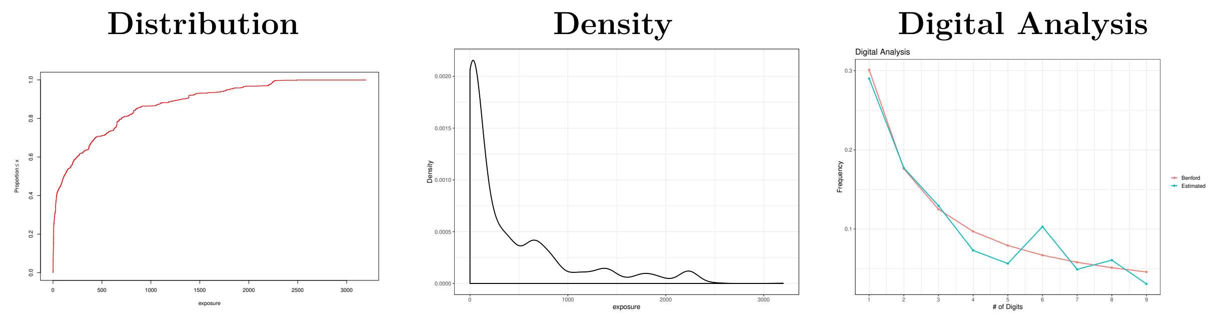


Figure 3.2: A simple comparison of the Sigmoidal like features of the fat-tailed, right skewed distribution for exposure, and first-digit frequency distribution from the exposure data with the expected distribution according to Benford's Law

son of first-digit frequency distribution analysis, according to Benford's Law, with exposure data distribution. The close fitting nature implies the data are uniformly distributed across several orders of magnitude, especially within the 1 year period.

Characteristics of the covariates

The characteristics of the operational risk portfolio are given by the following covariates: *UpdatedDay*, *UpdatedTime* - the day of the month and time of day the OpRisk incident occurs respectively; *TradedDay*, *TradedTime* - the day in the month and time of day the deal was originated respectively; The *LossIndicator* as indicated before is a binary variable consisting of two values: A 0, which indicates pending or near misses, and 1, if the incident results in a realised loss, meaning that there is significant p&L impact due to the OpRisk incident.

the *Desk* is the location in the portfolio tree the incident originated, it is a factor variable conisting of 10 categories; *CapturedBy*, the designated analyst who actions the incident, a factor variable consisting of 5 categories; *TraderId*, the trader who originates the deal, a factor variable with 7 categories; *TradeStatus*, the live status of the deal, a factor variable with 4 categories; *Instrument*, the type of deal, a factor variable with 23 categories; *Reason*, a description of the cause of the OpRisk incident, a factor variable with 19 levels; *EventTypeCategoryLevel*, 7 OpRisk event types as per Risk (2001), a factor variable with 5 categories; *BusinessLineLevel*, 8 OpRisk business lines as per Risk (2001), a factor variable with 8 categories.

The factor variables were transformed into dummy variables using the following commands:

```
# Remap factor variables and transform into numeric variables.
crs$dataset[["TNM_Desk"]] <- as.numeric(crs$dataset[["Desk"]])
crs$dataset[["TNM_CapturedBy"]] <- as.numeric(crs$dataset
                                              [["CapturedBy"]])
```

```
crs$dataset[["TNM_TraderId"]] <- as.numeric(crs$dataset[["TraderId"]])
crs$dataset[["TNM_Instrument"]] <- as.numeric(crs$dataset
                                               [["Instrument"]])
crs$dataset[["TNM_Reason"]] <- as.numeric(crs$dataset[["Reason"]])
crs$dataset[["TNM_EventTypeCategoryLevel1"]] <- as.numeric(crs$dataset
                                         [["EventTypeCategoryLevel1"]])
crs$dataset[["TNM_BusinessLineLevel1"]] <- as.numeric(crs$dataset
                                              [["BusinessLineLevel1"]])
```

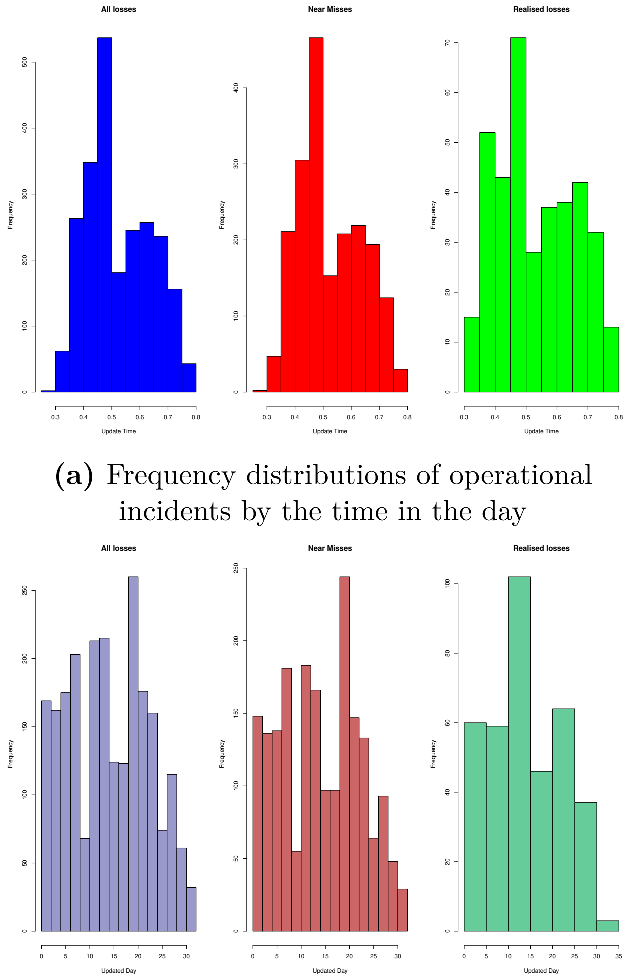
The continuous numerical variable *Loss*, shows the financial impact (severity) of the OpRisk incident in Rands. For the most part (i.e. 96.1% of the time) OpRisk incidents result in pending losses and/or near misses, most realised losses (2.3%) lie within the $[\mathbf{R}200, 00, \mathbf{R}300, 000]$ range. In the current portfolio there are also five p&L impacts higher than **R**2.5 **million**.

Characteristics of daily operational activity

The distribution of daily losses and/or pending/near misses by operational activities are represented in 3.3. Figure 3.3a shows that most operational events occur in times leading up to midday (i.e. 10:50AM to 11:50AM), the observed median is 11:39AM, and of these potential loss events, most realised losses occur closest to mid-day. The frequencies of the loss incidents in the analysed portfolio sharply decreases during the following period, i.e. from 12:10PM to 13:10PM, during which the least realised losses occur.

Figure 3.3b shows that operational activity increases in intensity in the days leading up to the middle of the month, i.e. 10^{th} - 15^{th}; the observed mean is 14.49 days, and of these potential loss events, realised losses especially impact on the portfolio during these days.

Similarly, the influence of trading desk's on the frequency of operational events can be analysed on the basis of the portfolio's bidimensional distribution by variables *Desk* and *LossIndicator*, which shows the proportions realised losses vs



(a) Frequency distributions of operational incidents by the time in the day

(b) Frequency distributions of operational incidents by the day in the month

Figure 3.3: The frequency distributions of All the losses, the realised losses, and pending/near misses of operational incidents by the day in the month when the indidents' occurred

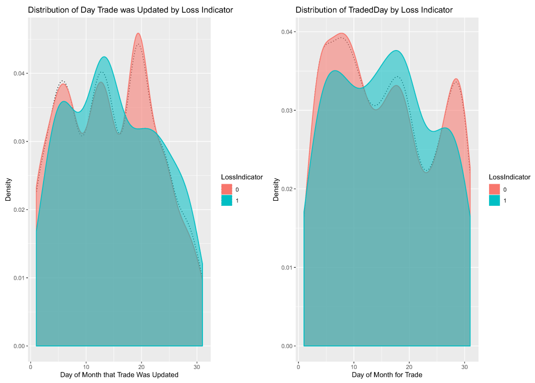


Figure 3.4: Density plots showing a comparison of realised vs pending losses and/near misses over a month for the day in the month the OpRisk incident was updated to the day in the month trades were traded/booked

pending and/or near misses for each particular desk. The bidimensional distribution of *Desk* and *LossIndicator* is presented in a contingency table, Table 3.3, in which it's considered useful to calculate proportions for each desk category.

Table 3.3: Occurence of realised losses: proportions on desk categories

	No. of transactions		
Desk	no Loss	Loss	Total
Africa	49	10	59
Bonds/Repos	113	31	144
Commodities	282	45	327
Derivatives	205	24	229
Equity	269	66	335
Management/Other	41	2	43
Money Market	169	52	221
Prime Services	220	62	282
Rates	336	53	389
Structured Notes	275	26	301

Thus, as illustratred in figure 3.5, from 23,5%; the highest proportion of realised losses per desk is the Money Market (MM) desk, the figures are decreasing, followed by Prime Services (22%); Bonds/Repos (21,5%); Equity (19,7%); Africa

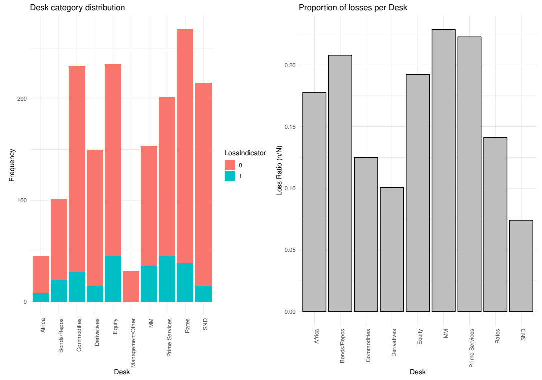


Figure 3.5: Histograms showing the proportions of realised losses vs all losses including pending and/or near misses by desk category

(16,9%); Commodities (13,8%); Rates (13,6%); Derivatives (10,5%); Structured Notes (SND) (8.6%), to the least proportion in the Management/Other, a category where only 4,7% of operations activities were realised as losses.

This behaviour can be extended beyond the trading desk, as represented in Figure 3.6, a mosaic plot grid presenting the structure of the OpRisk portfolio by Instrument, TraderId, CapturedBy [2] and the operational losses.

One can notice that the width of the bars corresponding to the different categories, i.e. Instrument, TraderId, CapturedBy, is given by their proportion in the sample. In particular, for the category 'at least one realised loss', in the top right mosaic of Figure 3.6 portrays a increase in "riskiness" trending up from Associate to AMBA, Analyst, Vice Principal, Managing Director, Director, up to the risky ATS category, which are automated trading system generated trades.

Figure 3.6 bottom right mosaic plot for technical support personnel for the category 'at least one realised loss', portrays a downward trend, slowing in riskiness from Unauthorised User downward to Tech Support, Mid Office, Prod Controller down to the least risky Prod Accountant. This intepretation makes sense given

[2] i.e. the type of financial instrument, the trader who originated the incident on the deal, and the role of the technical support personnel who is involved in the query resolution.

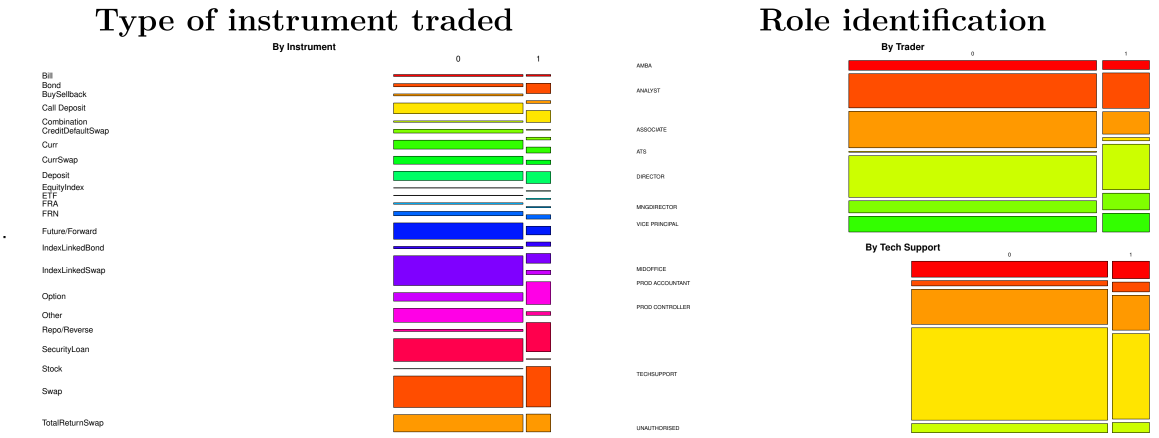


Figure 3.6: Mosaic grid plots for the bidimensional distribution by traded instrument, the trader originating the operational event, and by the technical support personnel involved in query resolution, against the dummy variable showing if a realised loss was reported.

unauthorised users are more likely to make impactful operational errors, technical support personnel would also be accountable for large impacts albiet for contrasting reasons, they are mandated to perform these deal adjustments which have unavoidable impacts associated with them, whereas the former group are unauthorised to perform adjustments therefore may lack the skill, or be criminally minded insiders acting on their own or in unison to enable their underhanded practices and intentions without raising any suspicion.

