

https://dspace.org/

Towards a framework for operational risk management in the banking sector

Hoohlo, Mphekeleli

2022
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/40754
Hoohlo, Mphekeleli. (2022). Towards a framework for operational risk management in the banking sector [PhD thesis, University of the Witwatersrand, Johannesburg]. WireDSpace. https://hdl.handle.net/10539/40754

Downloaded from a DSpace repository. https://dspace.org/

TOWARDS A FRAMEWORK FOR OPERATIONAL RISK MANAGEMENT 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. Major Professor	Thanti Mthanti, Ph.D. Committee Member
Odongo Kodongo, Ph.D. Committee Member	Thabang Mokoaleli-Mokoteli, Ph.D. Committee Member
Christopher Malikane, Ph.D. Committee Member	Paul Alagidede, Ph.D. Panel Member (WBS) and Academic Director

WITS UNIVERSITY
Johannesburg, Gauteng

2022

Copyright © Mphekeleli Hoohlo 2022

All Rights Reserved

ABSTRACT

The objective of this study is three-pronged. One, it investigates the factors that influence capital adequacy as measured by the covariates (exposure, frequency and severity) used in banking operations that accompany firms data-log loss reports. Two, it assesses the differential impact of discretionary (by adding artificial data) and non-discretionary (using real world data) loss disclosure on firms' value-at-risk. R software is used to determine the value-at-risk. GLM and GAMLSS techniques are employed and subsequent tests of significance derive aforementioned influential factors, accompanied by a data augmentation algorithm in Matlab software to determine the differential impact of artificial and real world operational loss disclosures on firms' performance in relation to meeting capital requirements. Three, it challenges firms' risk-neutral assumption inherent in operational risk practice, asserting that; in theory, banking operations are more risk averse. Rattle software is used in a k-means cluster analysis method to determine whether controls compensate for persistent losses due to the firms' natural risk aversion.

The research arrived at estimates on the number of losses and their sizes; whereby exposure positively influences the risk ceded by the bank having "learned" from possible variations in past data, therefore improving operational risk management frameworks by introducing ex ante forward-looking components, whereas the addition of artificial data points by data augmentation circumvents former dilemmas of large and rare events so one can do more "learning", notwithstanding the nature of the data's suspect quality as they are constructs not observations. Nevertheless, the artificial intelligent EBOR framework's performance improves on (Hoohlo (2014)'s applied data scaling and parametization techniques arrived at a proxy of about ZAR3B) former techniques for capital adequacy calculation of OpRisk opening up exploration modeling beyond historical accounts of significance to incorporate forward-looking aspects. Furthermore, checks and balances set up based on operational negligence slow down operational risk losses over time thereby establishing the move of firm risk tolerance levels away from risk neutrality, suggesting that banks are more risk averse.

(274 pages)

DECLARATION

I, Mphekeleli Hoohlo declare that the study towards a framework for operational risk in the banking sector is my own, unaided work. It is submitted in fulfillment of the requirements for the award of the Degree of Doctor of Philosophy in Management at the Wits Business School, University of the Witwatersrand, Johannesburg, South Africa. It has not been submitted before for any degree or examination in this or any other university. I further declare that I have properly acknowledged all sources of information used in this study.

Mphekeleli Hoohlo

October 6, 2022

DEDICATION

To my dearly departed father

SETSOMI GRATITUDE HOOHLO

from

a good guy

ACKNOWLEDGEMENTS

I acknowledge those acknowledgeable individuals and things.

CONTENTS

Abstract ii
Declaration iii
Dedication iv
Acknowledgments v
LIST OF TABLES xiii
LIST OF FIGURES xv
LIST OF ACRONYMS xviii

1 INTRODUCTION **1**

1.1 Purpose of the study 1
1.2 Fundamentals of ORMF's 3
1.2.1 Stylised facts and underpinning foundations 4
1.2.2 Behavioural foundations of OpRisk 8
1.3 Context of the study 9
1.3.1 Why `OpRisk?` 11
1.4 Gap in the literature and contribution to knowledge 12
1.4.1 Gaps 13
1.4.2 Contributions 15
1.5 Problem statement 18
1.5.1 Sub-problem 1 18
1.5.2 Sub-problem 2 19
1.6 Research objective 19
1.6.1 Sub-objectives and benefits 19
1.7 Significance of the study 22
1.8 Organisation of the study 23

2 LITERARY REVIEW AND ROAD MAP **26**

2.1 Introduction 26
2.2 Capital charge assessment framework 27
2.2.1 The Capital Adequacy Accord (Basel I) 28
2.2.2 New Capital Adequacy Accord (Basel II) 28
2.2.3 Basel III 30
2.3 Support for exposure-based operational risk (EBOR) modeling 31
2.3.1 Merton's model to Einemann et.al's EBOR model 32
2.4 Current OpRisk measurement framework (ORMF's) 33
2.4.1 Advanced Measurement Approach (AMA) 33
2.4.2 The standardised measurement approach (SMA) 35
2.5 Argument 36
2.6 Loss Distribution Approach: State-of-the-art way 37

2.6.1 A general class of LDA model characteristics 38
2.6.2 Computing the frequency distribution 39
2.6.3 Computing the severity distribution 41
2.6.4 Computing the combined loss distribution 42
2.6.5 Computing the total economic capital 44
2.7 Dependence Effects 46
2.7.1 Copulae 47
2.8 LDA model shortcomings 48
2.8.1 LDA integration methodologies 49
2.8.2 Benefits and Limitations 50
2.9 Modernising ORMF's: Historical development 51
2.9.1 Looking beyond current Oprisk modeling frameworks 52
2.9.2 Integration of EBOR and LDA models 54
2.9.3 Limitations of the EBOR model 55
2.10 EBOR methodology for capturing forward-looking aspects of ORM . 55
2.10.1 What is exposure? 56
2.11 A new class of EBOR models capturing forward-looking aspects . . . 57
2.12 GLM model specification for count data 58
2.12.1 Frequency model specification 59
2.12.2 Frequency model illustration 62
2.13 GAMLSS model specification for PnL impact data 63
2.13.1 Severity model specification 64
2.13.2 Severity model illustration 65
2.14 Review summary 66
2.15 Conclusion 68
2.15.1 Road map 69

3 CALIBRATION, DATA EXPLORATION AND EXPOSURE ANALYSIS **71**

3.1 Introduction 71
3.2 Notation and coding for generalized linear models 72
3.3 The generalized linear model 73
3.4 Exponential dispersion models 74
3.4.1 Steps in modeling from EDM's to the generalized linear model for count data 75
3.5 The variance function 76
3.6 Links and canonical links 77
3.6.1 Intepretation 78
3.7 Offsets in GLM theory 78
3.8 Offsets as a measure of exposure 81
3.9 Exploratory data analysis 83
3.9.1 The loss data collection exercise (LDCE) 84
3.10 Exposure analysis 87
3.10.1 Characteristics of exposure 88
3.11 Description of the dataset 91

3.11.1 Characteristics of the covariates 91
3.11.2 Characteristics of daily operational activity 93
3.12 Conclusion 101
3.12.1 Exposure response on risk influence 101
3.12.2 Visual exposure response analysis 101

4 AN EXPOSURE-BASED OPRISK ANALYSIS FOR FREQUENCY MODELING **103**

4.1 Introduction 103
4.2 A synthesis of research paradigms, the research design and concepts . 104
4.3 Applicability of EBOR design for capturing forward-looking aspects of ORM 106
4.4 Sub-problem 1i. 108
4.4.1 Sub aim (i): Exposure-based OpRisk GLM models 108
4.5 Derivation of the poisson model from the exponential family of distributions 109
4.6 A poisson regression operational hazard model 113
4.7 Logistic regression and GLM's: Loss frequency, Indicator variables . . 115
4.7.1 Estimation of some poisson regression model 116
4.7.2 Estimation of some binomial regression model 119
4.8 Model selection and multimodel inference 120
4.8.1 Data dredging 120
4.8.2 Discussion on single "best model" vs ensemble 122
4.9 Model performance evaluation 123
4.9.1 Confusion Matrix and Statistics 124
4.9.2 ROC Curve (receiver operating characteristic curve) 126
4.9.3 Discussion on multimodel inference (MuMIn) model performance 126
4.10 Conclusion 129

5 GAMLSS: A RATHER GENERAL AND FLEXIBLE FRAMEWORK FOR SEVERITY MODELING OF OPRISK **131**

5.1 Introduction 131
5.2 The 2013 OpRisk loss financial data 132
5.2.1 Modeling selection for severity of OpRisk losses using GAMLSS 133
5.3 Sub-problem 1ii. 136
5.3.1 Sub aim (ii): Exposure-based OpRisk GAMLSS models 136
5.4 Model selection strategy in GAMLSS framework 137
5.4.1 Component $\mathcal{D}$: Selection of the distribution 138
5.4.2 Component $\mathcal{G}$: selection of the link functions 141
5.4.3 Component $\mathcal{T}$ and Λ: selection of the terms and smoothing parameters in the model 142
5.5 Modeling of the GB2 distribution parameters 142
5.5.1 Step-wise selection of explanatory variables 143

5.5.2 Model selection . . . 144
5.6 Model diagnostics and checking . . . 144
5.6.1 Diagnostic plots . . . 146
5.7 Intepretation of the fitted model . . . 149
5.7.1 Dynamics of μ, σ and τ distribution parameters . . . 149
5.7.2 Pdfs and the fitted GB2 distribution . . . 153
5.8 Conclusion . . . 154

6 DATA AUGMENTATION AND CAPITAL ADEQUACY **157**

6.1 Introduction . . . 157
6.2 Overview of OpRisk generation . . . 157
6.2.1 Other fundamental aspects of OpRisk generation . . . 159
6.3 Research objective . . . 159
6.3.1 Sub-aim (iii): model OpRisk losses depending on augmented covariates . . . 160
6.4 Augmenting data for near misses, pending and realised OpRisk losses 160
6.5 Algorithm for population . . . 162
6.5.1 **Algorithm 5.1**: The dataset augmentation meta-algorithm for building the best amount of data observations required to train. . . . 162
6.6 Deploying the R model . . . 165
6.6.1 Deploying the estimated GLM frequency model . . . 165
6.6.2 Deploying the estimated GAMLSS severity model . . . 166
6.7 Further analysis and intepretation of augmented dataset . . . 167
6.7.1 Event loss (EL) type findings . . . 168
6.7.2 Business line (BL) findings . . . 170
6.8 Capital Calculation . . . 171
6.8.1 Computational analysis . . . 172
6.8.2 OpVaR basics and sampling criteria . . . 173
6.9 Computation of the aggregate OpRisk loss using R OpVaR package . 174
6.9.1 Experimental setup for using R OpVaR package . . . 174
6.9.2 Experimental design and analysis . . . 175
6.9.3 Experimental results . . . 177
6.9.4 Total loss estimation by Monte Carlo simiulation . . . 179
6.10 Conclusion . . . 179