In another mosaic plot, Figure 3.7, the bidimensional distribution of transactions by trader and realised vs pending losses, conditional on the trade status is presented and analysed. Here, and in the contingency table, Table 3.5, we can clearly see the following trends: In BO-BO confirmed status - an increase in realised losses from the leftmost TraderID (i.e. AMBA) to right, and the opposite for transactions performed in BO Confirmed status (both with two exceptions). In particular, the biggest number of realised losses in both BO and BO-BO Confirmed statuses occur due to automated trading systems (ATS) who also give rise to the exceptions mentioned.

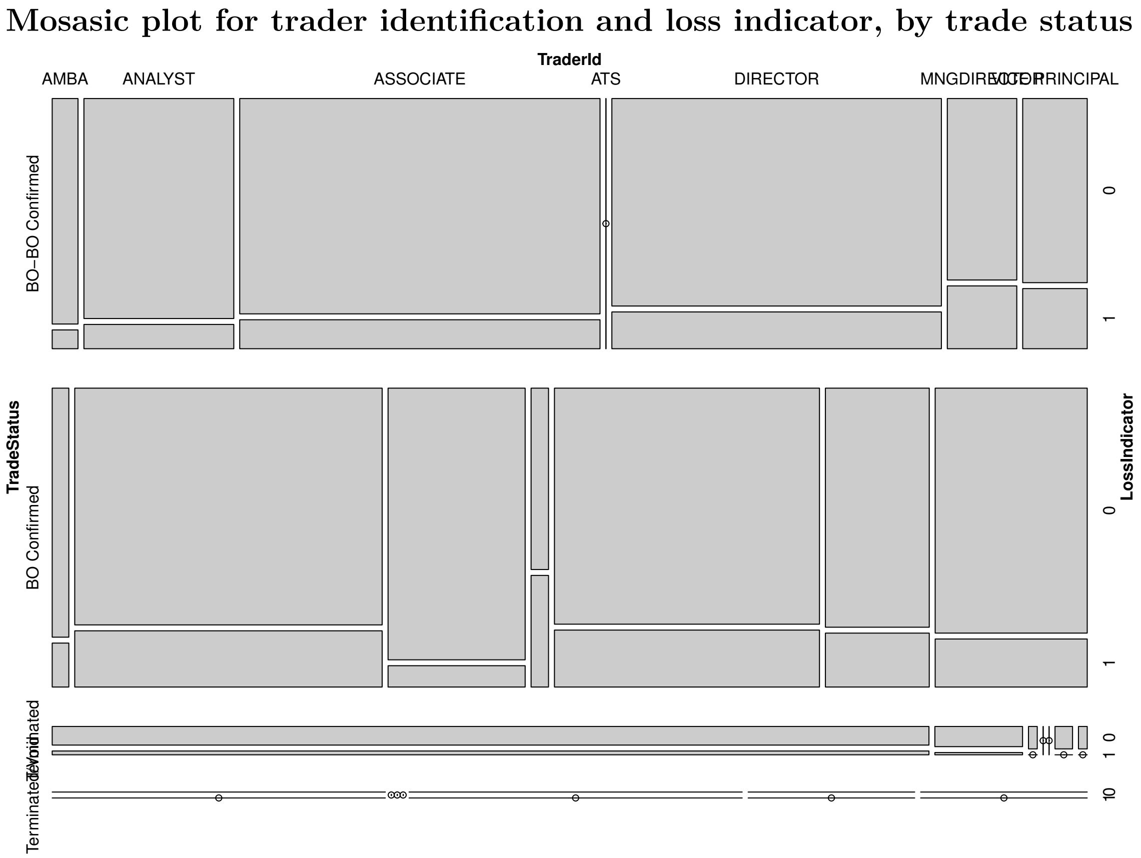

Mosaic plot for trader identification and loss indicator, by trade status

Figure 3.7: A mosaic plot representing the structure of the operational risk portfolio by trader identification (TraderId), the status ofthe trade (TradeStatus) and the number of realised losses vs pending or near misses

Table 3.5 and Figure 3.7 are obtained with the following commands:

```r
library(vcd)
STD <- structable(~TradeStatus + TraderId + LossIndicator
                                        , data = projdata)
MS01 <- mosaic(STD, condvars = 'TradeStatus', col=rainbow(20),
                  split_horizontal = c(TRUE, FALSE, TRUE))
```

Table 3.4 presents the most frequent category in the operational risk dataset for each possible covariate.

Crosstab of trader identification and loss indicator, by trade status

		Trader Identification						
TradeStatus	Loss Indicator	Amba	Analyst	Associate	ATS	Director	Mng Director	Vice Principal
BO-BO Confirmed	0	24	136	320	0	282	52	49
	1	2	15	43	0	50	18	16
BO Confirmed	0	17	299	153	13	257	102	153
	1	3	71	12	8	62	23	30
Terminated	0	83	9	1	0	0	2	1
	1	17	1	0	0	0	0	0
Terminated/Void	0	2	0	0	0	2	1	1
	1	0	0	0	0	0	0	0

Table 3.4: A contingency table showing the bidimensional distribution of transactions by trader identification vs realised and/or pending losses, conditional on the trade status

Modal classes for the categorical variables

Variable	Modal class or category	Name of modal class
Desk	Rates	DeskRates
CapturedBy	TECHSUPPORT	CapturedBy_TECHSUPPORT
TradeStatus	BO confirmed	TradeStatus_BO confirmed
TraderId	DIRECTOR	TraderId_DIRECTOR
Instrument	Swap	Instrument_Swap
Reason	Trade enrichment for system flow	Reason_Trade enrichment for system flow
EventTypeCategoryLevel	EL7	EventTypeCategoryLevel_EL7
BusinessLineLevel	BL2	BusinessLineLevel_BL2

Table 3.5: A contingency table showing the bidimensional distribution of transactions by trader identification vs realised and/or pending losses, conditional on the trade status

The estimation of some poisson regression generalised linear models (GLM's)

Section 3 introduced a GLM for the start of the expected number of operational events in the early stages. We aim to estimate the mean OpRisk frequency through a poisson classification model given by equation 3.3 using the glm function. The mean daily loss frequency in the risk correction statistics is estimated through the poisson regression model. Let us consider a model where the *LossIndicator* is the target variable: The following fits the model (the log link is canonical for the poisson distribution, and hence the R default) and checks it.

In calling the GLM we specify the target variable *LossIndicator*; the explanatory variables are composed of of numeric, continuous and categorical variables. Where the variable in the argument of a GLM is categorical , one chose to specify the modal class as the reference level. A user defined function "getmode" has been created; it selects the modal observation in each factor, and the dataset is reordered using the *relevel* function in R. These specifications were achieved in the following code chunk:

```
# Create function "getmode" which finds the modal class in
# the categorical variables
getmode <- function(x){
  u <- unique(x)
  as.integer(u[which.max(tabulate(match(x,u)))])
}
# Reorder the categorical variables so that the modal class
# is specified as the reference level
for (i in 5:(ncol(d1) - 3)){
     d1[[i]] <- relevel(d1[[i]], getmode(d1[[i]]))
}
```

Other GLM arguments are: The afore-mentioned link function poisson(link="log"); a data frame containing the OpRisk dataset, data=d1; and the r offset=log(exposure), i.e. the variable representing a component known apriori, coefficient= 1, introduced in the linear predictor (Covrig et al., 2015).

Firstly, consider a GLM in which is introduced two explanatory variables, one numerical variable, *UpdatedTime*, and another categorical variable *Desk*. This will be our global model. We will use *LossesIndicator* as the target variable, while these two unique variables will be explanatory variables:

```
freqfit1 <- glm(LossesIndicator ~ UpdatedTime + Desk, data=d1,
               family=poisson(link = 'log'), offset = log(exposure))
```

The output result of the estimation is presented below, where variables who were found to be significant predictors are indicated. The coefficients of the categorical variable *Desk* are reordered and weighted against the modal class: *DeskRates*. Interestingly the modal class does not show up in the results section (as the coefficient of the modal class = 0), given that the remaining classes are weighted against it.

```
## 
## Call:
## glm(formula = LossesIndicator ~ UpdatedTime + Desk, family = poisson(link = "lo
##     data = d1, offset = log(exposure))
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.8450  -0.5587  -0.2438  -0.0536   4.2780  
## 
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)             -9.2147     0.2855 -32.275  < 2e-16 ***
## UpdatedTime              1.7972     0.4749   3.784 0.000154 ***
## DeskAfrica               1.2515     0.3449   3.629 0.000285 ***
## DeskBonds/Repos          1.7758     0.2263   7.846 4.30e-15 ***
## DeskCommodities          0.8274     0.2027   4.082 4.47e-05 ***
## DeskDerivatives         -0.2071     0.2468  -0.839 0.401446    
## DeskEquity               1.3687     0.1849   7.403 1.33e-13 ***
## DeskManagement/Other    -1.2208     0.7204  -1.695 0.090135 .  
## DeskMM                   0.3910     0.1954   2.001 0.045431 *  
## DeskPrime Services     2.1217     0.1875  11.316  < 2e-16 ***
## DeskSND               -0.7055     0.2397  -2.943 0.003250 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2767.9  on 2329  degrees of freedom
## Residual deviance: 2461.7  on 2319  degrees of freedom
## AIC: 3225.7
##
## Number of Fisher Scoring iterations: 8
```

Using this bivariate model, the estimated quarterly OpRisk (LossIndicators) frequency of realised losses for each *Desk* category (excluding the insignificant ones) are:

* $0,002099618 = e^{-9.2147} \cdot e^{1.7972} \cdot e^{1.2515}$, for the combination of the **UpdateTime** and **DeskAfrica** category, which implies that frequency of realised losses for this combination of preditor variables is $3.4955824 (= \cdot e^{1.2515})$ fold (times) higher than the realised loss frequency of OpRisk causes in the reference desk category, viz. the **Rates** desk.

* $0,003546834 = e^{-9.2147} \cdot e^{1.7972} \cdot e^{1.7758}$, for the combination of the **UpdateTime** and **DeskBonds/Repos** category, which implies that frequency of realised losses for this combination of preditor variables is $5,90500325 (= \cdot e^{1.7758})$ fold higher than causes in the reference desk category.

* $0,001373903 = e^{-9.2147} \cdot e^{1.7972} \cdot e^{0.8274}$, for the combination, which implies that frequency of realised losses for this combination of preditor variables is $2,287363856 (= \cdot e^{0.8274})$ fold higher than the causes in the reference desk category.

* $0,002360693 = e^{-9.2147} \cdot e^{1.7972} \cdot e^{1.3687}$, for the combination, which implies that frequency of realised losses for this combination of preditor variables is $3,930238063 (= \cdot e^{1.3687})$ fold higher than the causes in the reference desk cate-

gory.

* $0,001373903 = e^{-9.2147} \cdot e^{1.7972} \cdot e^{0.8274}$, for the combination with **DeskMM**,an increase of 39%) w.r.t the baseline (the **Rates** desk)

* $0,005012603 = e^{-9.2147} \cdot e^{1.7972} \cdot e^{2.1217}$, for the combination, which implies that frequency of realised losses for this combination of preditor variables is $8,345312467 (= \cdot e^{2.1217})$ fold higher than the causes in the reference desk category.

The predicted mean frequency of realised losses for OpRisk incident i, for the model **freqfit1**, is given by:

$$\begin{aligned}
\mu_i &= \text{exposure}_i \cdot e^{-9.2147 \cdot \text{Intercept}_i} \cdot e^{1.7972 \cdot \text{UpdatedTime}_i} \cdot e^{1.2515 \cdot \text{DeskAfrica}_i} \\
&\cdot \; e^{1.7758 \cdot \text{DeskBonds/Repos}_i} \cdot e^{0.8274 \cdot \text{DeskCommodities}_i} \cdot e^{1.3687 \cdot \text{DeskEquity}_i} \\
&\cdot \; e^{0.3910 \cdot \text{DeskMM}_i} \cdot e^{2.1217 \cdot \text{DeskPrime Services}_i} \cdot e^{-0.7055 \cdot \text{DeskSND}_i} \qquad (3.8)
\end{aligned}$$

We now fit a more comprehensive model where we introduce more variables, in which show realised losses for quarterly OpRisk incidents for an all inclusive case. We will use “LossesIndicator” as the dependent variable, while the other variables will be predictor variables.

```
freqfit <- glm(LossesIndicator ~ UpdatedDay + UpdatedTime +
                 TradedDay + TradedTime + Desk + CapturedBy +
                 TradeStatus + TraderId + Instrument + Reason
               + EventTypeCategoryLevel1 + BusinessLineLevel1,
data=d1, family=poisson(link = 'log'), offset = log(exposure))
```

Which yields output (in summarised form):

```
Call:
glm(formula = LossesIndicator ~ UpdatedDay + UpdatedTime + TradedDay +
    TradedTime + Desk + CapturedBy + TradeStatus + TraderId +
    Instrument + Reason + EventTypeCategoryLevel1 + BusinessLineLevel1,
    family = poisson(link = "log"), data = d1, offset = log(exposure))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6205  -0.3700  -0.1056  -0.0295   4.0726

Coefficients:
                                                        Estimate Std. Error z valu
(Intercept)            -8.953252   0.604562 -14.809 < 0.0000000000000002 ***
UpdatedDay             -0.006976   0.008140  -0.857             0.391428
UpdatedTime             1.113913   0.564165   1.974             0.048331 *
TradedDay              -0.012303   0.006368  -1.932             0.053382 .
TradedTime              0.101378   0.637529   0.159             0.873656
DeskAfrica              1.899956   0.446050   4.260   0.0000204875303586 ***
DeskBonds/Repos         2.803220   0.334324   8.385 < 0.0000000000000002 ***
DeskCommodities         0.747527   0.364630   2.050             0.040355 *
DeskDerivatives         0.683199   0.374174   1.826             0.067867 .
DeskEquity              1.507079   0.321232   4.692   0.0000027113532659 ***
DeskManagement/Other   -2.054697   1.082815  -1.898             0.057755 .
DeskMM                  1.544054   0.453315   3.406             0.000659 ***
DeskPrime Services     -0.028783   0.960227  -0.030             0.976087
DeskSND                 0.766563   0.573844   1.336             0.181602
\vdots                  \vdots     \vdots     \vdots            \vdots

BusinessLineLevel1BL1   1.537103   0.636829   2.414             0.015792 *
BusinessLineLevel1BL3  -0.359123   0.514434  -0.698             0.485119
BusinessLineLevel1BL4  -1.384293   0.391691  -3.534             0.000409 ***
BusinessLineLevel1BL5  -1.169766   0.394350  -2.966             0.003014 **
BusinessLineLevel1BL6   1.250498   1.002141   1.248             0.212095
BusinessLineLevel1BL7   0.875839   1.746369   0.502             0.616005
BusinessLineLevel1BL9   4.214689   1.598598   2.636             0.008377 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2767.9  on 2329  degrees of freedom
Residual deviance: 1821.2  on 2252  degrees of freedom
AIC: 2719.2

Number of Fisher Scoring iterations: 13
```

Model selection and multimodel inference

The selection of the best-fit model from the list of possible combinations of predictor variables traditionally follows of a process removing/adding each variable progressively after each estimation, and propagating backward/forward, comparing goodnes of fit tests at each stage. For example, if we compare the values of the Aikaike information criteria (AIC) for the bivariate model **freqfit1** and the multivariate model **freqfit**, by AICs; we see that for the first model the AIC value is 3225.7 and 2719.2 for the second model, which suggests that the second model, **freqfit**, the model in which we considered an all inclusive list of 15 predictor variables is a better fit since the AIC reduces in magnitude the first, hence **freqfit** is preferable to the first.

In a similar way, we can estimate the models comparing each one which enables one to choose the most appropriate or "best" fit one, by first checking if the model is significant, i.e. if the Residual deviance and the corresponding number of degrees of freedom doesn't have a value significantly bigger than 1: In the latter model $\frac{1821.2}{2252} = 0.808703374$, and then retaining the one with the smaller AIC value.

Burnham and Anderson (2002) introduces information-theoretic approach that allows a data-based selection of a "best" model in the anaysis of our dataset, and a ranking and weighting of the remaining models. These approaches allow traditional (formal) statistical inference to be based on the selected "best" model, which is now based on more than one model (multimodel inference). To do this we are required to load the "MuMIn" package in R.

```
require(MuMIn)
```

Then, we use "dredge" function to generate models using combinations of the terms in the global model. The function will also calculate AICc values and rank models according to it. Note that AICc is AIC corrected for finite sample sizes. The process of analyzing data where the experimentalist has few or no a priori information, thus "all possible models" are considered by subjectively ad iteratively searching the data for patterns and "significance", is often called "data mining", “data snooping” or the term “data dredging”.

```
options(na.action=na.fail)
freqfits <- dredge(freqfit)
```

The function “MuMLn::dredge” returns a list of 4096 models, which is every combination of predictor variable in the global model freqfit. Model number 2942 is the best model and shows that all predictor variables included in the model have a positive effect on the target variable except for the preditor TrddD (**TradedDay**) which has a negative effect on the likelihood og a realised loss (target variable *LossIndicator*). Additionally, from the delta (=delta AIC) one cannot distinguish model 2942 from 3966 and 2878 since (using the common thumb rule) they have AIC < 2.

The top three models, models $2942, 3966$ & 2878 each include nine, ten and seven predictor variables respectively, and where a variable doesn’t have a value, it means that it was not included in the model, not that it does not have and effect. For example model 2942 returns a combination of the seven variables 1/2/3/4/5/6/7/8/10, corresponding to the following output predictor variables (abbreviated in the header row) below:

```
Model selection table
     (Intrc) BsLL1 Desk ETCL1 Instr Reasn TrddD   TrdrI TrdSt UpdtT
2942  -9.107     +    +     +     +     + -0.011660   +     + 1.2760000
```

Information from the AICc’s values suggest, that of the top three models have similar support, and their Akaike weights are not high relative to the $[0, 1]$ weight range; This is characteristic of the endemic nature of data dredging, as the literature suggests (Burnham and Anderson, 2002), and should generally be avoided to curb attendant inferential problems if a single model is chosen, e.g the risk of finding spurious effects, overfitting, etc. .Burnham and Anderson (2002) advises that model averaging is useful in finding a confirmatory result as estimates of precision should include model selection uncertainty. Even so, one can rule out many models on a priori grounds.

We now use “get.models” function to generate a list in which its objects are the fitted models. We will also use the “model.avg” function to do a model averaging based on AICc. Note that “subset=TRUE” will make the function calculate the average model (or mean model) using all models. However, we want to get only the models that have delta AICc < 2; we threfore use “subset=delta<2”

```
adelmodel <- (model.avg(get.models(freqfits, subset=delta<2)))
```

Now we have AICc values for our models and we have the average (mean) model.

Multimodel inference leads to more robust inferences, especially in the point of view that the selection of the model used to estimate the mean frequency must, at the same time, serve the ultimate root cause analysis objective of OpRisk control, that decide calculating capital requirement, in OpVaR measures, taking into account as many characteristics of the trading OpRisk dataset as possible, as well considering how the variables interact with each other.

Modelling population size of the OpRisk events

We have gained initial insights through data exploration in Section 3 and then built models. The next critical step is to evaluate our model. For this we need to use a testing dataset whose function is to provide error estimates of the final result. The testing dataset is not used in building or even fine tuning the models that we build, for the sake of model building define a training dataset and a validation dataset to test different parameter setings or different choices of variables in the data mining part of the project.

We have a population of $K = 2330$ OpRisk events over the first quarter Q12013, and of these events we have a number $N = 371$ of realised losses. N is a discrete random variable modelled as a Poisson variable with rate λ. Each loss X_i is another random variable with an underlying sverity distribution. How does the size K of the population enter the risk model?. It doesn't appear explicitly in the model (Parodi, 2014), however, it is taken into account during the creation of the model. Intuitively, the poisson rate λ is likely to be proportional to the current OpRisk sample size, or more specifically, it is the rate of some expected operational event over per specified time interval. Predicting test set results and evaluating the parameter λ

```r
av.pred <- predict(adelmodel, test_set)

MASS::fitdistr(exp(av.pred), "Poisson")
```

yields a daily rate of $\lambda = 0.001840831$ or 0.1840831% per day.

By a simple growth formula, five years of data (20 quarters) i.e., 3 months * 20 = 5 years:

$$\begin{aligned} 5yr_Population &= Initial_Population * (1+\lambda)^n \\ 5yr_Population &= 2330 * (1+0.18009498)^20 \\ 5yr_Population &= 63929 \end{aligned} \tag{3.9}$$

corresponds to a 5yr population of 63929 observations. What remains is to use the extrapolation script to generate the simulated dataset.

The estimation of some generalised additive models for location scale and shape (GAMLSS) for severity of loss

We introduce a Box-Cox Power Exponential distribution (BCPE), which is a four parameter distribution, for fitting a GAMLSS to estimate the (non-linear nature) mean OpRisk loss severity using the gamlss function. The mean daily loss severities in the risk correction statistics is estimated through the BCPE gamlss model.

The pdf of the BCPE distribution is defined as:

$$
\begin{aligned}
f(y|\mu,\sigma,\nu,\tau) &= (y^{(\nu-1)/\mu^n u)} \cdot \frac{\tau}{\sigma} \cdot \frac{e^(-0.5 \cdot |\frac{z}{c}|^\tau)}{(c \cdot 2^(1 + \frac{1}{tau})} \cdot \Gamma(\frac{1}{\tau})) \\
\text{where} \quad c &= [2^(\frac{-2}{\tau}) \cdot \frac{\Gamma(\frac{1}{\tau})}{\Gamma(\frac{3}{\tau})}]^{0.5}, \quad \text{where if} \quad \nu! = 0, \quad \text{then} \\
Z &= \frac{(\frac{y}{\mu})^\nu - 1}{\nu \cdot \sigma}, \quad \text{else} \quad z = \frac{log\frac{y}{\mu}}{\sigma}, \\
\text{for} \quad y > 0 \quad &, \quad \mu > 0, \sigma > 0, \nu = \text{(-Inf,+Inf)} \quad \text{and} \quad \tau > 0.
\end{aligned}
\tag{3.10}
$$

The BCPE adjusts the obove density $f(y|\mu,\sigma,\nu,\tau)$, resulting from the condition $y > 0$. See Stasinopoulos, Rigby, Heller, Voudouris, and De Bastiani (2017) . We now consider a model where the *Loss* is the target variable: The following fits the model and checks it.

```
library(gamlss)
sf <- gamlss(Losses~cs(UpdatedDay + UpdatedTime + TradedDay +
                TradedTime + Desk + CapturedBy + TradeStatus
               + TraderId + Instrument + Reason +
                 EventTypeCategoryLevel1 + BusinessLineLevel1),
sigma.formula=~cs(UpdatedDay + UpdatedTime + TradedDay +
                TradedTime + Desk + CapturedBy + TradeStatus
               + TraderId + Instrument + Reason +
               EventTypeCategoryLevel1 + BusinessLineLevel1),
nu.formula=~cs(UpdatedDay + UpdatedTime + TradedDay + TradedTime
               + Desk + CapturedBy + TradeStatus + TraderId +
              Instrument + Reason + EventTypeCategoryLevel1 +
              BusinessLineLevel1),
 tau.formula=~cs(UpdatedDay + UpdatedTime + TradedDay + TradedTime
                + Desk + CapturedBy + TradeStatus + TraderId +
                Instrument + Reason + EventTypeCategoryLevel1 +
                BusinessLineLevel1),
data=D1, mu.start = NULL,  sigma.start = NULL, nu.start = NULL,
                              tau.start = NULL, family=BCPE)
```

CHAPTER 4

METHODS FOR MODELING OPRISK DEPENDING ON COVARIATES

Introduction

This section of the paper concentrates on combining various supervised learning techniques with extreme value theory (EVT) fitting, which is very much based on the Dynamic EVT-POT model developed by Chavez-Demoulin et al. (2016). This can only happen due to an abundance of larger and better quality datasets and which also benefits the loss distribution approach (LDA) and other areas of OpRisk modeling. In Chavez-Demoulin et al. (2016), they consider dynamic models based on covariates and in particular concentrate on the influence of internal root causes that prove to be useful from the proposed methodology.