7 THEORETICAL INVESIGATIONS INTO THE QUANTIFICATION OF MODERN ORMF **183**

7.1 Introduction . . . 183
7.2 Analysis and interpretation issues with behavioral finance theory . . . 184
7.2.1 Expected utility theory . . . 185
7.3 A new class of ORMF models approach . . . 187
7.3.1 Prospect theory . . . 188

7.4 Sub-problem 2. 189
7.4.1 Sub aim (iv): Interpretation issues using cluster analysis . . . 190
7.5 Theoretical foundations of risk-neutrality in risk quantification 190
7.5.1 Risk-neutral measure $\mathbb{Q}$ 192
7.6 K-means clustering of OpRisk loss data 193
7.7 Exploratory data analysis 194
7.7.1 Choice of appropriate dataset attributes 195
7.8 The estimation of k-means clustering solutions 196
7.8.1 Determining the number k of clusters present 196
7.8.2 Determining a final clustering solution 197
7.9 Descriptive analytics - intepreting the clusters 199
7.9.1 K-means clustering models for OpRisk loss data 199
7.9.2 Model results using discriminant coordinates 202
7.10 Conclusion 203

8 DISCUSSIONS, CONCLUSION AND FUTURE RECOMMENDATIONS **205**

8.1 Introduction 205
8.2 Holistic profile overview 205
8.2.1 Qualitative profile of the loss distribution data collection exercise (LDCE) 207
8.2.2 Qualitative profile on forward-looking aspects compared to normal models 207
8.2.3 Qualitative profile on risk averse attitudes in banking instead of risk-neutrality 208
8.3 Discussion 209
8.3.1 OpRisk loss representation: Exposure 209
8.3.2 OpRisk loss frequency: Exposure-based OpRisk GLM model . 211
8.3.3 OpRisk loss severity: Exposure-based OpRisk GAMLSS models 212
8.3.4 OpRisk losses modeling and capital calculation depending on augmented covariates 213
8.3.5 OpRisk loss representation on risk tolerance: A k-means clustering algorithm 215
8.4 Limitations of the study 217
8.5 Contributions and Recomendations 218
8.5.1 Conclusion 219

APPENDICES **226**

8.6 Appendix A: R Code and Data Preparation for Chapter 3 234
8.6.1 Data preparation for understanding the raw frequency and raw severity data mentioned in section 3.9 on page 83 of collected internally over the period between 1 January 2013 and 31 March 2013 i.e., Q12013 at an investment bank in SA 234

8.6.2 Data preparation providing the numbers of OpRisk events collected through pre-processing by following the LCDE as described in section 3.11 on 91 over Q12013, limited to the covariates selected to fit into the models 234
8.6.3 The algorithm only accepts numerical data and so categorical data is transformed into numeric. This is done using an approach where each value of a categoric variable is turned into a variable itself. Multi-level categoric variables are recoded by building dummy variables corresponding to each level by the following commands: 235
8.6.4 Scatterplots and Histograms plots of loss severities and frequency counts against selected explanatory variables showing basic summary statistics of intra-day trading activity 236
8.6.5 Characteristics of exposure: Exposure data is used for several of the steps of the process in frequency and severity modelling 236
8.6.6 Histograms for frequency characteristics of daily and monthly operations for daily losses and/or pending/near misses 240
8.6.7 Density plots of overlaid trade proportions of realised losses vs pending losses/near misses 240
8.6.8 Histograms of overlaid trade proportions of realised losses vs pending loss/near misses 241
8.6.9 Mosaic grid plots for the bidimensional distributions by traded `Instrument`, `TraderId` and number 242
8.6.10 Mosaic grid plots for the structure of OpRisk portfolio by traded `TradeStatus`, `TraderId` by the number of realised losses vs pending loss/near misses 243
8.7 Appendix B: R Code for Chapter 4 244
8.7.1 Data partitioning of the pre-processed OpRisk dataset into Training/Validation/Testing proportions, in preparation for machine learning model building treatments. The original dataset is partitioned into three random subsets initiated by a random number sequence with a randomly selected seed. 244
8.8 Models 246
8.8.1 Estimation of some poisson regression models for OpRisk loss frequency distribution: To build the model we pass on to the model building function `glm` i.e., the formula that describes the model to build. 246
8.8.2 Let us now fit a broader GLM to be our global model, still `LossesIndicator` as the dependent variable, while the rest of the predictive variables will be predictor variables. 247
8.8.3 Estimation of some binomial regression modelS: The formula that describes the model to build. 248
8.9 Model selection and multimodel inference: MuMIn 250
8.9.1 "All possible models" are considered by subjectively ad iteratively searching the data for patterns and "significance". . . . 250

8.9.2 We 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. 250
8.9.3 Evaluate model performance on the test dataset 250
8.10 Appendix C: R Code for Chapter 5 259
8.11 Appendix C: R Code for Chapter 5 (GAMLSS) 259
8.11.1 Data exploration of OpRisk loss severity dataset in preparation for GAMLSS machine learning treatment. The raw data and the pre-processed datasets are initiated and analysed. 259
8.11.2 Data partitioning of the pre-processed OpRisk dataset into Training/Validation/Testing proportions, in preparation for machine learning model building treatments. The original dataset is partitioned into three random subsets initiated by a random number sequence with a randomly selected seed. 260
8.11.3 Selection of the distribution. 262
8.11.4 Selection of terms. 262
8.11.5 The estimation of some GAMLSS for OpRisk loss severity distribution: To build the model we pass on to the model building function `gamlss` i.e., the formula that describes the model to build. 263
8.12 Appendix D: R Code for Chapter 6 266
8.12.1 Evaluate Poisson rate 266
8.13 Data augmentation code: Extrapolation simulation in Matlab 266
8.13.1 Deploying the GLM and GAMLSS Models 268
8.13.2 Deploying the estimated GLM frequency model 269
8.13.3 Deploying the estimated GAMLSS severity model 271
8.13.4 Experimental setup for using R OpVaR package 272
8.14 Appendix E: R Code for Chapter 7 274
8.14.1 Results: Rattle program code 274

LIST OF TABLES

Table Page

3.1 Exponential family distributions and their parameters (De Jong, Heller, et al., 2008a) 76

3.2 The generalized linear model link functions with their associated units of intepretation. Note: This list is not exhaustive and there are likely more GLMs that are used within previous research 79

3.3 The contents of the traded transactions of the associated risk correction events. 86

3.4 Table generated by Rattle R script (8.6.3) transforming BL Variable to 'Dummy' 86

3.5 Occurence of realised losses: proportions on desk categories 94

3.6 Summary statistics for all losses as per Instrument type 97

3.7 A contingency table showing the bidimensional distribution of transactions by trader identification vs realised and/or pending losses, conditional on the trade status 100

3.8 A contingency table showing the bidimensional distribution of transactions by trader identification vs realised and/or pending losses, conditional on the trade status 100

6.1 Testing the goodness-of-fit of the distributions given in a sevdist or a freqdist object to the loss data 177

6.2 Matrix of correlations of BL/ET cells 179

6.3 The specified dependence copula 3 = Clayton copula is specified . . . 180

6.4 Combining the Frequency and Severity Distributions. 181

7.1 A table showing Within cluster sum of squares sizes and average.within & average.between statistics 199

7.2 K-means cluster sizes (means and centers) of the OpRisk loss banking data 200

8.1 Add caption 230

8.2 Add caption 231

8.3 Add caption 232

8.4 Add caption 233

8.5 Building 'Dummy' variables corresponding to each level 237

8.6 Building 'Dummy' variables corresponding to each level 238

LIST OF FIGURES

Figure		Page
1.1	Losses suffered from 2011 to 2016 from Oprisk	7
3.1	A simple comparison of the Sigmoidal like features of the fat-tailed, right skewed distribution for the Exposure variable, and first-digit analysis of frequency distribution from the exposure data with the expected distribution according to Benford's Law	90
3.2	Numerical grid display	92
3.3	Two numerical solutions: Histograms showing the distribution of UpdatedTime & UpdatedDay by LossIndicator.	95
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	96
3.5	Desk category by realised losses	97
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 target variable `LossIndicator` showing realised losses vs near misses/pending losses reported.	98
3.7	Portfolio structure by trader, trade status and number of realised losses	99
4.1	Confusion (Error) matrices	125
4.2	Comparison of area under the curve (AUC) for the Poisson, Binomial GLM ROC curves for the Training and Validation datasets	127
4.3	ROC Curves for GLM over validation and testing set	128
5.1	Loss against explanatory variables	134
5.2	Summary of fitted distribution	145
5.3	Summary of fitted models	146

5.4 Normalized quantile residuals from model GB2 148

5.5 Two diagnostic plots: Worm plot showing adequacy in modeling the $GB2$ distribution. 150

5.6 Dynamics of the GB2 distribution parameters of the fitted model . . 152

5.7 simple plots of the additive contribution the `BusinessLineLevel1` term and the `Instrument` term in modelling the variation of exposure (σ), and the kurtosis (τ) 153

5.8 Pdf and fit plots: Plots showing adequacy fit of the $GB2$ distribution model. 155

6.1 Event type root cause analysis for all losses 169

6.2 Event type root cause analysis for realised losses 169

6.3 Histograms of business line root cause analysis for amended, terminated and/or voided deals (a) 11,760 cases of pending, nearly missed and/or realised losses, and (b) 1,345 realised losses of the total 11,760 losses 171

6.4 The 3-Dimensional grid of the BL/ET matrix for 7 event types and 8 business lines 175

6.5 Fitted distribution parameters of the fitted severities model 178

7.1 Graphically displayed correlations by strength and a plot of OpRisk loss severities vs exposure 195

7.2 Finding the optimal number of k groups by Plotting the Within-groups sums of squares (Sum(WithinSS) - Silhouette Statistic(SS):Sum) vs. the number of clusters extracted. the sharp decreases suggest that a three-cluster solution may be a good fit to the data. 198

7.3 visualizing the 4-means clustering solution 202

7.4 visualizing the 4-means clustering solution 203

7.5 Discriminant coordinate plot for the four (4) group clusters of the OpRisk loss banking data 204

8.1 Loss data collection exercise (LDCE) categorisation flowchart 229

8.2 Data dredging 251

8.3 Model averaging components 252

8.4 Poisson GLM Amodel summary statistics 253

8.5 Poisson GLM Amodel summary statistics continued 254

8.6 Confusion Matrices comparison 257

8.7 Comparison of area under the curve (AUC) for the Poisson, Binomial GLM ROC curve for training, validation and testing datasets 258