Motivated by the abundance of data and better data quality, these new data-intensive techniques offer an important tool for ORM and at the same time supporting the call from industry for a new class of EBOR models that capture forward-looking aspects of ORM (Embrechts et al., 2018). Three different machine learning techniques viz., decision trees, random forest, and neural networks, will be employed using R. A comprehensive list of user defined variables associated with root causes that contribute to the accumulation of OpRisk events (frequency) has been provided, moreover, a lot can be gained from this dataset as it also bears the impacts of these covariates on the severity of OpRisk.

Modeling Oprisk: The loss distribution approach (LDA)

Twenty-one key risk indicators (kri's) with eight feature groups including person identification, trade origination, root causes and market value sensitivities are in the chosen covariates. For each risk event there is information about: trading risk exposure, trading characteristics, causal factor characteristics and the losses created by these factors. The development, training and validation of the machine learning (ML) models lends itself to this new type of data and requires a higher degree of involvement across operations. Moreover, at this level of granularity the different types of data is particularly suited to exposure-based treatment, and other forward-looking aspects within the OpRisk framework, for improved forecasts of OpRisk losses.

The aggregated operational losses can be seen as a sum S of a random number N individual operational losses

$$(X_1, \dots, X_N)$$

. The total required capital is the sum of VaR of each BL/ET combination calibrated through the underlying mathematical model whose analytic expression is given by:

$$\mathbf{G}_{\vartheta(t)}(x) = Pr[\vartheta(t) \leq x] = Pr\left(\sum_{n=1}^{N(t)} X_n \leq x\right), \qquad \text{where} \quad \vartheta(t) = \sum_{n=1}^{N(t)} X_n. \tag{4.1}$$

$\mathbf{G}(t)$ can only be obtained numerically using the Monte Carlo method, Panjer's recursive approach, and the inverse of the characteristic function (Frachot et al. (2001); Aue and Kalkbrener (2006); Panjer (2006); & others).

Research Objective 2

To test the accuracy of several classes of data-intensive techniques in approximating the weights of the risk factors; i.e., the input features of the model viz., TraderID, UpdatedDay, Desk, etc. of the underlying value-adding processes, against traditional statistical techniques, in order to separately estimate the frequency and severity distribution of the OpRisk losses from historical data. As a consequence, capital estimates should be able to adapt to changes in the risk profile e.g., upon the addition of new products or varying the business mix of the bank (e.g., terminations, voids, allocations, etc.) to provide sufficient incentives for ORM to mitigate risk (Einemann et al., 2018).

Theoretical investigations for the quantification of modern ORM

Within the variety of relations among risk preferences, people have difficulty in grasping the concept of risk-neutrality. In a market where securities are traded, risk-neutral probabilities are the cornerstone of trade, due to their importance in the law of no arbitrage for securities pricing. Mathematical finance is concerned with pricing of securities, and makes use of this idea: That is, assuming that arbitrage activities do not exist, two positions with the same pay-off must also have an identical market value (Gisiger, 2010). A position (normally a primary security) can be replicated through a construction consisting of a linear combination of long, as well as short positions of traded securities. It is a relative pricing concept which removes risk-free profits due to the no-arbitrage condition.

This idea seems quite intuitive from an OpRisk management perspective. The fact that one can take internal historical loss data and use this to make a statement on the `OpRisk` VaR measure for the population, is based on the underlying assumption of risk neutrality. Consider a series of disjoint risky events occurring at times

τ to $\tau + 1$. We can explore the concept of a two state economy in which value is assigned to gains and losses, rather than to final assets, such that an incremental gain or loss can be realised at state $\tau + 1$, contingent on the probability which positively impacts on the event happening.

Risk-neutral measure $\mathbb{Q}$

Risk-neutral probabilities simply enforce a linear consistency for views on equivalent losses/gains, with regard to the shape of the value function. The shape the graph depicts a linear relationship based on responses to gains/losses and value. The risk neutral probability is not the real probability of an event happening, but should be interpreted as (a functional mapping) of the number of loss events (frequency).

Suppose we have: Θ = Gain/Loss; $\nu(x)$ = risk event happening; and X = Individual gain/loss (or both), then;

$$\begin{aligned} \Theta = \; & \textstyle\sum_{i=1}^{n} \Pr[\nu(x_i)] * X_i \qquad (4.2) \\ \text{where} & \\ & \textstyle\sum_{i=1}^{n} \Pr[\nu(x_i)] = 1 \qquad \text{and} \qquad \Pr[\nu(x_i)] \geq 0 \quad \forall i \end{aligned}$$

Note that the random variable Θ is the sum of the products of frequency and severity for losses (in `OpRisk` there are no gains).

This formula is used extensively in actuarial practices, for decisions relating to quantifying different types of risk, in particular in the quantification of value-at-risk (VaR) (a risk measure used to determine capital adequacy requirements, commonly adopted in the banking industry). A quantile of the distribution of the aggregate losses is the level of exposure to risk, expressed as VaR.

People exhibit a specific four-fold behaviour pattern when facing risk (Shefrin, 2016). There are four combinations of gain/loss and moderate/extreme probabilities, with two choices of risk attitude per combination. OpRisk measurement focuses on only those casual factors that create losses with random uncertainty, for the value adding processes of the business unit.

CHAPTER 5

THEORETICAL INVESTIGATIONS INTO THE QUANTIFICATION OF MODERN ORMF'S

Theoretical investigations for the quantification of modern ORM

Within the variety of relations among risk preferences, people have difficulty in grasping the concept of risk-neutrality. In a market where securities are traded, risk-neutral probabilities are the cornerstone of trade, due to their importance in the law of no arbitrage for securities pricing. Mathematical finance is concerned with pricing of securities, and makes use of this idea.

That is, assuming that arbitrage activities do not exist, two positions with the same pay-off must also have an identical market value (Gisiger, 2010). A position (normally a primary security) can be replicated through a construction consisting of a linear combination of long, as well as short positions of traded securities. It is a relative pricing concept which removes risk-free profits due to the no-arbitrage condition.

This idea seems quite intuitive from an OpRisk management perspective. The fact that one can take internal historical loss data and use this to make a statement on the `OpRisk` VaR measure for the population, is based on the underlying assumption of risk neutrality. Consider a series of disjoint risky events occurring at times τ to $\tau + 1$. We can explore the concept of a two state economy in which value is assigned to gains and losses, rather than to final assets, such that an incremental gain or loss can be realised at state $\tau + 1$, contingent on the probability which positively impacts on the event happening.

Risk-neutral measure $\mathbb{Q}$

Risk-neutral probabilities simply enforce a linear consistency for views on equivalent losses/gains, with regard to the shape of the value function. The shape the graph depicts a linear relationship based on responses to gains/losses and value. The risk neutral probability is not the real probability of an event happening, but should be interpreted as (a functional mapping) of the number of loss events (frequency).

Suppose we have: Θ = Gain/Loss; $\nu(x)$ = risk event happening; and X = Individual gain/loss (or both), then

$$\Theta = \textstyle\sum_{i=1}^{n} \Pr[\nu(x_i)] * X_i \tag{5.1}$$

where

$$\textstyle\sum_{i=1}^{n} \Pr[\nu(x_i)] = 1 \qquad \text{and} \qquad \Pr[\nu(x_i)] \geq 0 \quad \forall i$$

Note that the random variable Θ is the sum of the products of frequency and severity for losses (in `OpRisk` there are no gains).

This formula is used extensively in actuarial practices, for decisions relating to quantifying different types of risk, in particular in the quantification of value-at-risk (VaR) (a risk measure used to determine capital adequacy requirements, commonly adopted in the banking industry).

A quantile of the distribution of the aggregate losses is the level of exposure to risk, expressed as VaR. People exhibit a specific four-fold behaviour pattern when facing risk (Shefrin, 2016). There are four combinations of gain/loss and moderate/extreme probabilities, with two choices of risk attitude per combination. OpRisk measurement focuses on only those casual factors that create losses with random uncertainty, for the value adding processes of the business unit.

Cluster analysis

Cluster analysis (CA) is an unsupervised machine learning technique, which sets out to group combinations of covariates according to levels of similarity into clusters. The CA algorithm attempts to optimise homogeneity within data groups, and heterogeneity between groups of observations. Thus, in the context of ORM, CA regroups these combinations of covariates into clusters (so that features within each group are similar to one another, and different from features in other groups), ordering and prioritising the root causes of losses.

A new and challenging argument can be demonstrated through clustering correlated data objects in the OpRisk dataset, by asserting that clustering should show more than one distinct group. In addition, the more groups of distinct clusters, losses are expected to drop, and losses in distinct clusters should also show a decreasing trend over time, with intensifying exposure. Ultimately, subtle patterns of frequencies and associated severities of losses in the OpRisk data can be revealed.

The OpRisk dataset is subdivided for training patterns, validated and tested with the *k*-means clustering algorithm. To achieve this the *k*-means algorithm randomly subdivides the data in k groups. Firstly, each groups mean is found by clustering the centers in the input variable-space of the training patterns. In each cluster within each group, the significant variables' coefficients which determine cluster have set centers closest to the cluster centers generated by the *k*-means clustering algorithm applied to the input vectors of the training data (Flake, 1998). These clusters have centers closest:- as defined by a differential metric i.e., the Euclidean distance, to a relationship (e.g. a linear combination of coefficients and variables) which most accurately predicts the target variable.

Research Objective 3

To identify potential flaws in the loss distribution approach (LDA) model of ORM by employing CA. The *classical* LDA model, through a mathematical framework derives a negative pay-off function (loss) based on a risk-neutral measure $\mathbb{Q}$. The study addresses weaknesses in the current LDA model framework, by assuming managerial risk-taking attitudes are more risk averse.

More precisely, the goal is to use CA to learn deep hierarchies of features[1] found during operations, to then determine whether risk adverse techniques overcompensate for persistent loss event types over time.

Description of the dataset

The characteristics of the traded transactions or of the associated risk correction event are given by the following variables: Trade, UpdateTime, UpdatedDay, TradedTime, TradedDay, Desk, CapturedBy, TradeStatus, TraderId, Instrument, Reason behind the risk correction event, Nominal, FloatRef floating rate reference for fixed income products, ResetDate and ResetRate, Theta, Loss severity, four EventTypeCategoryLevel viz., EL1 - IF, EL4 - CPBP, EL6 - BDSF, and EL7 - EDPM & all seven associated BusinessLineLevel, and the LossIndicator. The exposure variable shows the length of the time interval from the initial moment when the risk event happened, until the occurrence of a risk correction.

The data is limited to the training dataset over the interval 01 January - 31 March 2013, in Figure 5.1, portrays detail of the trend of OpRisk losses against exposures for each of the 1631 observations and 16 variables. In the first plot, trans-

[1] A typical approach taken in the literature is to use an unsupervised learning algorithm to train a model of the unlabeled data and then use the results to extract interesting features from the data [@coates2012learning]

actions with small exposures are concentrated in the first quadrant where HFLS losses persist. This is in line with the sentiment in risk management circles, that small exposures are not actively managed and hence risk mitigation is not a priority. As a result many of the unforeseeable LFHS losses occur here, as they are not anticipated and therefore slip through OpRisk defences, who more often than not, do not mitigate against these events.

Loss severities decrease with increasing exposures, as seen by the lowering variabilities (and colour concentration of the exposure) between loses and exposures. This support the view that more impactful past losses invoke active risk management and mitigation, as risk managers overcompensate for these severities in their management practices i.e., they are more risk averse. In addition there are graphically displayed correlations (which work for numerical explanatory variables only), which are ordered by their strengths. There is a weak positive relationship between exposure and UpdatedDay, TradedTime & TradedDay; a weak negative relationship with UpdatedTime.

Exploratory data analysis

The estimation of k*-means clustering algorithm*

A cluster analysis will identify groups within a dataset. The target variable is LossIndicator, a binary variable indicating a 1 if a realised loss occurs and 0 for those pending or near misses. The K-means clustering algorithm will search for K clusters (specified by the user). The resulting k clusters are represented by the mean or average values of each of the variables. Let us consider a model where the LossIndicator is the target variable: The user whose task it is to specify k, may guess right or in practice they may obtain a priori, the knowledge of how to select the appropriate k in advance.