8.8 Normalized quantile residuals from model GB2 264

8.9 Normalized quantile residuals from models ZAIG & ZAGA 265

LIST OF ACRONYMS

ACT Advanced Credibility Theory
AI artificial intelligence
AIC Akaike information criteria
AMA Advanced Measurement Approach
ANN artificial neural network
ARMA autoregressive moving average
ATS automated trading systems
auc area-under-curve
BCBS Basel Committee on Banking Supervision
BEICF business environment and internal control factors
BIA Basic Indicator Approach
BIS Bank for International Settlements
BL business line
BO back office
CA Cluster analysis
CaR Capital-at-Risk
CARA constant absolute risk aversion
EBOR exposure-based operational risk
EC economic capital
EDM exponential dispersion model
EI exposure indicator
EL expected loss
ESF Expected Shortfall/Conditional VaR
ET event type
EUT expected utility theory
EVT extreme-value theory
FI financial institution
FO front office
FX foreign exchange
GAM generalized additive model
GAMLSS generalized additive models for location scale and shape
GLM generalised linear model.
GLR generalized linear regression. The GLR model reduces to LR provided the error term (random component of the response variable y) has a normal distribution
HFLS high frequency low severity
ILM internal loss multiplier
IMA internal measurement approach
IPO initial public offerings
KRI key risk indicators
LDA loss distribution approach
LDCE loss data collection exercise
LFHS low frequency high severity
LGE loss-given-event

LMS Lambda, Nu and Sigma
LV Las Vegas
MADAM mean and dispersion additive models
MC Monte Carlo
ML machine learning
MO middle office
MOPL middle office profit & loss
OpCaR operational risk capital-at-risk
OpRisk operational risk
OpVaR operational risk value-at-risk
ORMF operational risk management framework
ORX operational riskdata exchange
OVB omitted variables bias
POT Peaks Over Threshold
PT Prospect Theory
RC regulatory capital
ROC receiver operating characteristic
rv random variable
RWA risk weighted assets
SA South Africa
SEC Securities Exchange Commission
SMA Standardised Measurement Approach
SOB start-of-business
SS Silhouette Statistic
StA Standardised Approach
TGSS trade generation and and settlement system
UL unexpected loss
UOM unit of measurement
VaR value-at-risk

CHAPTER 1
INTRODUCTION

Purpose of the study

This study identifies two banking industry specific problems in the measurement and management of operational risk (OpRisk) in recognition of it's increasingly destructive nature; now even more pronounced in the coronavirus (Covid-19) pandemic related context, and builds onto existing theories and methods in attempts to relinquish potentially infectious organisational deficiencies that follow from it: Moreso, in an era where OpRisk has earned its status as an existential threat to the stability of financial institutions (FI) and modern day commerce. The researcher tackles the first problem case through a series of supervised machine learning (ML) steps towards a better framework for OpRisk in the banking sector by applying a generalised linear model (GLM) classification problem, and generalized additive models for location scale and shape (GAMLSS) regression problem suitable for exposure-based operational risk (EBOR) treatments. EBOR models incorporate forward-looking aspects in the inherently backward-looking character of models based on historical loss data, effectively replacing historical loss severity curves obtained from historical loss counts by forward-looking measures using event frequencies based on actual OpRisk exposures (Einemann, Fritscher, & Kalkbrener, 2018).

Forward-looking mechanisms are concerned with the planning for future OpRisk levels based on judgments on the subject-matter, in adapting to uploaded information from to point-in-time data changes, such as portfolio restructuring or changes in the business mix in real-time (e.g. divestments), and/or incorporating business environment and internal control factors (BEICF's) directly into the operational risk

management framework (ORMF) or capital model. Preliminary work on EBOR models was undertaken by Einemann et al. (2018), who extend to the ORMF requirements by Basel II regulations to fully capture exposures or forward-looking aspects, thereby widening the AMA framework beyond BEICF's, and are therefore better linked to the loss-generating process and control environment.

Secondly, the framework leverages a computationally intensive data augmentation technique for use in a comparability exercise between the afore-mentioned *artificially intelligent* (AI) data-intensive agile methodologies versus *traditional* statistical estimation. In the same vein, the study speaks to the classical understanding of existing ORMF's to date, which are limited to the assumption that FI's are risk-neutral; evidenced in theory underlying the *classical* LDA model, which derives a negative pay-off function through a mathematical framework in the quantification of value-at-risk (VaR) (refer to definition 2.2.2.1): A risk measure used to determine the capital adequacy requirement commonly adopted in the banking industry based on a *risk-neutral* measure $\mathbb{Q}$.

In lieu of the afore-mentioned, the study finally seeks to invalidate the risk-neutral assumption by means of evidence-based discoveries revealed through a clustering algorithm arising naturally in the unknowns of the data by means of a prescribed model, which applies unsupervised learning techniques to counteract the basis to the second research problem, proposing that FI's are more risk-averse. In theory, a risk manager who experiences persistent/excessive losses due to particular OpRisk events in the past, would over-compensate cover for these particular risk types in the present in anticipation of possible losses the future, and this would be the updated, new learnt behaviour expected to show in reduced losses in data for these event types over time. This determination is best understood analysing subtle patterns between data features and trends in the allocated risk capital estimates to determine what is going on.

Fundamentals of ORMF's

Congruent to Cruz (2002), in the current study the researcher alludes to the notion that most banks' estimates for their risk's are divided into credit risk (50%), market risk (15%) and OpRisk (35%). Cruz (2002) postulates 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 that risk, that is the how, manifests in the widely varying degrees in risk sensitivity of a risk practitioner (Hemrit & Arab, 2012) and other organisational leaders, and it therefore encompasses approaches and theories that focus on how they will choose when faced with a decision based on how comfortable they are with the situation and the variables that are present.

Organisation's assign different risk tolerance or risk appetite levels to different types of risk based on many factors, including the risk sensitivity. Risk appetite is a high level determination of how much risk a firm is willing to accept taking into account the risk/return attributes; it is often taken as a forward looking view of risk acceptance. Risk tolerance is a more specific determination of the level of variation a bank is willing to accept around business objectives that is often considered to be the amount of risk a bank is prepared to accept. Through the use of consistent measurement scales the same risk tolerance or interchangeably risk appetite levels should apply regardless of the type of risk or its source.

Definition 1.2.0.1. ***Operational Risk*** *(OpRisk) is defined as:* **"The risk of direct or indirect loss resulting from inadequate or failed internal processes, people and systems or from external events".**

(Securitisation, 2001).

In planning for future OpRisk concerns, subject matter experts question whether it would ever be possible to build a "super framework" that accounts for every single

determinant of loss (Yan, Guszcza, Flynn, & Wu, 2009). Ultimately, no matter how robust of an ORMF framework design in the past, there are extreme situations where supervisors have always had to exercise judgment e.g., where a bank has large legal exposure or in provisioning. These aspects are commonly known as forward-looking and arise through subjective analysis made by seasoned professionals in their expert subject-matter capacities depending on the scenario.

Forward-looking components aim make the OpRisk modeling apparatus risk sensitive by integrating the internal component (historical data) with subjective components, based on the idea that in estimating the parameters of the subjective and historical distributions separately and combining them to construct the overall capital estimate; plus internal losses exclusively would surely limit risk sensitivity relative to an approach that included other sources of information. For example, in the case where a bank has large legal exposure due to pending litigations, potential loss events require an estimate for the exposure and supervisors in the bank typically have an educated view on the probability and range of outflow, which can then be translated into the model parameters event probability and the stochastic loss-given-event ratio. As such, any ORMF must include the use of internal data and relevant external data, scenario analysis and BEICF's.

Stylised facts and underpinning foundations

The major managerial concern for businesses is in their lack of an universally accepted standard to decide on the choice amongst the several OpRisk management/measurement designs in circulation, resulting in unfounded mitigation measures without a solid managerial discretionary basis at worst, and with dire repercussions from a misunderstanding of the extent of their OpRisk exposure. This is indicative of inefficiencies and inadequacies industry-wide, needless to say those of the operational

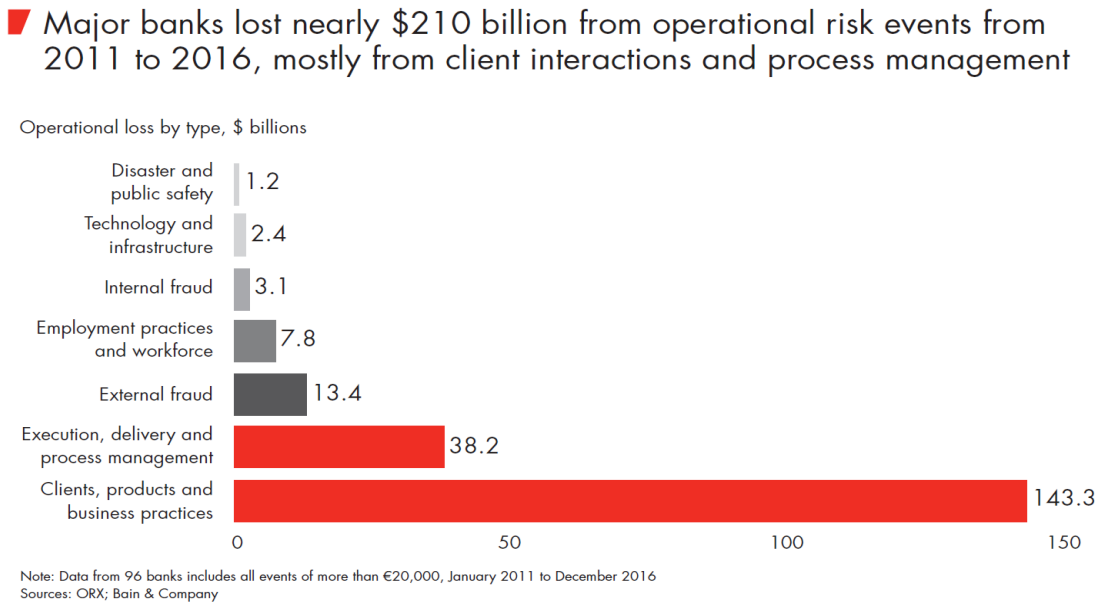
risk function's importance, which tends to look bad overall undermining regulatory authority, thence the endemic inability to successfully account for susceptibility to it, particularly in the banking sector following a number of very costly and highly publicized operational events that have lead to catastrophic losses for the banks in question. OpRisk became popular following a now famed fraudulent trading incident which was responsible for catastrophic losses that lead to the collapse of Barings Bank (the UK's oldest bank) in 1995. The term *OpRisk* began to be used extensively after the afore-mentioned and similar types of OpRisk events became common.

A *rogue* trader, Nick Leeson (Panjer, 2006) risked the banks' survival rather than expose his trading losses by consciously deceiving senior management to hide unethical rogue trading acts, was found responsible for unethical trading practices when he created illegal trades in his account. He then used his position in the front and back offices of the bank to hide his losses. Worse still, he went further in his fraudulent activities incurring greater risks to the bank, by lying in order to give a false impression of his profits. This is in line with the behavioural notion alluded to by Shefrin (2016), that in most risk-bearing decision-making situations people would rather incur greater risks to hold on to their current position and things they already have, than the risks they would have taken to get into that position in the first place.