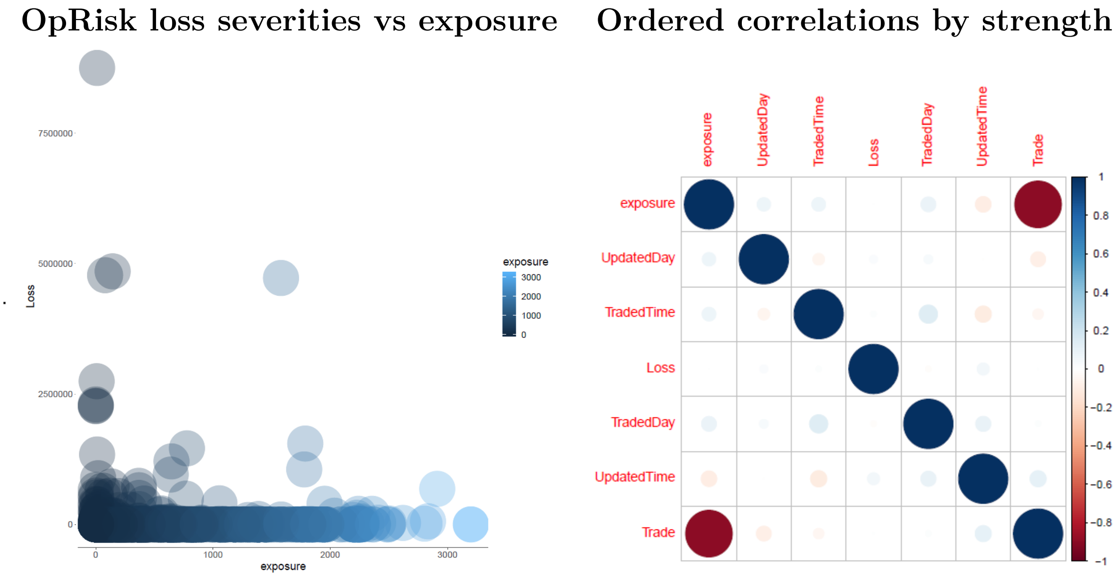


Figure 5.1: Graphically displayed correlations by strength and a plot of OpRisk loss severities vs exposure

Rather than the trial and error method which involves guessing k values and successively computing minimum separation between centers, there are several data mining techniques found in the literature, that can be used to determine the optimal k (Rousseeuw, 1987). The output plot for the estimation of the optimal k is presented in Figure **??** below. We have iterated over cluster sizes from 2 to 10 clusters. The program KMeans resets the random number seed to obtain the same results each time. where the optimal k found to be significant close to $k = 10$.

The plot displays the 'sum(withinss)' for each clustering and the change in this value from the previous clustering. The Sum(WithinSS) (blue line) as a performance metric indicates that beyond $k = 4$ clusters the model overfits: Its computes the absolute error which is initially large, then monotonicaly decreases to the point $k = 4$, it then begins to increase subsequent to the point where the Diffprevious Sum(WithinSS) (red line) intersects viz., at $k = 4$ clusters, which means $k = 4$ is the local optimal number of clusters i.e., beyond which the iterative relative errors converges faster than the absolute errors and successively reduces as k increases

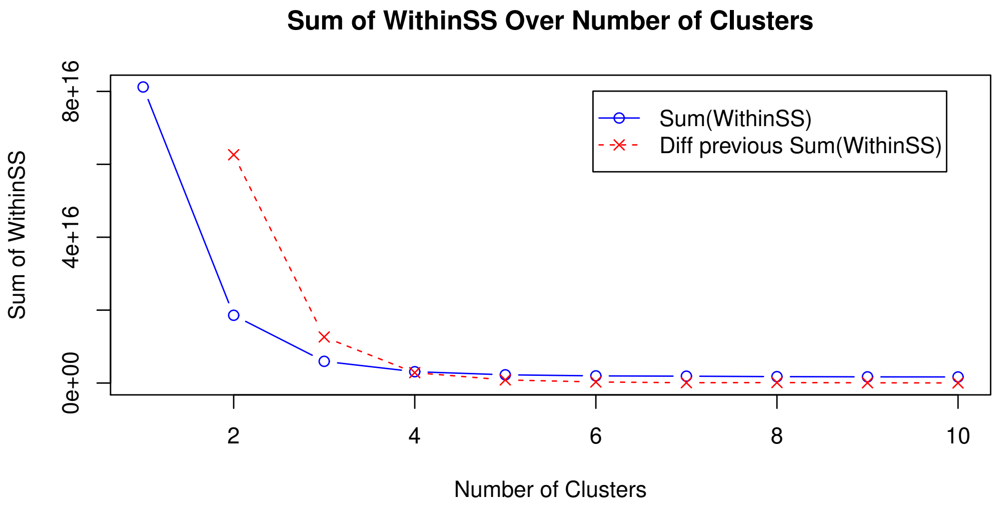


Figure 5.2: Finding the optimal number of k groups by the Silhouette Statistic SS: Sum is a measure to approximate the optimal number of k groups by the Silhouette Statistic SS

from 4 to 10.

Rattle program code

Results

```
Cluster sizes:


[1] "478 404 570 179"


Data means:


      Trade  UpdatedDay UpdatedTime   TradedDay  TradedTime

0.762016409 0.448559166 0.486589314 0.487369712 0.601539912

       Loss    exposure

0.003232348 0.121083376
```

```
Cluster centers:

      Trade UpdatedDay UpdatedTime TradedDay TradedTime        Loss
1 0.8106844  0.3943515   0.4123358 0.2912134  0.8556825 0.004692829
2 0.8716248  0.4900990   0.5409218 0.7948845  0.8270263 0.002132631
3 0.8378683  0.4493567   0.5264944 0.4160234  0.2165842 0.002308103
4 0.1431301  0.4970205   0.4351758 0.5443203  0.6397973 0.004757466
    exposure
1 0.08060460
2 0.06359981
3 0.07134609
4 0.51729829

Within cluster sum of squares:

[1]  84.88017  89.27845 148.89661  59.37208

Time taken: 1.86 secs

Rattle timestamp: 2018-12-13 07:22:48 User
```

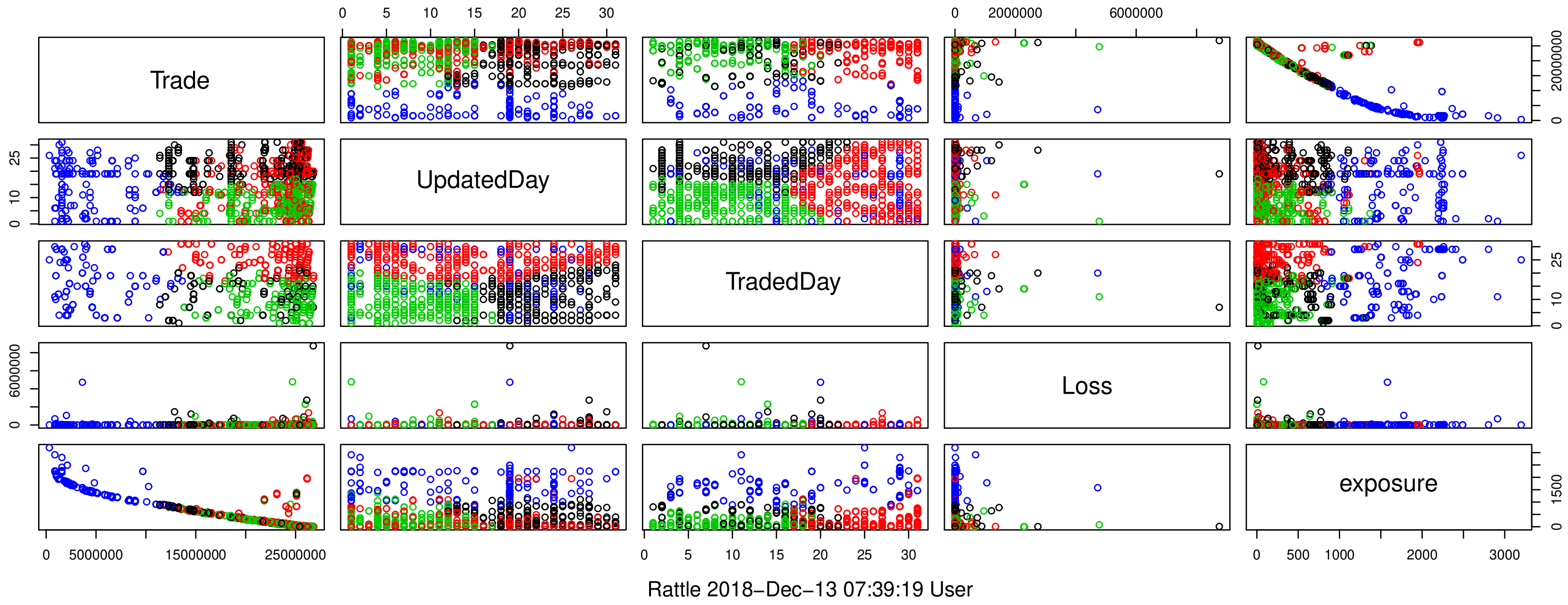


Figure 5.3: A scatterplot matrix for the *k*-means clustering of size 4, and the covariates of frequency loss events consisting of 369 loss event frequencies amounting to R 61 534 745 P&L severity of loss impact.

CHAPTER 6

RESULTS

The power of intuitive understanding will protect you from harm until the end of your days. — Lao Tzu

Introduction

Results

```
## Error in eval(expr, envir, enclos): object 'da36361.0001' not found

## Error in .f(.x[[i]], ...): object 'cigrec' not found

## Error: Column indexes must be at most 15 if positive, not 16, 17, 18

## Error: Column indexes must be at most 15 if positive, not 19, 20, 21, 22, 23, 2

## Error in mutate_impl(.data, dots): Evaluation error: object 'cigrec' not found.

## Error in library(here): there is no package called 'here'

## Error in here("Data/NSDUH_2014_Results.rda"): could not find function "here"

## Error in library(survey): there is no package called 'survey'

## Error in svydesign(ids = ~1, strata = ~vestr, weights = ~analwt_c, data = d1):

## Error in svyglm(self ~ religious + age2 + irsex + newrace2 + irfamin3 + : could

## Error in svyglm(peer ~ religious + age2 + irsex + newrace2 + irfamin3 + : could

## Error in svyglm(dep ~ religious + age2 + irsex + newrace2 + irfamin3 + : could

## Error in coef(obj): object 'svy_a1' not found

## Error in rownames(est1) = c("Respondent", "Peer", "Depression"): object 'est1'

## Error in data.frame(est1): object 'est1' not found

## Error in svyglm(model, design = design, family = "quasibinomial"): could not fi

## Error in vcov.default(object): object does not have variance-covariance matrix

## Error in rownames(est2) = c("Tobacco", "Rx", "Marijuana", "Illicit"): object 'e
## Error in data.frame(est2): object 'est2' not found
```

```
## Error in loadNamespace(name): there is no package called 'anteo'
```

Conclusions

CHAPTER 7

DISCUSSION

> *A model is a simplification or approximation of reality and hence will not reflect all of reality. ... While a model can never be "truth," a model might be ranked from very useful, to useful, to somewhat useful, to, finally, essentially useless.* — Burnham and Anderson, 2002

General Discussion

Findings from the Three Chapters

Limitations

Future Research

Conclusions

REFERENCES

Acharyya, M. (2012). Why the current practice of operational risk management in insurance is fundamentally flawed: Evidence from the field. In *ERM symposium, april* (pp. 18–20).

Agostini, A., Talamo, P., and Vecchione, V. (2010). Combining operational loss data with expert opinions through advanced credibility theory. *The Journal of Operational Risk*, *5*(1), 3.

Altman, M. (2008). Behavioral economics, economic theory and public policy.

Aue, F., and Kalkbrener, M. (2006). LDA at work: Deutsche bank's approach to quantifying operational risk. *Journal of Operational Risk*, *1*(4), 49–93.

Badescu, A. L., Lan, G., Lin, X. S., and Tang, D. (2015). Modeling correlated frequencies with application in operational risk management.

Barberis, N., and Thaler, R. (2003). A survey of behavioral finance. *Handbook of the Economics of Finance*, *1*, 1053–1128.

Burnham, K., and Anderson, D. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* Springer-Verlag.