It was later discovered that Nick was placing illegal bets in the Asian-markets, while covertly keeping these contracts out of sight from senior management to cover up his illegal trading activities (Cruz, Peters, & Shevchenko, 2015). 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 (Martin, 2009). In most, if not all of case involving OpRisk hazard, 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 worldwide that have threatened the stability financial systems due to OpRisk losses. Hefty fines worthy of bankrupting entire corporate entities often have to be imposed on the guilty culprits sometimes resulting in irreparable damage to banks' overall business and reputations, such that widespread regulatory scrutiny has been heightened as a result of a number of scandalous operational events. Kennett & Carrivick (2018) recognise from information offered via the operational riskdata exchange (ORX) global banking loss database that historically, gross loss sizes have been predominantly high and volatile, characterised by a period (from 2011 to 2016) driven by the occurence of large loss events. This coincides with the afore-mentioned post-crisis period, predominantly in 2012, followed by a comparatively stable period in 2017 where industry Oprisk losses saw sizes decrease significantly. Figure 1.1 illustrates comparative distribution of severity of losses per loss event type reported from 2011 to 2016 during which major banks suffered nearly $210 billion worldwide from Oprisk alone.

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 are hidden. For example, the January 2016 "Dark Pool" trading penalties suffered by Barclays Bank PLC amounting to about $70mn and Credit Suisse ($85mn), imposed by the United States (US) based securities exchange commission (SEC). In a case closer to home, Gous (2019) reports ongoing investigations launched in April 2015 for price fixing and "widespread" collusion between banking insiders in South Africa (SA), of the market allocation for foreign exchange (FX) currency pairs viz., USD/ZAR rates, a case which now has been referred to SA based competition tribunal for prosecution, as late as February 2017. Three local banks viz., Absa bank, Standard bank & Investec are implicated in the scandal along with 14 others; some of which have already been fined within juristictions they reside (StaffWriter, 2017), may be liable to payment of an admistrative penalty equal to 10% of their annual turnover.



Source: https://www.bain.com/contentassets/f0199ad9887e402cb37cd1fd316f5ee3/bain_brief_how_banks_can_manage_operational_risk.pdf

Figure 1.1
Histogram showing a breakdown of gross losses focusing on OpRisk loss events in comparison to each other recorded from 2011 up to and including 2016

This investigation led by the local based competition commission uncovered irregularities when rogue traders manipulated the price of the rand through buying and selling US dollars in exchange for the rand at fixed prices between 2007 and 2013. According to the competition commission, currency traders allegedly had been colluding or manipulating the price of the rand through these buy and sell orders to change supply of the currency in contravention of the competition act (Gous, 2019). These acts compromise risk management's advisory service and pedigree, and arouse huge interest as, with the SA incident, distorting the rand value has major implications on the living standards of SA's, felt down to the man in the street. There are many attitudes that can potentially infect organisational processes, the most persistent of these attitudes stem from human failings that are exploitable (Barberis & Thaler, 2003); such as the human conditions' propensity to being deceitful during periods of distress. Furthermore, this kind of behavior sometimes leads to catastrophic operational losses resulting in a mismatch between business' expectations and the value

risk management practice is delivering, which is a prevailing precondition worldwide and remains a major challenge.

Behavioural foundations of OpRisk

In efforts at establishing loss generating mechanisms in the behavioural context, Hemrit & Arab (2012) argue that common and systematic operational errors in those hypothetical situations when put in practice pose presumptive evidence that OpRisk practitioners are almost always unaware of the consequences when loss events are created, assuming they have no reason to disguise their preferences. Consider the case of a risk practitioner in a financial system who assumes that he/she is intent on and accurately executing tasks and analysing an observed subject trusting the validity and relying on visual information that their sense of sight reveals alone. 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.

In fact, throughout his effort the researcher purports OpRisk controls will be advanced herein, which are deemed to set banks apart to the competition as supported by experimental evidence, on the basis of the above-mentioned behavioural traits viz., an OpRisk practitioner (an individual legal entity or a group) is a non-willing agent or cause for OpRisk losses. In most scenarios, information needs to attract investor attention before it can be processed in the organisation and incorporated through methods into prices via trading. Attention, however, is a limited cognitive resource (Kahneman, 1973) posing shortcomings and obscuring OpRisk challenges observed by the risk practitioner, and opening up possibilities for operational errors. The researcher illustrates that behavioural finance theories are compelled to take some of this behaviour into account when trying to explain in the context of a model, how investors maximise a specific utility/value function.

In a theoretical paper, Wiseman & Catanach Jr (1997) discussed several organizational and behavioural theories, such as prospect theory (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 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 obtaining more complete explanations of risk preferences and risk taking behavior (Wiseman & Catanach Jr, 1997).

Wiseman & 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 & Catanach Jr (1997) investigate more comprehensive economic theories viz. PT and the behavioural theory of the firm, that prove relevant to complex organizations who present a more fitting measure for OpRisk. Upon further investigation, Barberis & Thaler (2003) reveal that in finance, behavioral theory explains whether certain financial phenomena can be viewed as the result of less than fully rational thinking. Their argument goes, that through integrating OpRisk management into behavioral theory it may be possible to improve our understanding of firm level RC by refining the resulting OpRisk models to account for these behavioral traits. Thus implying that people's economic preferences described in the model have an economic incentive to improve the OpRisk RC measure.

Context of the study

The main challenge in OpRisk modeling is poor data quality usually comprised of very few data points. This is largely characterised by high frequency low severity

(HFLS) and low frequency high severity (LFHS) data types often culminating toward situations of heteroskedasticity in the data, resulting from combinations of small data sample sizes with wide ranges of event type losses to be modeled (Larsen, 2019), rendering bias in maximum likelihood estimators (m.l.e's). OpRisk's LFHS are risks where the probability of an extreme loss is very small but costly, and HFLS risks are where frequency plays a major role in the OpRisk capital charge calculation. Consequently, this leads to HFLS losses at the lower end of the loss spectrum being ignored by management due to their perceived insignificance and are therefore less likely to be reported, whereas LFHS losses comprise of sensitive company information, hence more often than not are well guarded and kept secret internally with organisations, and therefore are not likely reported.

This poses regulatory concerns, whose reforms are designed and fines imposed around (such) non-compliance 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 due to effects of OpRisk lapses 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 facing South African investment banks and those foreign counterparts who have been found in their countries of origin, and duly compensated for their misdealings, 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, 2014).

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 therefore not able to process all the information in the given time span, they could lose control leading to noncompliance, disincentivising support for regulation, particularly Basel III recovery and resolution processes. In latter days more often than not, OpRisk loss cases reflect lack of sufficient controls driving current OpRisk management catastrophies. The agent on this end of the spectrum of the risk management strategy, which mitigates risk and enforces regulation dependent on visual and information controls is better off, and stands to be improved over an agent on the other extreme, who does not react at all to changes in the system environment.

According to Cruz (2002), FI's wish to measure the impact of operational events upon profit and loss (PnL), these events depict the idea of explaining the *volatility of earnings* due to OpRisk data points which are directly observed and recorded. Due to commonly held beliefs, Aue & Kalkbrener (2006) interpret the main challenges next

generation LDA models are posed with is in their handling of statistical validation of qualitative adjustments to measure OpRisk capital adequaecy, citing incapabilities due to ill-conceived justification for its direct application towards the requirement for RC. In the advent of AI and recent methodological developments viz., ML techniques, able to handle qualitative data types with robustness in the context of validity and reliability, it is believed that the advantages of conducting deeper experimental investigations into the handling of OpRisk's RC estimation outweigh past disadvantages.

This resolves the problem associated with the context dependent nature of OpRisk as an apparent gap in the literature. By seeking to incorporate new data intensive machine learning (ML) approaches to help understand the data, the framework analyses response variables that are decidedly non-normal, including categorical outcomes and discrete counts. Furthermore, such studies go a long way toward shedding light on understanding forward-looking aspects of the business environment and internal control factors (BEICF's), and their effects on firm-level OpRisk RC

Gap in the literature and contribution to knowledge

To meet the conventional caliber required of the PhD report, the research must address a comparative gap in the operational risk literature. A fair amount of work has been conducted over the years around the failures in OpRisk modeling. Most of these studies have produced models with varied shortcomings and inadequacies; from Ganegoda & Evans (2013)'s demonstrative model for severity of OpRisk losses illustrating corrective measures for reporting and scale biases through the use external databases, to the auto-regressive covariate dependent OpRisk modeling attempts through Chavez-Demoulin, Embrechts, & Hofert (2016) and Embrechts, Mizgier, & Chen (2018) extreme-value theory (EVT) models which operate under the traditional premise that past values have an effect on current values, culminating to the recent

EBOR model recommended by Einemann et al. (2018).

EBOR models extend historic severity curves by measures of current exposures as well as using event frequencies based on actual exposures instead of historic loss counts (Einemann et al., 2018). The ultimate task of defining the ideal *exposure* or forward-looking measure for the operational event data is specifically dealt with in this study. In fact, a significant gap in OpRisk literature is in the current lack of a strongly risk-sensitive *exposure* measure (cf. market and credit risk) (Aue & Kalkbrener, 2006).

Gaps

These models in their current form are not suitable for the treatment of the wide spectrum of OpRisk types commonly occurring to forecast OpRisk losses accurately. The Dynamic peaks over threshold-extreme value theory (POT-EVT) based models developed by Chavez-Demoulin et al. (2016) and Embrechts et al. (2018) attempt to address calls from industry for new class of EBOR models that capture the dynamic nature of OpRisk losses, but lack the learning capabilities a future method combining ML technology offer, and are exclusively backward-looking due to their auto-regressive nature comprising of a traditional (statistical) analysis that future values are estimated based on a weighted sum of their covariates, dependent on historical data.

Furthermore, Ganegoda & Evans (2013)'s model cannot be fully trusted due to their exclusive focus on loss severities, dismissing frequency modeling when making unsubstantiated subjective assumptions of what distributional properties fit loss frequencies. The loss frequency is as vital a component as any part of the ORMF framework, and Ganegoda & Evans (2013)'s Poisson assumption with no empirical/methodological basis is misleading at best as it assumes all losses given a loss

event result in total losses, and may not serve as the optimal fit at worst.

These along with other studies singled out specific risk types, such as litigation risk (Einemann et al., 2018), to model OpRisk due to favourable properties inherent in those risk profiles which are designed to suit their model's specification, and sometimes only apply if the data comes in specific packaging. For example, litigation risk profiles are "concentrated" which means they're loss pnl impacts are very likely realised at a somewhat known/predetermined point/period in the future, which is especially interesting, and actually pertinent to a design capturing forward-looking realisations of OpRisk losses in the model. However, since most OpRisk loss risk profiles are not necessarily concentrated and the data forms don't strictly adhere to the their tailor-made requirements, these models regularly fail in practice.