Cameron, A. C., and Trivedi, P. K. (2013). *Regression analysis of count data* (Vol. 53). Cambridge university press.

Chau, V. (2014). *Robust estimation in operational risk modeling* (Master's thesis).

Chavez-Demoulin, V., Embrechts, P., and Hofert, M. (2016). An extreme value approach for modeling operational risk losses depending on covariates. *Journal of Risk and Insurance*, *83*(3), 735–776.

Chavez-Demoulin, V., Embrechts, P., and Nešlehová, J. (2006). Quantitative models for operational risk: Extremes, dependence and aggregation. *Journal of Banking & Finance*, *30*(10), 2635–2658.

Committee, B., and others. (2010). Basel iii: A global regulatory framework for more resilient banks and banking systems. *Basel Committee on Banking Supervision, Basel.*

Committee, B., and others. (2011). Operational risk–Supervisory guidelines for the advanced measurement approaches. *Basel: Bank for International Settlements.*

Covrig, M., Mircea, I., Zbaganu, G., Coser, A., Tindeche, A., and others. (2015). Using r in generalized linear models. *Romanian Statistical Review*, *63*(3), 33–45.

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

& Sons New York,

De Jong, P., Heller, G. Z., and others. (2008). Generalized linear models for insurance data. *Cambridge Books*.

Denuit, M., Maréchal, X., Pitrebois, S., and Walhin, J.-F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons.

Dorval, M. (2013). *Achieving Basel III compliance: how to tackle it and business issues* (pp. 1–12). Retrieved from http://www.risktech-forum.com/research/achieving-basel-iii-compliance-how-to-tackle-it-and-business-issues

Einemann, M., Fritscher, J., and Kalkbrener, M. (2018). Operational risk measurement beyond the loss distribution approach: An exposure-based methodology.

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. *Risk Management: Value at Risk and Beyond*, *1*, 176–223.

Embrechts, P., Mizgier, K. J., and Chen, X. (2018). Modeling operational risk depending on covariates. an empirical investigation.

Flake, G. W. (1998). Square unit augmented radially extended multilayer perceptrons. In *Neural networks: Tricks of the trade* (pp. 145–163). Springer.

Frachot, A., Georges, P., and Roncalli, T. (2001). Loss distribution approach for operational risk.

Frees, E. W., and Sun, Y. (2010). Household life insurance demand: A multivariate two-part model. *North American Actuarial Journal*, *14*(3), 338–354.

Friedman, M., and Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy*, *56*(4), 279–304.

Galloppo, G., and Previati, D. (2014). A review of methods for combining internal and external data.

Gigerenzer, G., and Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143.

Gisiger, N. (2010). Risk-neutral probabilities explained.

Hemrit, W., and Arab, M. B. (2012). The major sources of operational risk and the potential benefits of its management. *The Journal of Operational Risk*, *7*(3), 71–92.

Hoohlo, M. (2015). *A new internal data measure for operational risk: A case study of a south african bank* (PhD thesis).

Jongh, R. de, De Wet, T., Raubenheimer, H., and Venter, J. H. (2015). Combining scenario and historical data in the loss distribution approach: A new procedure that incorporates measures of agreement between scenarios and historical data.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697.

Kahneman, D., and Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part i* (pp. 99–127). World Scientific.

King, J. L. (2001). Operational risk: Measurement and modelling (the wiley finance series).

Kuhnen, C. M., and Knutson, B. (2005). The neural basis of financial risk taking. *Neuron*, *47*(5), 763–770.

List, J. A. (2004). Neoclassical theory versus prospect theory: Evidence from the marketplace. *Econometrica*, *72*(2), 615–625.

Mignola, G., Ugoccioni, R., and Cope, E. (2016). Comments on the basel committee on banking supervision proposal for a new standardized approach for operational risk.

Morgenstern, O., and Von Neumann, J. (1953). *Theory of games and economic behavior*. Princeton university press.

Opdyke, J. D. (2014). Estimating operational risk capital with greater accuracy, precision, and robustness. *arXiv Preprint arXiv:1406.0389*.

Panjer, H. H. (2006). *Operational risk: Modeling analytics* (Vol. 620). John Wiley & Sons.

Parodi, P. (2014). *Pricing in general insurance*. CRC Press.

Peters, G., Shevchenko, P. V., Hassani, B., and Chapelle, A. (2016). Should the advanced measurement approach be replaced with the standardized measurement approach for operational risk?

Risk, B. O. (2001). Supporting document to the new basel capital accord. *Consultative Document, January*, *200*.

Risk, B. O. (2016). Standardised measurement approach for operational risk. *Consultative Document, June*.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Shefrin, H. (2016). *Behavioral risk management: Managing the psychology that drives decisions and influences operational risk*. Springer.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible regression and smoothing: Using gamlss in r*. Chapman;

Hall/CRC.

Tom, S. M., Fox, C. R., Trepel, C., and Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, *315*(5811), 515–518.

Urbina, J., and Guillén, M. (2014). An application of capital allocation principles to operational risk and the cost of fraud. *Expert Systems with Applications*, *41*(16), 7023–7031.

Wiseman, R. M., and Catanach Jr, C. (1997). A longitudinal disaggregation of operational risk under changing regulations: Evidence from the savings and loan industry. *Academy of Management Journal*, *40*(4), 799–830.

Wood, S. N. (2017). *Generalized additive models: An introduction with r*. Chapman; Hall/CRC.

APPENDICES

–>

Appendix A: R Code for Chapter 3

Required: R Packages from CRAN

Required: R Packages from GitHub

```r
if (!require(MarginalMediation)){
  devtools::install_github("tysonstanley/MarginalMediation")
  library(MarginalMediation)
}
```

Exploratory Data Analyses

Data Preparations for tables and figures around page

```
file_loc <- "C:/Users/User/Documents/OpRiskPHDGitHub/OpRisk_PHD_Dissertate
/OpRisk_PHD_Dissertation"
setwd(file_loc)
list.files(file_loc)

frequency <- openxlsx::read.xlsx("Raw_Formatted_Data.xlsx",
                                 check.names = TRUE, sheet = "Frequency")
severity <- openxlsx::read.xlsx("Raw_Formatted_Data.xlsx",
                                check.names = TRUE, sheet = "Severity")
projdata <- openxlsx::read.xlsx("OPriskDataSet_exposure.xlsx",
                                check.names = TRUE, sheet = "CleanedData")
```

```
pander::pander(tapply(projdata$Loss, INDEX = projdata$Instrument, function(x)
  c(N = length(x), Mean = mean(x), SD = sd(x), Min = min(x), Max = max(x))))

tablex <- do.call("rbind", lapply(split(projdata$Loss, projdata$Instrument),
function(x) c(N = length(x), Mean = mean(x), SD = sd(x), Min = min(x),
              Max = max(x))))
tablex <- cbind(Instrument = rownames(tablex), tablex)
tablex <- as.data.frame(tablex)
tablex$Mean <- as.numeric(as.character(tablex$Mean))

tablex <- tablex[order(tablex$Mean), ]

stargazer::stargazer(tablex)

openxlsx::write.xlsx(tablex, "tablex.xlsx", rownames = TRUE)
```

Figure on page

```
# Exploratory data analysis for Update Time
### summary statistics
summary(projdata$UpdatedTime)
### Histograms
par(mfrow=c(1,3))
### ALL Losses
hist(projdata$UpdatedTime, col = "blue", main = "All losses", xlab =
       "Update Time", ylab = "Frequency")
### Near Misses/Pending Losses
hist(projdata$UpdatedTime[projdata$LossIndicator == 0], col = "red", main =
       "Near Misses", xlab = "Update Time", ylab = "Frequency")
### Realised losses
hist(projdata$UpdatedTime[projdata$LossIndicator == 1], col = "green", main
     = "Realised losses", xlab = "Update Time", ylab = "Frequency")
par(mfrow=c(1,1))
```

```
###
plot(projdata$UpdatedTime, log(projdata$Loss+0.000000001), ylim = c(6, 18),
     col = "navy", xlab = "Updated Time", ylab = "Log. Loss")
do.call("rbind", lapply(split(projdata$Loss, projdata$UpdatedTime),
function(x) c(N = length(x), Mean = mean(x), SD = sd(x), Min = min(x), Max = max(x))))
```

Figures , , and on pages , and , respectively.

```
# Instrument
unique(projdata$Instrument)
table(projdata$LossIndicator, projdata$Instrument)
plot(table(projdata$LossIndicator, projdata$Instrument),
    main="By Instrument", col=rainbow(20), las=1, cex.axis=1.0)

# Trader
unique(projdata$TraderId)
table(projdata$TraderId, projdata$LossIndicator)
round(addmargins(prop.table(table(projdata$TraderId, projdata$LossIndicator)
                                                               , 1), 2)*100, 1)
plot(table(projdata$LossIndicator, projdata$TraderId), main="By Trader",
                                      col=rainbow(20), las=1, cex.axis=1.0)

# Captured By
table(projdata$CapturedBy, projdata$LossIndicator)
round(addmargins(prop.table(table(projdata$CapturedBy, projdata$LossIndicator)
                                                                 , 1), 2)*100, 1)
plot(table(projdata$LossIndicator, projdata$CapturedBy), main="By Tech Support"
                                    , col=rainbow(20), las=1, cex.axis=1.0)

do.call("rbind", lapply(split(projdata$Loss, projdata$CapturedBy), function(x)
  c(N = length(x), Mean = mean(x), SD = sd(x), Min = min(x), Max = max(x))))
```

Examples from Chapter 3

Data Preparation

Data preparation using the OpRisk Loss Collection Data Exercise (LCDE), as described in Chapter 2.

```r
# Set parameter values
crv$seed <- 42 # set random seed to make your partition reproducible
crv$taining.proportion <- 0.7 # proportion of data used for training
crv$validation.proportion <- 0.15 # proportion of data used for validation

# Load data for frequency of LossIndicator analysis
d <- read.csv("OPriskDataSet_exposure.csv",
              sep=";",
              dec=",",
              na.strings=c(".", "NA", "", "?"),
              strip.white=TRUE, encoding="UTF-8")

exposure <- d[,ncol(d)]
class(exposure)
length(exposure)

summary(d)

d1 <- d %>%
  group_by(UpdatedDay,
           UpdatedTime,
           TradedDay,
           TradedTime,
           Desk,
           CapturedBy,
           TradeStatus,
           TraderId,
           Instrument,
           Reason,
           EventTypeCategoryLevel1,
           BusinessLineLevel1) %>%
  transmute(LossesIndicator = LossIndicator,
            Losses = Loss,
            exposure = exposure)

# Load data for severity of losses analysis

D <- read.csv("OPriskDataSet_exposure_severity.csv",
              sep=";",
              dec=",",
              na.strings=c(".", "NA", "", "?"),
              strip.white=TRUE, encoding="UTF-8")
```

```
exposure <- D[,ncol(D)]

D1 <- D %>%
  group_by(UpdatedDay,
           UpdatedTime,
           TradedDay,
           TradedTime,
           Desk,
           CapturedBy,
           TradeStatus,
           TraderId,
           Instrument,
           Reason,
           EventTypeCategoryLevel1,
           BusinessLineLevel1) %>%
transmute(LossesIndicator = LossIndicator,
            Losses = Loss,
            exposure = exposure)
```

GLM Models

```
getmode <- function(x){
  u <- unique(x)
  as.integer(u[which.max(tabulate(match(x,u)))])
}

for (i in 5:(ncol(d1) - 3)){
     d1[[i]] <- relevel(d1[[i]], getmode(d1[[i]]))
}
```

```
freqfit <- glm(LossesIndicator ~ UpdatedDay + UpdatedTime + TradedDay
               + TradedTime + Desk + CapturedBy + TradeStatus +
               TraderId + Instrument + Reason +
               EventTypeCategoryLevel1 + BusinessLineLevel1, data=d1,
               family=poisson(link = 'log'), offset = log(exposure))
```

```
options(na.action=na.fail)
freqfits <- dredge(freqfit)
adelmodel <- (model.avg(get.models(freqfits, subset=delta<2)))
```

GAMLSS Model

```
sf <- gamlss(Losses~cs(UpdatedDay + UpdatedTime + TradedDay + TradedTime
      + Desk + CapturedBy + TradeStatus + TraderId + Instrument + Reason
      + EventTypeCategoryLevel1 + BusinessLineLevel1),
sigma.formula=~cs(UpdatedDay + UpdatedTime + TradedDay + TradedTime + Desk
      + CapturedBy + TradeStatus + TraderId + Instrument + Reason +
        EventTypeCategoryLevel1 + BusinessLineLevel1),
nu.formula=~cs(UpdatedDay + UpdatedTime + TradedDay + TradedTime + Desk +
      CapturedBy + TradeStatus + TraderId + Instrument + Reason +
      EventTypeCategoryLevel1 + BusinessLineLevel1),
 tau.formula=~cs(UpdatedDay + UpdatedTime + TradedDay + TradedTime + Desk +
      CapturedBy + TradeStatus + TraderId + Instrument + Reason +
      EventTypeCategoryLevel1 + BusinessLineLevel1),
data=D1, mu.start = NULL,  sigma.start = NULL, nu.start = NULL,
                                            tau.start = NULL, family=BCPE)
```

Appendix B: R Code for Chapter 6

Required: R Packages from CRAN

```r
if (!require(tidyverse)){
  install.packages("tidyverse")
  library(tidyverse)
}
if (!require(furniture)){
  install.packages("furniture")
  library(furniture)
}
if (!require(here)){
  install.packages("here")
  library(here)
}
if (!require(devtools)){
  install.packages("devtools")
  library(devtools)
}
if (!require(survey)){
  install.packages("survey")
  library(survey)
}
```

Required: R Packages from GitHub

```r
if (!require(MarginalMediation)){
  devtools::install_github("tysonstanley/MarginalMediation")
  library(MarginalMediation)
}
```