This study recognises Ganegoda & Evans (2013)'s proposal of proportionalities of OpRisk losses to means, variations around means and tail indices of the loss distribution, through the introduction of so called GLM's and GAMLSS methods (John Ashworth Nelder & Wedderburn, 1972). They represent a very general class of regression model which are comprehensive in nature, such that all distributional parameters can be modelled as linear and/or non-linear, parametric,semi-parametric and/or smooth non-parametric functions of covariates. This study contributes to that body of work by extending the methods put forward by Ganegoda & Evans (2013) by the inclusion of the additional estimation of the loss frequency distribution as well as an analysis beyond just a parametric linear GAMLSS models.

Furthermore, Ganegoda & Evans (2013), Chavez-Demoulin et al. (2016) and Embrechts et al. (2018)'s models are limited to publicly available data and therefore suffer from a lack in their accounting for internal controls and risk culture. In comparison, this study uses data points extracted using digital information directly from the internal loss generating mechanisms within the bank, thus created and collected at an individual trade-by-trade, event-wise level, free from potentially erroneous as-

sumptions and manipulations and processed through the loss data collection exercise (LDCE). Additionally, a first in the form of a reproducible data augmentation algorithm consisting of simulated OpRisk losses is implemented by adding more data through a tool that reflects the non-obvious loss-generating mechanisms as well as risk mitigants, circumventing the HFLS & LFHS issue of data scarcity and hence "variance-bias" dilemma associated with increased variance of the resulting m.l.e based estimators, as more data becomes available.

Contributions

Filling in these gaps affirms a meaningful contribution to knowledge of the theoretical, methodological and empirical kind. More precisely there are two (2) theoretical contributions, two (2) empirical contributions, and a methodological contribution summarised as follows:

[1]. In addition to the expansion to both Ganegoda & Evans (2013) and Embrechts et al. (2018)'s proposals, in this study the influence of a moderating variable viz., the *exposure* variable - a monotonic transformation of the mean intepreted through a proportionality constant, otherwise defined as the offset variable, which offers properties eliminating omitted variables bias (OVB), removing some of the constraints regarding needed concentrated risk profiles in the afore-mentioned EBOR model.

[2]. GLM's application of "learning"[1] derives from the multiple ways to reconstruct a given finite number of observations resulting in different scenarios in the future; thanks to a long series of trials allowing the optimization of weights linking

[1]Machine learning (ML) efficiency is a function of the number of data points, it introduces a sensitivity minimising the unscaled deviance (which is trivially equivalent to maximizing the likelihood in the sense that deviances are interpreted as weighted sums of distances of estimated means from observations, see Annex 1 8.2) and hence estimating standard errors.

predictors to targets, Thus making them ideally placed to explore the influence of time lag as defined by the *exposure* i.e., the offset variable d, see Equation 2.35, on the likelihood of the LGE. Offset serves as a function used to measure performance in the multiplicative poisson model over time.

[3]. Ganegoda & Evans (2013), Chavez-Demoulin et al. (2016), Embrechts et al. (2018),and Einemann et al. (2018) affirm that no amount of capital is realistically reliable in the quantification of OpRisk particularly through the modeling approach of capital adequacy, due to weaknesses in current VaR methods which don't have learning defined by the detection and updating of risk tolerances based on changing past behaviours, thereby factoring out misunderstood behavioural risk attitudes such as "loss aversion". Introducing GLM's and GAMLSS techniques advances the recent literature that a better estimation model unrestricted to a "single best" model permitting optimally fitting heavy-tailed distributions without compromising OpVaR model stability i.e., the classical "bias-variance" trade-off (Larsen, 2019), thereby moving away from a traditional "ARMA" based statistical theory which assume risk neutrality to open up another line of research, which suggests a better modeling method for more efficient human behaviour focusing the theoretical lens to give a better understanding of how past losses affect risk attitudes.

[4]. The new framework overcomes the systemic modeling problem consisting of a combination of small sample sizes with the wide spectrum of loss events to be modeled (HFLS and LSHF heteroskedasticity dilemma) and simultaneously eliminates the OpVaR model instability highlighted by Larsen (2019), by applying an unsupervised learning data augmentation methodology. The methodology deploys the properties of GLM and GAMLSS model to create expansive loss forecasts bearing forward-looking properties due to "learning", and OVB which incorporates LGE scenarios rather than dismiss the many alternatives that were not imagined, the possibilities of unaccounted inefficiencies or finding value in unusual places.

[5]. This study challenges potential flaws in LDA models whose induction based derivation i.e., the EBOR model is a special case of the LDA model, assumes that bank's OpRisk appetites obey the principles of expected utility theory (EUT), and are therefore risk neutral. An unsupervised learning Cluster analysis (CA) models what accurately ascribes to decision rules that in making predictions about what OpRisk loss events might result in the future, and is used to establish whether or not; the researchers assertion that the rate of OpRisk's hazard slows, aligning more to prospect theory (PT) i.e., the forecasted future for a specific risk type should be overprovided for in the event that particular loss type has already occurred, and this should show in reduced loss forecast toward the determination of risk adjusted capital to buffer for that risk, which fits normal behavioural patterns around individuals psychological makeup and is consistent with risk aversion.

Conventional OpRisk frameworks in banking commonly exhibit control and system deficiencies whereby information processing is slow and have the tendency to rely on manual, uncertain, unpredictable and unrealistic management and measurement methods, obscuring reporting and resulting in undesirable pre-market conditions. The OpRisk management's function should be able to assist in the ability to mitigate risks through acquiring and/or refining risk management solutions which deliver reliable and consistent benefits of improved operational controls and enhanced risk management frameworks.

Taking combinations and extensions to the modeling frameworks by Ganegoda & Evans (2013), Chavez-Demoulin et al. (2016), Embrechts et al. (2018) and Einemann et al. (2018), this study supports the call for an industry standard in the form of the new EBOR-LDA POT-EVT framework: Building hybrid frameworks toward EBOR, POT-EVT, and LDA integration, we derive a modeling strategy which facilitates the migration of OpRisk subtypes from classical to artificially intelligent (AI) EBOR based treatments which seek to drive capital and risk management such

that capital will decrease in tandem with risk mitigation steps. In this way, the thesis supports the assertion in the prevailing phenomenon of increasing OpRisk hazard growth that tightening OpRisk's management controls should result in OpRisk's hazard slowing.

Problem statement

In our prevailing banking phenomena of increasing OpRisks, the problem consists of questioning whether a firm's susceptibility to OpRisk hazard's growth, results in the degree of OpRisk losses slowing due to tightning OpRisk controls and enhancements to OpRisk frameworks. It would be prudent not to declare things are improving if the evidence is not quite firm that this is true. This is essentially a check for situations, from these data, whether there is evidence of the unchecked spread of negligent behaviour leading to operational loss events or not; or on the contrary, those situations other than the unrestricted spread of these "rogue" events consequently driving OpRisk losses i.e., which may requiring a re-thinking our approach to improving OpRisk controls and enhancing OpRisk management frameworks.

Sub-problem 1

The existing models in OpRisk measurement for which historical loss distributions are the best predictors of future losses, who by design lack the learning capabilities attributed to artificial intelligence (AI), hence assume that we do not learn from past losses and therefore suffer from this as a clear limitation. This is problematic for "predictable" risk types due to their model's practice of undercapitalising known risks before occur, and overcapitalising for risks after the losses materialise, creating inappropriate capital estimates (Group et al., 2013).

Sub-problem 2

Furthermore, we challenge the weaknesses in current OpRisk theory which assumes banks are risk neutral (as evidenced in chapters 1 and 7, sections 1.1 & 7.5 respectively), asserting they are more risk averse. Consequently, "predictive" future losses can be determined who's estimated RC adapts to changes in the risk profile of the bank i.e., with the introduction of new products or in changes to the business mix of the portfolio (e.g. mergers and aquisitions, trade terminations, allocations or disinvestments), providing sufficient incentives for OpRisk management to mitigate risk.

Research objective

1. To circumvent the known HFLS & LFHS problem that comes from a combination of small sample sizes with a wide range of loss types, and associated difficulty of "upward bias" from estimating errors due to m.l.e fitting of heavy-tailed distributions, along with the resulting OpVaR model instability to ultimately realise the move toward an EBOR framework for OpRisk in the banking industry, balancing out both the existence of a heavy-tailed distribution fit on the one hand, and the stability, or variance, of the model derived from a small-data environment as it is applied to new OpRisk loss data on the other.

Sub-objectives and benefits

The afore-mentioned problematic concerns along with evidence showing up in overcapitalised capital buffers losses after they had occurred as claimed in past ex-

periences, which result from OpVaR model instabilities are required to resolve the articulated problem/s of the study, particularly with reference to sub-problem 1.5.1, to motivate all efforts in the development of:

- GLM EBOR modeling framework which not only captures past losses but also how exposures to forward-looking aspects affect risk attitudes.
- GAMLSS models accounting for variations around the mean of severities in order to account for heteroskedacity; notwithstanding an inappropriate assumption of a constant scale parameter for different ET/BL categories of OpRisk that do not capture tail behaviour properly.

Sub aim (i): Exposure-based OpRisk GLM models

To quantify OpRisk loss frequencies by introducing GLM modeling toward a new EBOR framework for OpRisk management, who are "predictive" due to learning capabilities, capturing exposures to forward-looking aspects in the OpRisk loss prediction problem of how past losses affect may risk attitudes. EBOR treatments effectively look into deep hierarchies in the features of covariates in the investment banking risk forecast business problem for frequencies, in the form of a transformation matrix, $e^{\beta_0 \mathbf{X}^T}$ and which "act on" and initial condition[2] to give a by forward-looking measure as a solution using event frequencies based on actual operational risk (OpRisk) exposures in the business environment and internal control risk factors (BEICF) thereof.

Sub aim (ii): Exposure-based OpRisk GAMLSS models

To quantify OpRisk loss severities by introducing forward-looking GAMLSS modeling toward a new EBOR framework for OpRisk management, who are "predictive" due to learning capabilities, effectively replacing historical loss severity curves

[2]Defined as the exposure variable

obtained from historical loss counts, by models accounting for variations around the mean severities; notwithstanding how inappropriate assumptions of a constant scale parameter (variance) are due to the wide spectrum of loss events to be modeled, which creates heterogeniety in the different ET/BL unit of measurement (UOM) categories of OpRisk, and therefore does not capture tail behaviour properly, especially for those more flexible options for distribution fits. Ideally, a heavy tailed distribution with appropriate goodness of fit is desired, which reveals "odds against" knowing if one UOM's losses are more or less variable than another.