Extrapolation code for simulation in matlab

Notably, the code for the predict condition was run via the Matlab Terminal:

```
% Updated Time
% generate the vector DD

DDD = 1:31;

% generate the vector VVV

VVV = 1:12;

% generate the vector UUU

UUU = 2013 : -1 : 2006;

% making the full thrity five days vector
% Years
Thirty_five_days = [UUU';UUU';UUU';UUU(1:end-1)'];
%Months
Thirty_five_days2 = [VVV';VVV';VVV(1:7)'];
% Days
Thirty_five_days3 = DDD';

% The updated time algorithm

% initializing the time matrix
for i = 1 : length(UUU)

    UUU_trans{i} = num2cell(zeros(1,12));

end

for i = 1 : length(UUU)
    for j = 1 : length(UUU_trans{1,1})

        UUU_TRANS{1,i}{1,j} = num2cell(zeros(31,4));
    end

end

% The number of random numbers
H = 1000;
UPD =.789155092592539;
format long
% filling in the time matrix
for i = 1 : length(UUU)
```

```
    for j = 1 : length(UUU_trans{1,1})

        UUU_TRANS{1,i}{1,j}(:,end) = num2cell((1:31)');
        UUU_TRANS{1,i}{1,j}(:,end-1) = num2cell(VVV(j));



    end

end

UUU = num2cell(UUU);
%         UUU = sortrows(UUU,2);


for i = 1 : length(UUU)
    for j = 1 : length(UUU_trans{1,1})
        for k = 1 : length(UUU_TRANS{1,5}{1,1})
            % PART 1
            UUU{i,j} =num2cell((((i)^(0)).*((j)^(0)).*rand(1,31)));
            UUU{i,j} = UUU{i,j}';
            UUU{i,j}(:,2) = num2cell(Thirty_five_days(:,1));
            UUU{i,j} = sortrows(UUU{i,j},1);

            %PART 2
            UUU2{i,j} =num2cell((((i)^(0)).*((j)^(0)).*rand(1,31)));
            UUU2{i,j} = UUU2{i,j}';
            UUU2{i,j}(:,2) = num2cell(Thirty_five_days2(:,1));
            UUU2{i,j} = sortrows(UUU2{i,j},1);
            % PART 3
            UUU3{i,j} =num2cell((((i)^(0)).*((j)^(0)).*rand(1,31)));
            UUU3{i,j} = UUU3{i,j}';
            UUU3{i,j}(:,2) = num2cell(Thirty_five_days3(:,1));
            UUU3{i,j} = sortrows(UUU3{i,j},1);
            % PART 1
            UUU_TRANS{1,i}{1,j}(k,end-3) = UUU{i,j}(k,2);

            % PART 2
            UUU_TRANS{1,i}{1,j}(k,end-2) = UUU2{i,j}(k,2);
            %PART3
            UUU_TRANS{1,i}{1,j}(k,end-1) = UUU3{i,j}(k,2);
            rH{i,j} = num2cell(((i)^(0).*(j)^(0)).*rand(H,1));
            yH{i,j} = rH{i,j}(cell2mat( rH{i,j}) <= UPD);
            gH{i,j} = num2cell(cell2mat( yH{i,j}(1:31)));
            UUU_TRANS{1,i}{1,j}(k,end) = gH{i,j}(k,1);
        end
    end
```

```
end

UPDATED_TIME = UUU_TRANS;


%% Traded time

% generate the vector DD

DDDT = 1:31;

% generate the vector VVV

VVVT = 1:12;

% generate the vector UUU

UUUT = 2013 : -1 : 2006;

% making the full thrity five days vector
% Years
Thirty_five_daysT = [UUUT';UUUT';UUUT';UUUT(1:end-1)'];
%Months
Thirty_five_days2T = [VVVT';VVVT';VVVT(1:7)'];
% Days
Thirty_five_days3T = DDDT';

% The updated time algorithm

% initializing the time matrix
for i = 1 : length(UUUT)

    UUU_transT{i} = num2cell(zeros(1,12));

end

for i = 1 : length(UUUT)
    for j = 1 : length(UUU_transT{1,1})

        UUU_TRANST{1,i}{1,j} = num2cell(zeros(31,4));
    end

end

% The number of random numbers
HT = 1000;
```

```
UPDT =.789155092592539;
format long
% filling in the time matrix
for i = 1 : length(UUUT)
    for j = 1 : length(UUU_transT{1,1})

        UUU_TRANST{1,i}{1,j}(:,end) = num2cell((1:31)');
        UUU_TRANST{1,i}{1,j}(:,end-1) = num2cell(VVVT(j));



    end

end

UUUT = num2cell(UUUT);
%         UUU = sortrows(UUU,2);


for i = 1 : length(UUUT)
    for j = 1 : length(UUU_transT{1,1})
        for k = 1 : length(UUU_TRANST{1,5}{1,1})
            % PART 1
            UUUT{i,j} =num2cell((((i)^(0)).*((j)^(0)).*rand(1,31)));
            UUUT{i,j} = UUUT{i,j}';
            UUUT{i,j}(:,2) = num2cell(Thirty_five_daysT(:,1));
            UUUT{i,j} = sortrows(UUUT{i,j},1);

            %PART 2
            UUU2T{i,j} =num2cell((((i)^(0)).*((j)^(0)).*rand(1,31)));
            UUU2T{i,j} = UUU2T{i,j}';
            UUU2T{i,j}(:,2) = num2cell(Thirty_five_days2T(:,1));
            UUU2T{i,j} = sortrows(UUU2T{i,j},1);
            % PART 3
            UUU3T{i,j} =num2cell((((i)^(0)).*((j)^(0)).*rand(1,31)));
            UUU3T{i,j} = UUU3T{i,j}';
            UUU3T{i,j}(:,2) = num2cell(Thirty_five_days3T(:,1));
            UUU3T{i,j} = sortrows(UUU3T{i,j},1);
            % PART 1
            UUU_TRANST{1,i}{1,j}(k,end-3) = UUUT{i,j}(k,2);

            % PART 2
            UUU_TRANST{1,i}{1,j}(k,end-2) = UUU2T{i,j}(k,2);
            %PART3
            UUU_TRANST{1,i}{1,j}(k,end-1) = UUU3T{i,j}(k,2);
            rHT{i,j} = num2cell(((i)^(0).*(j)^(0)).*rand(HT,1));
            yHT{i,j} = rHT{i,j}(cell2mat( rHT{i,j}) <= UPDT);
            gHT{i,j} = num2cell(cell2mat( yHT{i,j}(1:31)));
            UUU_TRANST{1,i}{1,j}(k,end) = gHT{i,j}(k,1);
        end
    end

end

TRADED_TIME = UUU_TRANST;

%% RULE for correcting the traded time
SIZZZE = size(UUU_TRANS{1,3}{1,2});
for i = 1 : 8
    for j = 1 : length(UUU_transT{1,1})
        for k = 1 : length(UUU_TRANST{1,5}{1,1})


            if  TRADED_TIME{1,i}{1,j}{k,1} >= UPDATED_TIME{1,i}{1,j}{k,1}

                TRADED_TIME{1,i}{1,j}{k,1} = UPDATED_TIME{1,i}{1,j}{k,1};
            end

            if TRADED_TIME{1,i}{1,j}{k,1} >= UPDATED_TIME{1,i}{1,j}{k,1}...
                    && TRADED_TIME{1,i}{1,j}{k,2} >= UPDATED_TIME{1,i}{1,j}{k,2}


                TRADED_TIME{1,i}{1,j}{k,1} = UPDATED_TIME{1,i}{1,j}{k,1};
                TRADED_TIME{1,i}{1,j}{k,2} = UPDATED_TIME{1,i}{1,j}{k,2};
            end



            if TRADED_TIME{1,i}{1,j}{k,1} >= UPDATED_TIME{1,i}{1,j}{k,1}...
                    && TRADED_TIME{1,i}{1,j}{k,2} >= UPDATED_TIME{1,i}{1,j}{k,2}...
                    && TRADED_TIME{1,i}{1,j}{k,3} >= UPDATED_TIME{1,i}{1,j}{k,3}


                TRADED_TIME{1,i}{1,j}{k,1} = UPDATED_TIME{1,i}{1,j}{k,1};
                TRADED_TIME{1,i}{1,j}{k,2} = UPDATED_TIME{1,i}{1,j}{k,2};
                TRADED_TIME{1,i}{1,j}{k,3} = UPDATED_TIME{1,i}{1,j}{k,3};
            end


            if TRADED_TIME{1,i}{1,j}{k,1} >= UPDATED_TIME{1,i}{1,j}{k,1}...
                    && TRADED_TIME{1,i}{1,j}{k,2} >= UPDATED_TIME{1,i}{1,j}{k,2}...
                    && TRADED_TIME{1,i}{1,j}{k,3} >= UPDATED_TIME{1,i}{1,j}{k,3}...
                    && TRADED_TIME{1,i}{1,j}{k,4} >= UPDATED_TIME{1,i}{1,j}{k,4}
                TRADED_TIME{1,i}{1,j}{k,1} = UPDATED_TIME{1,i}{1,j}{k,1};
                TRADED_TIME{1,i}{1,j}{k,2} = UPDATED_TIME{1,i}{1,j}{k,2};
                TRADED_TIME{1,i}{1,j}{k,3} = UPDATED_TIME{1,i}{1,j}{k,3};
                TRADED_TIME{1,i}{1,j}{k,4} = UPDATED_TIME{1,i}{1,j}{k,4};

            end
        end
    end

end

%% The traded time table
% Fill the updated time
for i = 1 : 8

   TABLE_TRADED_TIME{1,i} = vertcat(TRADED_TIME{1,i}{1,1},...
       TRADED_TIME{1,i}{1,2}, TRADED_TIME{1,i}{1,3},...
       TRADED_TIME{1,i}{1,4}, TRADED_TIME{1,i}{1,5},...
       TRADED_TIME{1,i}{1,6}, TRADED_TIME{1,i}{1,7},...
       TRADED_TIME{1,i}{1,8}, TRADED_TIME{1,i}{1,9},...
       TRADED_TIME{1,i}{1,10}, TRADED_TIME{1,i}{1,11},...
       TRADED_TIME{1,i}{1,12});

end

% The final concatenation
FINAL_TABLE_TRADED_TIME = vertcat(TABLE_TRADED_TIME{1,1},...
    TABLE_TRADED_TIME{1,2},TABLE_TRADED_TIME{1,3},...
    TABLE_TRADED_TIME{1,4},TABLE_TRADED_TIME{1,5},...
    TABLE_TRADED_TIME{1,6},TABLE_TRADED_TIME{1,7},...
    TABLE_TRADED_TIME{1,8});


%% The updated time table
% Fill the updated time
for i = 1 : 8

   TABLE_UPDATED_TIME{1,i} = vertcat(UPDATED_TIME{1,i}{1,1},...
       UPDATED_TIME{1,i}{1,2}, UPDATED_TIME{1,i}{1,3},...
       UPDATED_TIME{1,i}{1,4}, UPDATED_TIME{1,i}{1,5},...
       UPDATED_TIME{1,i}{1,6}, UPDATED_TIME{1,i}{1,7},...
       UPDATED_TIME{1,i}{1,8}, UPDATED_TIME{1,i}{1,9},...
       UPDATED_TIME{1,i}{1,10}, UPDATED_TIME{1,i}{1,11},...
       UPDATED_TIME{1,i}{1,12});
```

```
end

% The final concatenation
FINAL_TABLE_UPDATED_TIME = vertcat(TABLE_UPDATED_TIME{1,1},...
    TABLE_UPDATED_TIME{1,2},TABLE_UPDATED_TIME{1,3},...
    TABLE_UPDATED_TIME{1,4},TABLE_UPDATED_TIME{1,5},...
    TABLE_UPDATED_TIME{1,6},TABLE_UPDATED_TIME{1,7},...
    TABLE_UPDATED_TIME{1,8});

% Find the unique traded times, Sort the traded times and assign
% unique trade number to each in ascending order
UNIQ = sortrows(unique(cell2mat(FINAL_TABLE_TRADED_TIME),'rows'),[1 2 3 4]);
UNIQO = sortrows(unique(cell2mat(FINAL_TABLE_TRADED_TIME),'rows'),[1 2 3 4]);

% generate a random number
raNDgen1 = (324434 : 26835144)';
raNDgen2 = rand(length(raNDgen1),1);

% merge the two vectors
raNDgen = [raNDgen1, raNDgen2];
% sort according to the second column
Sort_raNDgen = sortrows(raNDgen,2);

% Cut at 2976 and sort according to the first column
Sort_raNDgen1 = Sort_raNDgen(1: length(FINAL_TABLE_TRADED_TIME), :);
Sort_raNDgen2 = sortrows(Sort_raNDgen1,1);

% assign the computed trade numbers to the corresponding trade times
UNIQO(:,5) = Sort_raNDgen2(:,1);
% size of UNIQO
SASS = size(UNIQO);
% finding the position of the sorted traded times in the original times
for i = 1 : length(FINAL_TABLE_TRADED_TIME)
POS{i,1} = num2cell(find(ismember(cell2mat(FINAL_TABLE_TRADED_TIME(:,1:end)),UNIQO(i,1:4
end


%% THE_FINAL_TRADED_TIME
THE_FINAL_TRADED_TIME = zeros(length(FINAL_TABLE_TRADED_TIME), SASS(2));

for i = 1 : length(FINAL_TABLE_TRADED_TIME)

THE_FINAL_TRADED_TIME(cell2mat(POS{i,1}),:) = UNIQO(i,:);

end

 Headers = OPriskDataSetexposure(1,:);
```

```
MATRIX = zeros(length(FINAL_TABLE_TRADED_TIME),length(Headers));

%% The traded time

% converting time into usal time formats
% there are 24 hours in the a day , to fins the hour
rt = 24.*THE_FINAL_TRADED_TIME(:,end-1);
hh = round(rt);

% the minutes
rr = 60.*abs(rt - hh);

mm = round(rr);

% the seconds
rg = 60.*abs(rr - mm);

ss = round(rg);

% Updated time as a vectors
Vec_tedTime = [THE_FINAL_TRADED_TIME(:,1:end-2), hh, mm, ss];

% converting the date back to string
formatOut = 'yyyy-mm-dd HH:MM:SS PM';
Vec_tradedTimeSTRING = datestr(Vec_tedTime(:, 1:end),formatOut);

%% The updated time
% converting time into usal time formats
% there are 24 hours in the a day , to fins the hour
rt = 24.*cell2mat(FINAL_TABLE_UPDATED_TIME(:,end));
hh = round(rt);

% the minutes
rr = 60.*abs(rt - hh);

mm = round(rr);

% the seconds
rg = 60.*abs(rr - mm);

ss = round(rg);

% Updated time as a vectors
Vec_updatedTime = [cell2mat(FINAL_TABLE_UPDATED_TIME(:,1:end-1)), hh, mm, ss];

% converting the date back to string
```

```
formatOut = 'yyyy-mm-dd HH:MM:SS PM';
Vec_updatedTimeSTRING = datestr(Vec_updatedTime(:, 1:end),formatOut);

%% generate the compatible columns
% capturedBy
UNI_STRINGS = unique(OPriskDataSetexposure(2:end,9));
% TraderID
UNI_STRINGS1 = unique(OPriskDataSetexposure(2:end,11));

for j = 1 : length(UNI_STRINGS)

    CapturedBy{j} = OPriskDataSetexposure(strcmp(OPriskDataSetexposure(:,9),...
        UNI_STRINGS(j))==1,9);
    % percentage proportion
    LEngC(j) = length(CapturedBy{j})./length(OPriskDataSetexposure);
    format long
   N_STR(j) = ceil(LEngC(j).* length(THE_FINAL_TRADED_TIME));

   N_STRRR{j} = num2cell(zeros(N_STR(j),1));

end


for j = 1 : length(UNI_STRINGS)
   N_STRRR{j}(:,1) = (UNI_STRINGS(j,1));
end
%
CAPTUREDBY_TOTAL = vertcat(N_STRRR{1,1},N_STRRR{1,2},...
    N_STRRR{1,3},N_STRRR{1,4},N_STRRR{1,5});

CAPTUREDBY_TOTAL = CAPTUREDBY_TOTAL(1:length(THE_FINAL_TRADED_TIME));
CAPTUREDBY_TOTAL = [CAPTUREDBY_TOTAL, num2cell(rand(length(CAPTUREDBY_TOTAL),1))];
CAPTUREDBY_TOTAL = sortrows(CAPTUREDBY_TOTAL,2);
%%


%% generate the compatible columns
% TraderID
Tra_UNI_STRINGS = unique(OPriskDataSetexposure(2:end,11));
% TraderID
Tra_UNI_STRINGS1 = unique(OPriskDataSetexposure(2:end,11));

for j = 1 : length(Tra_UNI_STRINGS)

    TraderID{j} = OPriskDataSetexposure(strcmp(OPriskDataSetexposure(:,11),...
        Tra_UNI_STRINGS(j))==1,11);
    % percentage proportion
    LEngC(j) = length(TraderID{j})./length(OPriskDataSetexposure);
    format long
   TRAN_STR(j) = ceil(LEngC(j).* length(THE_FINAL_TRADED_TIME));

   TRAN_STRRR{j} = num2cell(zeros(TRAN_STR(j),1));

end


for j = 1 : length(Tra_UNI_STRINGS)
   TRAN_STRRR{j}(:,1) = (Tra_UNI_STRINGS(j,1));
end

TRADERID_TOTAL = vertcat(TRAN_STRRR{1,1},TRAN_STRRR{1,2},...
    TRAN_STRRR{1,3},TRAN_STRRR{1,4},TRAN_STRRR{1,5},...
    TRAN_STRRR{1,6},TRAN_STRRR{1,7});

 TRADERID_TOTAL = TRADERID_TOTAL(1:length(THE_FINAL_TRADED_TIME));
TRADERID_TOTAL = [TRADERID_TOTAL, num2cell(rand(length(TRADERID_TOTAL),1))];
TRADERID_TOTAL = sortrows(TRADERID_TOTAL,2);
%% Business lines

BL1 = [cellstr('BL1') cellstr('BL1') ;...
    cellstr('Credit Derivatives') cellstr('Investment Banking')]';

BL2 = [cellstr('BL2') cellstr('BL2') cellstr('BL2') cellstr('BL2')...
    cellstr('BL2') cellstr('BL2') cellstr('BL2') cellstr('BL2')...
    cellstr('BL2') cellstr('BL2') cellstr('BL2') cellstr('BL2')...
    cellstr('BL2');...
    cellstr('Rates') cellstr('MM') cellstr('Equity')...
    cellstr('Commodities') cellstr('Africa')...
    cellstr('Options') cellstr('Bonds/Repos')...
    cellstr('Forex') cellstr('Prime Services')...
    cellstr('Credit Derivatives') cellstr('Management')...
    cellstr('Group Treasury') cellstr('SND')]';

BL3 = [cellstr('BL3') cellstr('BL3') cellstr('BL3') ;...
    cellstr('Africa') cellstr('MM') cellstr('SND')]';

BL4 = [cellstr('BL4') cellstr('BL4') cellstr('BL4')...
     cellstr('BL4') cellstr('BL4') cellstr('BL4');...
    cellstr('ACBB') cellstr('Credit Derivatives') cellstr('Funding')...
     cellstr('MM') cellstr('Portfolio Management') cellstr('SND')]';

 BL5 = [cellstr('BL5') cellstr('BL5') ;...
    cellstr('Credit Derivatives') cellstr('MM')]';
```

```
BL6 = [cellstr('BL6') cellstr('BL6') ;...
    cellstr('Management') cellstr('Prime Services')]';

BL7 = [cellstr('BL7') cellstr('BL7') ;...
    cellstr('Portfolio Management') cellstr('SND')]';

BL9 = [cellstr('BL9') cellstr('Portfolio Management')];

%% generate the compatible columns
% Business line
BUB_UNI_STRINGS = unique(OPriskDataSetexposure(2:end,22));
%
for j = 1 : length(BUB_UNI_STRINGS)

    Bus{j} = OPriskDataSetexposure(strcmp(OPriskDataSetexposure(:,22),...
        BUB_UNI_STRINGS(j))==1,22);
    % percentage proportion
    LEngB(j) = length(Bus{j})./length(OPriskDataSetexposure);
    format long
    Bu_STR(j) = ceil(LEngB(j).* length(THE_FINAL_TRADED_TIME));

    BUs_STRRR{j} = num2cell(zeros(Bu_STR(j),1));

end


for j = 1 : length(BUB_UNI_STRINGS)
    BUs_STRRR{j}(:,1) = (BUB_UNI_STRINGS(j,1));
end
%
BUS_TOTAL = vertcat(BUs_STRRR{1,1},BUs_STRRR{1,2},...
    BUs_STRRR{1,3},BUs_STRRR{1,4},BUs_STRRR{1,5},...
    BUs_STRRR{1,6}, BUs_STRRR{1,7}, BUs_STRRR{1,8});

BUS_TOTAL = BUS_TOTAL(1:length(THE_FINAL_TRADED_TIME));
BUS_TOTAL = [BUS_TOTAL, num2cell(rand(length(BUS_TOTAL),1))];
BUS_TOTAL = sortrows(BUS_TOTAL,2);

%% BUSINESS LINES

BUSINESS_LINESL = [BL1;BL2;BL3;BL4;BL5;BL6;BL7;BL9];

for j = 1 : length(THE_FINAL_TRADED_TIME)

    BUSINESS_LINES{j} = num2cell((j.^0).*zeros(length(BUSINESS_LINESL),3));
    BUSINESS_LINES{j}(:,3) = num2cell((j.^0).*rand(length(BUSINESS_LINESL),1));
    BUSINESS_LINES{j}(:,1:2) = BUSINESS_LINESL(:,1:2);
    POSITION{j} = find(strcmp(BUSINESS_LINES{j}(:,1),...
        BUS_TOTAL(j,1))==1, 1, 'last' );
    % the desk
    DESK(j,1) = BUSINESS_LINESL(POSITION{j},2);
end
%%
% fill in the matrix
MATRIX(:,1) = (THE_FINAL_TRADED_TIME(:,end));
MATRIX(:,7) = (THE_FINAL_TRADED_TIME(:,end-1));
MATRIX(:,6) = (THE_FINAL_TRADED_TIME(:,end-2));
MATRIX(:,4) = cell2mat(FINAL_TABLE_UPDATED_TIME(:,end));
MATRIX(:,3) = cell2mat(FINAL_TABLE_UPDATED_TIME(:,end-1));


MATRIX = num2cell(MATRIX);

MATRIX(:,8) = DESK(:,1);
MATRIX(:,22) = BUS_TOTAL(:,1);
MATRIX(:,9) = CAPTUREDBY_TOTAL(:,1);
MATRIX(:,11) = TRADERID_TOTAL(:,1);
% fill in the updated time and the traded time
MATRIX(:,2) = cellstr(Vec_updatedTimeSTRING);
MATRIX(:,5) = cellstr(Vec_tradedTimeSTRING);


% % Exporting the results to Excel
% filename = 'Hoohlo.xlsx';
% writetable(cell2table(MATRIX ,...
%     'VariableNames',Headers),...
%     filename,'Sheet',1,'Range','A1')
```

CURRICULUM VITA