Sub-aim (iii): model OpRisk losses depending on augmented covariates

To deploy GLM and GAMLSS models wherein approximate weights of input features consisting of a comprehensive list of user defined input variables (the risk factors) modeled through heuristics derived in the underlying value-adding processs, with associated root causes, are used to estimate OpRisk loss frequencies and severities (both target variables) sequentially. Moreover, these target variables contribute to the frequency and severity of OpRisk losses borne out through an augmentation algorithm forecasting a new OpRisk loss dataset . As a consequence, the framework overcomes the systematic OpVaR model instability and small-size data difficulty, replacing capital esitmate's "upward bias" of overcapitalizing realised losses after they occur and therefore 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.

Sub-aim (iv): Interpretation Issues using cluster analysis

To identify potential flaws in the mathematical framework for the loss distribution approach (LDA) model of ORM, whose derivation of OpRisk losses is 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. The wish is 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 making 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.

Significance of the study

This study fills a gap in that advancing OpRisk 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 et al., 2018). The current *traditional* approach consists of a 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 et al. (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 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. 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

Chapter 1 begins with the purpose of the study, followed by an overview or background of OpRisk theory and its importance in banking. The EBOR model concept is introduced and arguments presented around topical debates in OpRisk LDA modeling, further contexualised by the gap in literature and the contribution the study makes. It provides the statement of the research problems and, research question and objectives to guide the work, and concludes by an account of significance.

Chapter 2 gives an overview of theoretical foundations of OpRisk, followed by a review of the LDA model used to calculate an estimate of the OpVaR measure. The three most high level risk factors are two target variables loss frequency and loss severity, and exposure operating as a moderating variable. A further breakdown analysing the EBOR models approach follows promising to remediate some of the LDA shortcomings and on how EBOR components presented in the study offer a novel approach in contrast to the current literature demonstrating its value and opening up the gap in finer detail.

The general literature review is succeeded by four interlinked chapters, the purpose of each besides acting as successive building blocks, is to provide evidence-based findings in comparison with existing theories and literature, focusing on four (4) propo-

sitions and two (2) hypothesis statements each contributing to resolve specific problems and questions going on in line with the OpRisk literature. In these chapters the application of machine learning techniques on the observed data take centre-stage demonstrating how issues in OpRisk capital requirement estimation are more effectively resolved.

Chapter 3 introduces the GLM and GAMLSS concepts, and the methodological approach to be followed detailing an analysis of the exposure variable in the GLM context. It further describes the data exploration and data screening approaches, looking into addressing the exploratory variable validity from the dataset, presenting Proposition 3.7.1 and Hypothesis 3.9.1, before moving on to data analysis and exposure variable analysis. Chapter 3 concludes with a lead up to Chapter 4 which proposes; with the exposure as an offset variable, a new GLM research methodology in which a combination of ML techniques and statistical theory underlying ORMF's would benefit measurement of capital requirements for OpVaR.

Chapter 5 details the loss frequency concept and design of the new ORMF with the application of EBOR techniques using the GLM's in more detail, and the empirical determinants due to the different components of OpRisk measurement under the new class of EBOR models. It provides Sub-problem 4.4, sub-aim 4.4.1, and Proposition 4.6.1 that guide this chapter. The latter part of Chapter 5 presents the results of the GLM EBOR OpRisk model their analogies to the loss frequency aspect of the OpRisk loss system as a whole. Moreover the results from the GLM regression are intepreted and presented w.r.t the proposition tested in Proposition 4.6.1.

Chapter 5 details the loss severity concept and design of the new ORMF to the portfolio of a wide range OpRisk losses found in the dataset and their integration into an LDA framework through an application of GAMLSS EBOR techniques. Due to GAMLSS's very general class characteristics, emphasis is illustrated around the model selection strategies and step-wise selection exercises that make up the different

components of OpRisk measurement under this new class of OpRisk model. It provides Sub-problem 5.3, sub-aim 5.3.1, and Proposition 5.4.1 that guide this chapter. The latter part of chapter 5 presents the results of the GAMLSS EBOR OpRisk model and their analogies to the loss severity aspect of the OpRisk loss system as a whole. Moreover the results from the GAMLSS regression are intepreted and presented w.r.t the proposition tested in Proposition 5.4.1.

In Chapter 6 the framework's final design takes form: An amalgamation of the previous GLM and GAMLSS components is brought together through intensive algorithm, basing OpRisk generation partly on the heuristics of it's causes and sources, developed in the form of a dataset augmentation technique, as laid out in the body of the chapter, and partly via model deployment, forecasts for both loss frequency & severity target variables are presented. Furthermore, a growth rate is incorporated in the augmentation technique which is also detailed. Chapter 6 provides Research objective 6.3, sub-aim 6.3.1, research question 6.5.1, and 6.7.1 that guide this chapter. The latter part of Chapter 5 presents the computation of total CaR and the results obtained by the OpVaR system as a whole are presented.

Chapter 7 recognises the dual rooted nature of Oprisk theory, nestled in the fields of finance and behavioural science by taking an interdisciplinary perspective on the key concept of risk attitudes, specifically in relation risk neutrality vs "loss aversion". After arguing that banks are more risk averse, the latter part of chapter 7 goes about illustrating the afore-mentioned principles as suggested in Sub-problem 7.4, by providing evidence in the form of an unsupervised learning algorithm viz., cluster analysis (CA), and presents results and arguments in support of Proposition 7.6.1

Chapter 8 recaps and discusses findings in chapters 3 through 7 in comparison with existing theories. In this concluding chapter, final contributions to knowledge are provided along with recommendations and suggestions for future research.

CHAPTER 2
LITERARY REVIEW AND ROAD MAP

Introduction

Acharyya (2012)'s look into literary sources for OpRisk theory, makes the assertion that there is insufficient academic literature that looks to characterize its theoretical roots, as it is a relatively new discipline, by choosing instead to focus on proposing a solution to the quantification of OpRisk. 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 & 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. As such, this chapter not only provides a discussion on the non-triviality of the topic of OpRisk assessment, but also attempts to understand its properties in the context of risk theory, and in the thinking of practitioners and academics in this field.

Capital charge assessment framework

This section justifies the basis for an urgent intervention rooted in supervisory and/or regulatory foundations. 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. In 2001, the Risk (2001) consultative package provided an overview of the proposed framework for OpRisk regulatory capital (RC) charge for OpRisk upon the realisation of financial institutions' (FIs) OpRisk component, which constitutes a substantial risk component to other risks. Consequently in 2006, BCBS adopted a common definition for OpRisk RC, saying;

Definition 2.2.0.1. At the discretion of their national authority, banks ***regulatory capital (RC)*** *charge for OpRisk is:* **The minimum capital charge permissible for the sole purpose of meeting a proportion of the capital requirements for operational risks.**

(Banking Supervision (BCBS), 2006).

There are to date 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.

New Capital Adequacy Accord (Basel II)

The framework for Basel II was implemented in June 2006. The rationale for Basel II is to introduce 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 whereby additional capital 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 (Hoohlo, 2014).

Regarding the sequence Basel I & Basel II: Regulation begins as a qualitative recommendation which requires banks to have an assets-to-capital multiple of at least

20, then auditors' discretion is replaced by market perception of capital which focuses on ratios in which both on-balance sheet and off-balance sheet items are used to calculate the bank's total RWA, meaning there is a market risk capital charge: Risk-based capital i.e., capital-to-RWA of 8%, and 4% with core capital in Basel I and II for all items in the trading business line; then lastly on tail risk. In other words, exciting new static risk management approaches are introduced which involve calculating an $\alpha = 99.9\%$ percentile left tail confidence interval to measure operational risk value-at-risk (VaR) and convert it into a RC charge.

Definition 2.2.2.1. ***Value-at-risk (VaR)*** *is defined as:* **The predicted worst-case loss at a specific confidence level, α over a certain period of time, t.**

(Laubsch & Ulmer, 1999).

Mathematically it can be defined as

$$\begin{aligned} \mathrm{VaR}_\alpha(Z) &= \inf\{z \in \mathbb{R} \| \mathbb{P}(Z \leq z) \geq \alpha\}, \\ &= \inf\{z \in \mathbb{R} \| \mathrm{F}_Z(z) \geq \alpha\} \end{aligned} \tag{2.1}$$

Where $\mathrm{F}_Z(z) = \mathbb{P}(Z \leq z)$ is the cumulative distribution function (c.d.f), or simply distribution function (d.f) associated with the random variable (rv) i.e., the loss distribution. In the continuous case, this reduces to

$$\mathrm{VaR}_\alpha(Z) = \mathrm{F}_Z^{-1}(\alpha) \tag{2.2}$$

Also, it is probably better to use the Expected Shortfall (ESF)/Conditional VaR instead of the VaR per se as this is the new VaR measure suggested by Basel i.e., the allocation of risk capital to BL/ET combination risk cells is based on ESF contributions; which is numerically evaluated in the tail of the aggregate loss distribution if

a sample list of the aggregate distribution X has been calculated at the group level. The tail focused allocation technique is particularly well suited to model risk capital allocations to individual exposures, due to the fact that that each sample x of X_k, where X_k denotes the loss distribution in cell k, can be reduced down in granularity of the m loss events .i.e., $\sum_{k=1}^{m} x_k$, defined as its contribution to the tail of the aggregate loss distribution (Aue & Kalkbrener, 2006): more precisely,

$$\mathrm{ESF}_i(\alpha) = \mathbb{E}[X_k \| X > \mathrm{VaR}_\alpha(X)] \tag{2.3}$$

And α has been chosen such that:

$$\mathrm{ESF}(\alpha) = \mathbb{E}[X \| X > \mathrm{VaR}_\alpha(X)] = \mathrm{VaR}_0.9998(X) \tag{2.4}$$

In particular, standard risk measures like VaR and ESF are based on the Monte Carlo (MC) simulation of the loss distribution which is a numerical representation of a simple closed form of the total event distribution function. Also this measure is better than the classical VaR, notwithstanding and henceforth, reference to the wording VaR will be made to characterize all VaR measures in this study.

Basel III

Basel III establishes tougher capital standards through more restrictive capital definitions, higher risk weighted assets[1] (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 et al., 2010): 1] A future of more capital through incremen-

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

tal 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.

Support for exposure-based operational risk (EBOR) modeling

Many times losses of operational nature are mistakenly attributed to credit defaults or market risk related movement. Merton (1974) structural model is one credit risk model amongst other approaches to model credit risk e.g., the reduced form approach, however, a construct for credit risk modelling uses in OpRisk is borne out of the structural model found in Merton (1974), whereby a theory for pricing corporate debt is presented. Merton (1974) postulates the corporate bond's value is dependent on the volatility of the firms value at a given interest rate i.e., the risk structure of interest rates (Rosa, 2012) under possible gains or losses to investors when there is a significant (unanticipated) probability of default.

The transition from Merton (1974) credit risk model to Oprisk is founded in a seminal paper by Rosa (2012), through her illustration of modern risk exposure whereby the afore-mentioned phenomena arises: An exposure-based method is derived using a credit risk structural model to predict a series of operational losses arising from events of operational nature. In the aftermath triggering off the filing of litigations involving initial public offerings (IPO's), but nevertheless related to credit or market risk losses defined by the credit risk or market risk exposures as opposed to standard OpRisk types whose exposure was undefined. This credit risk model adapted to OpRisk defines what is now called the *exposure-based* operation risk (EBOR) model.

Merton's model to Einemann et.al's EBOR model

In an adaptation of the Merton (1974)'s concept to OpRisk as illustrated by Rosa (2012) to OpRisk losses involving these IPO's, an EBOR model emerges out of a method which requires new types of data incorporating "predictive" factors by the use of a combination of statistical modeling and scenario analysis: Thus, allowing for the inclusion of forward-looking aspects of BEICF's in an EBOR model. An early representation of this EBOR model according to Rosa (2012) is expressed as a common credit risk model given as:

$$EL = p \cdot LGE \tag{2.5}$$

Where EL is the expected loss, p is the probability of the realised loss occuring, and LGE is the loss-given-event. Oprisk events are divided into event types ET_j $j = 1, 2, \ldots$ and business lines BL_i $i = 1, 2, \ldots$.

Other necessary inputs are: the severity exposure E_{ij}, or the maximum possible loss for an event in the BL/ET combination risk cell ij, an exposure indicator (EI) capturing the scale of the bank's activities in the risk cell; and a stochastic estimate L_{ij} specified since only a fraction of the exposure is typically lost, such that $LGE = L \cdot E$. Now, suppose Y_{ij} denotes loss for business line BL_i $i = 1, 2, \ldots$ and event type ET_j $j = 1, 2, \ldots$ for BL/ET combination ij; and p_{ij} denotes the probability that the OpRisk event will occur over the next period $[T, T + \tau]$, then the total loss Y_{ij} is given by:

$$Y_{ij} = p_{ij} \cdot L_{ij} \cdot E_{ij} \qquad \text{where} \quad LGE = L \cdot E \quad \text{by} \quad 2.5 \tag{2.6}$$

Conceptually, this factor based quantification model for capital requirements can be extended to also include future events.

Current OpRisk measurement framework (ORMF's)

Basel II describes three methods of calculating capital charge for OpRisk RC viz., the standardised approach (StA), 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, 2014). In the IMA, the bank uses it's own internal models to calculate OpRisk loss.

The internal measurement approach (IMA) allows banks to use their internally generated risk estimates, Under Basel II the IMA is a first attempt at capital charge calculation for OpRisk: "to directly capture a bank's underlying risk by using the bank's internal loss data as key inputs for capital calculation" (Mori, Harada, et al., 2001). It has similarities to the Basel II model for credit risk, where a loss event is a default in the credit risk jargon.

Advanced Measurement Approach (AMA)

The AMA is an IMA method which applies estimation techniques of OpRisk capital charge derived from a bank's internal risk measurement system (Cruz, 2002). The Basel II & Basel III capital adequacy rules permit model-based calculation methods for capital, including the advanced measurement approach (AMA) for OpRisk capi-

tal. Under the AMA, OpRisk events are divided into seven event types $j = 1, \ldots, 7$ and eight business lines $i = 1, \ldots, 8$ (Risk, 2001) forming a loss decomposition of 56 BL/ET combination sub-risks.

The EBOR model as a special case of the advanced measurement approach (AMA)

Total capital charge is computed as:

$$\mathcal{C}_{OpRisk}^{AMA} = \sum_{i=1}^{8}\sum_{j=1}^{7}\gamma_{ij}\epsilon_{ij} \tag{2.7}$$

$$\begin{aligned} where \quad \epsilon_i &: \quad \text{is the expected loss for business line } i, \text{ risk type } j \\ \text{and} \quad \gamma_{ij} &: \quad \text{is an internal scaling factor} \\ \Leftrightarrow \epsilon_{ij} \equiv Y_{ij} &= p_{ij} \cdot L_{ij} \cdot E_{ij} \quad \text{from} \quad 2.6 \end{aligned} \tag{2.8}$$

The scaling factor γ_{ij} represents a regulatory parameter that is used to transform Y_{ij} into a capital charge. At its most basic level, the EBOR model is a special case of the AMA.

Basel II proposed measurement of OpRisk to define capital requirements against unexpected bank losses: According to a well established theory, the unexpected loss (UL) for the event type j in the business line i corresponds to difference between the quantile for the predetermined confidence level α minus the mean, otherwise stated as expected loss (EL).

The BCBS proposes to define $RC = UL$. This involves simulations based on historical data to establish frequency and severity distributions for losses. In this case the $RC = UL$, is a value-at-risk (VaR) measure.

$$\text{UL} = \inf\{x|\mathbf{G}_{X_k}(x) \geq \alpha\} - \text{EL} \tag{2.9}$$

$$= \text{VaR}_\alpha(X_k) - \text{EL} \tag{2.10}$$

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 measure of RC.

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. 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 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. 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.

The standardised measurement approach (SMA)

The flexibility of internal models was expected to narrow down 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

[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.

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, & Cope (2016) and Peters, Shevchenko, Hassani, & 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 exhibit 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

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

[3]Meaning as of the publication (Risk, 2016) 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

(internal) measurement approach designs, or should the focus of financial regulators shift toward improving on what they see fit within current existing OpRisk 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 in exchange for being able to use a simple formula across the whole industry?

This chapter sought to interrogate this view and others like it, as well as provide an overview of some of the antecedents of OpRisk in light of views of its measurement and management in the banking industry. A topical consequence descendent of regulatory rigour over the years, are new regulatory requirements (Basel III) that remove some forerunning advanced measurement approach (AMA) designs for OpRisk to settle for simpler standardised measurement approaches (SMA), shaking up OpRisk fraternity enthusiasts. However, one may argue for the exact motive for doing this work, that besides regulatory capital requirement, institutions have to compute their true economic capital for their internal capital management and allocation among their business lines and for their optimal value creation.

Loss Distribution Approach: State-of-the-art way

The loss distribution approach (LDA) model is based on actuarial techniques and is generally accepted as the industry (AMA) standard for OpRisk estimation. LDA models require quant level expertise in order for one to accept the statistical relationships linking the actual (perceived) risk exposures. What it has done is to provide the most realistic risk profiles of a company to date (Einemann et al., 2018), based on partitioning OpRisk loss data into sufficiently homgeneous sets, typically corresponding to combinations of business lines (BL) and event types (ET), and to

calibrate a frequency and severity distribution for each BL/ET combination.

LDA models cover risks that are well reflected through historical events and *exposure* data is used in several of the steps of the process in frequency and severity modeling (Cruz et al., 2015). The risk-type cells, reflects Cruz et al. (2015), can be selected at the actual loss generating process level, however most banks use the LDA for BL/ET risk-type cells. The LDA is the most comprehensive modelling approach and is the focus of this study forthwith.

The Loss Distribution Approach (LDA) is an AMA method whose main objective is to provide realistic estimates to calculate RC for OpRisk 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 measurement value for the OpRisk.

A general class of LDA model characteristics

We begin by defining some concepts:

- Consider a matrix consisting of business lines BL and OpRisk event types ET. In line with Basel II, the bank estimates for each business line/event type (BL/ET) combination, the probability function (p.f.) of a single event severity[4] (PnL impact) and the event frequency[5] over a period, say for the next three months. More precisely, according to Frachot, Georges, and Roncalli (2001), in each cell of the BL/ET matrix combinations separate distributions for loss frequency and loss severity are modeled and following an actuarial approach aggregated to produce a loss distribution at the group level.

[4]refers to the PnL impact resulting from the frequency of events

[5]refers to the number of events that occur within the specified time period (daily buckets) T and $T + \tau$

- More precisley, let $N_1, \ldots, N_m$ represent random variables for the loss frequencies i.e., the values $N_k \in \mathbb{N} > \nvDash$ for each $k \in \{1, \ldots, m\}$. Let the individual operational losses $(S_{k1}, \ldots, S_{kN_k})$ denote independent and identically distributed (i.i.d) rv's of the severity distribution S_k, which are independent of N_k, be samples obtained from BL/ET combination types k. The aggregate loss variable X_k in this cell is given as

$$X_k = \sum_{l=1}^{N_k} S_{kl} \tag{2.11}$$

  The aggregate OpRisk loss can be seen as a sum $\mathbf{X} = X_1 + X_2 + \cdots + X_m$; the aggregate loss distribution at group level is then defined by

$$\mathbf{X} = \sum_{k=1}^{m} X_k \tag{2.12}$$

$$= \sum_{k=1}^{m} \sum_{l=1}^{N_k} S_{kl} \tag{2.13}$$

- 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. Daily buckets are chosen in order to ensure data points are sufficient for ML/statistical analysis.

Computing the frequency distribution

- $\mathbf{N}_{ij}$ represents the number of times of OpRisk event failures between times T & $T + \tau$ in BL/ET combination ij, where subscripts i refers to the BL and j to ET. More generally the rv N_k[6] has distribution function (d.f)[7] $\mathbf{P}_k(n/\theta_0)$, where

[6] $N_1 = N_{ij}$, where subscript $i = 1$ & $j = 1$ i.e., N_{11} corresponds to dealing with ET_1 e.g., internal Fraud and BL_1 e.g., Corporate Finance

[7] The term distribution function (d.f) is monotonic increasing function of n which tends to 0 as $n \Longrightarrow -\infty$, and to 1 as $n \Longrightarrow \infty$

θ_0 is an unknown parameter to be estimated in some way, the two best known methods used are the maximum likelihood estimators (m.l.e) or the method of moments.

- The d.f. $\mathbf{P}_k(n/\theta_0)$ for the loss frequency, is defined as the probability that N_k takes a value less than or equal to n, where n is a small sample from the entire population of observed frequencies, i.e.

$$\mathbf{P}_k(n) = Pr\left(N_k \leq n\right) \quad \text{k} = 1, 2, \ldots, \tag{2.14}$$

And a corresponding probability density function (p.d.f)[8]: A d.f term associated with the rv defined by ordinary summations of the N_k's in the discrete case anagolous to the Stieltjes Integrals (of functions relating to the rv N_k and quantile n thereof) defined in the continuous case. The term for p.d.f (called probability function) is the also called the probability mass function (p.m.f) given by the probability that the variable takes the value n, i.e.

$$p_k(n) = Pr\left(N_k = n\right), \quad \text{k} = 1, 2, \ldots, \tag{2.15}$$

- The r.h.s of equation (2.14) is the summation of the r.h.s of equation (7.3), we derive a relation for the d.f associated with given rv's N_k:

$$\mathbf{P}_k(n) = \sum_{l=1}^{N_k} p_k(l) \tag{2.16}$$

Where $N_k \in \mathbb{N}_{>\not\vdash}$ for each $k \in \{1, \ldots, m\}$ between T and some terminal time $T + \tau$.

[8]A non–negative function $p(n)$ 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_k takes discrete values of n with finite probabilities.

Computing the severity distribution

- Suppose S_{kl} is a random variable representing the amount of one loss event in a cell of the BL/ET combination matrix. Define next period's loss in each cell $k = 1, \ldots, m$, where k is the number of BL/ET combinations, S_{kl}^{T+1}. According to equation 2.11, next period's aggregate loss severity X_k^{T+1}

$$X_k^{T+1} = \sum_{l=1}^{N_k^{T+1}} S_{kl}^{T+1} \quad \text{k} = 1, \ldots, m \tag{2.17}$$

  Therefore, it follows by equation 2.12, the LDA model for the amount of the total operational loss in BL_i, $i = 1, \ldots, 8$, loss type ET_j, $j = 1, \ldots, 7$ over a given time T & $T + \tau$, over the future:

$$\begin{aligned} \mathbf{X}^{T+\tau} &= \sum_{k=1}^{m} X_k^{T+1} \\ \mathbf{X}^{T+\tau} &= \sum_{i=1}^{8} \sum_{j=1}^{7} \sum_{l=1}^{N_k^{T+1}} S_{kl}^{T+1} \end{aligned} \tag{2.18}$$

- The rv's S_{ij}^l are i.i.d and independent of N_{ij}. A fixed number in a chosen business line (e.g. Corporate Finance) for a particular loss type (e.g. Internal Fraud) would be denoted by S_{11}^l, representing random samples of the PnL impacts in the $k = 1$ cell in the BL/ET matrix. The loss severity distribution is denoted by $\mathbf{F}_k(x)$ where x is a small sample from the entire population of loss severity S_{kl}. Since loss severity variate S is continuous (i.e. can take on any real value), we define a level of precision h such that the probability of S being within $\pm h$ of a given number x tends to zero. The loss severity, S_k has a (d.f.) $\mathbf{F}_k(x/\theta_1)$, where θ_1 is an unknown parameter to be estimated in some way, the two best

known ways being the m.l.e and/or the method of moments. By definition

$$\mathbf{F}_{S_k}(x) = \int_0^{x_\alpha} f_{S_k}(x)dx \tag{2.19}$$

where $f_{s_k}(x)$ is the probability density function (p.d.f.).

- Therefore by numerical approximation; i.e., the Stieltjes integral definition (Evans, Merran, Hastings, & Peacock, 2001), whereby:

$$\int_0^{x_\alpha} \phi(x) f_{S_k}(x)dx \qquad \text{corresponds to} \quad \sum_0^{x_\alpha} \phi(x) f_{S_k}(x) \tag{2.20}$$

Implying that the severity p.d.f

$$\begin{aligned} \mathbf{F}_{S_k}(x) &= \mathbb{P}[X_k \leq x] \\ \implies &= \mathbb{P}\left(\sum_{l=1}^{N_k} S_{kl} \leq x\right) \end{aligned} \tag{2.21}$$

following by equation 2.11. Once again, the subscript S_{kl} identifies the random variable for severity (PnL impact) of one loss event while the argument x is an arbitrary sample of the loss severities.

Computing the combined loss distribution

- With the frequency choice of $\mathbf{P}_k$ and severity $\mathbf{F}_{S_k}$ loss distributions derived, we define a compound loss distribution $\mathbf{G}_{X_k}(x)$: The distribution of the rv X_k, where the sample is drawn by combining frequency and severity loss distributions following an actuarial approach i.e., from a distribution which is a product a sample from the frequency d.f. (expression 2.16, which is deterministic) the p.d.f (expression 2.21), therefore it is probabilistic. We see a fundamental

relation corroborated by Frachot, Georges, and Roncalli (2001), Cruz (2002), Embrechts, Klüppelberg, and Mikosch (2013), & others:

$$\mathbf{G}_k(x) = \begin{cases} \sum_{n=1}^{\infty} p_k(n)\mathbf{F}_k^{n\star}(x) & \text{x} > 0 \\ p_k(0) & \text{x} = 0 \end{cases} \tag{2.22}$$

where $\star$ is the *convolution* operator on d.f.'s, $\mathbf{F}^{n\star}$, the n-fold convolution of $\mathbf{F}$ with itself. [9] The elements of the convolution function (when $x > 0$) survive; since p kills off the elements of X_n which do not satisfy the condition $X_n < x$, otherwise $G_k(x)$ is 0.

- The compound loss distribution 2.22, $\mathbf{G}_k(x)$ cannot be represented in analytical form; approximations, expansions, and recursions of numerical algorithms used to obtain solutions have been proposed, viz., Monte Carlo simulations, Panjer's recursive approaches, and/or taking the inverse of the characteristic function [Frachot, Georges, & Roncalli, 2001; Aue and Kalkbrener (2006); Panjer (2006); & others]. In the LDA separate distributions of frequency and severity are derived from loss data then combined by Monte Carlo simulation.

- The method consisting of taking a set $\langle \mathbf{X}_1, \ldots, \mathbf{X}_l \rangle$, otherwise known as the ideal generated by elements $\mathbf{X}_1, \ldots, \mathbf{X}_l$ which are l simulated values of the random variable X_k for $s = 1, \ldots, S$ as defined by Fraleigh (2003). It is named in the 1940's after a popular gambling location going by the coined Monte Carlo (MC) term, and owes its popularity to this place due to the gambling background context and their similarities to games of chance. It is a method of estimating the value of an unknown quantity using the a key fact that a random

[9] The convolution of two functions $f(x)$ and $g(x)$ is the function

$$\int_0^x f(t)g(x-t)dt \tag{2.23}$$

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

sample from the population the unknown is drawn from tends to exhibit the same properties as the population, as a principle of inferential statistics. In other words, the way a MC simulation 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.

Computing the total economic capital

The general theory for measuring and allocating risk capital is independent of specific risk types: It is the basis from which standard risk measures are founded. To compute total economic capital (EC) allocation involves developing solutions using common definitions of risk and EC methodologies to handle VaR measures by aggregating multiple OpRisks into a single metric for all VaR measures. Risk capital calculated at the aggregate level forms the basis from which the allocation of risk capital to individual events is derived. The VaR aggregation challenge relates to the incorporation of OpRisk to estimate the amount of capital necessary to absorb potential OpRisk losses associated with VaR measures. Aue & Kalkbrener (2006) show why a consistent framework for measuring risk and firm performance hindges, in fact, on a uniform application of risk theory to capital adequacy calculation for market, credit and Oprisk.

There are generally only two basic approaches for creating this single metric; the second of which is specifically subject to the data disposition: viz, 1] Aggregating across entities by VaR measures and subsequently aggregating these measures into a groups, 2] First aggregate VaR measures by entity and subsequently across entities. To go about these approaches, especially the 2^{nd} more favourable approach; systems used for VaR aggregation must be able to collect and analyse data constantly and

accurately, in addition to having the ability to aggregate and analyse huge volumes of data for assessing VaR measures.

Capital-at-Risk (CaR)

Definition 2.6.5.1. ***Total capital-at-risk (CaR)*** *is defined as:* the predicted worst-case loss for operational risk.

(Frachot, Georges, & Roncalli, 2001).

Most institutions compute CaR as the sum of EL and UL, therefore in using the relationship as defined in equation 2.9 and the obove-mentioned definition for CaR

$$\begin{aligned} \text{CaR} &= \text{UL} + \text{EL} \\ &= \text{G}^{-1}_{X_k}(\alpha) \qquad (2.24) \\ & \qquad (2.25) \\ \text{where,} \quad \text{G}^{-1}_{X_k}(\alpha) &= \inf\{x \in \mathbb{R} \| \mathbb{P}(X_k \leq x) \geq \alpha\} \qquad (2.26) \end{aligned}$$

Loosely, this defines CaR as a VaR measure: i.e., VaR_α, where x is a quantile for the confidence level α minus the mean ($\alpha \in (0,1)$ fixed) of the combined loss distribution $\mathbf{G}$.

The loss data collection exercise (LDCE) doesn't make it possible to always have a collection of all past event losses (Frachot et al., 2001), instead, sometimes when OpRisk losses are generated only summmed up losses are made available i.e., most times the LCDE contains individual OpRisk losses otherwise contains aggregates of losses. In the LDA estimation for the CaR measure for aggregated losses, specifically when dealing with the severity distribution, it's estimation is not a straightforward parametization estimation process which makes use of the popular method of moments, or maximum likelihood estimation (m.l.e). More precisely, The data structure would require analytical expressions for $\mathbf{X}_k$ to be available in order to permit it.

If the underlying portfolio is limited in granularity, risk capital is allocated to a

small number of portfolio constituents whose risk management strategy precludes tail focused allocation techniques like ESF which are based on a high quantile features designed to highlight risk concentrations (Einemann et al., 2018). Other risk measures exist but have the restrictive constraint requiring finite mean, contrary to the combined loss distribution which is generally very heavy tailed. Decidedly desiring alternative techniques gives more weight to the body of the underlying distributions.

Dependence Effects

Dependence matters due to the effect of the addition of risk measures over different risk classes (cells in the BL/ET matrix). Economic capital allocation or capital-at-risk CaR = $G^{-1}_{X_k}(\alpha)$, a VaR measure with an α confidence level for operational risk, which considers dependence between individual cells is benefited in a way that recognises the risk-reducing impact of correlation effects between the individual risks of the BL/ET combinations. The choice of VaR comes with a subadditive property implying the combined VaR measure for two or more BL/ET combinations can be no greater than the sum of their individual VaR measures before they were merged, and OpRisk VaR models are known to exhibit dynamic dependence between loss processes.

More precisely, Hoohlo (2014) explicates that correlations between individual frequency distributions, $\mathbf{N}_{ij}$, and of different types of risk in the particular individual cells of the BL/ET matrix are correlated through a choice of a copula. Due to special dependence effects; i.e., their ability to handle otherwise relatively complicated multivariate distributions through an interesting alternative with no computational difficulties, copulas are normally used and can always be found for the joint model.

Copulae

Copulas are a very powerful statistical tool, and their application for operational risk management has been proposed extensively (Aue & Kalkbrener (2006); Chavez-Demoulin, Embrechts, & Nešlehová (2006); Frachot et al. (2001); Frachot, Moudoulaud, & Roncalli (2004); & Others), since it fulfills one of its main objectives: Conveniently incorporates correlation into a function that combines each of the frequency (marginal) distributions to produce a single bivariate cumulative distribution function (Hoohlo, 2014).

Copulas introduce simplicity in an otherwise relatively complicated multivariate distribution i.e., when dimensions of each frequency in order to replicate observed correlations in the observed data exceeds two. It is sufficient to note that the copula is unique if the marginal distributions are continuous.

In an excerpt from Hoohlo (2014):

> 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:
>
> $$\mathbf{G}(x_1, \ldots, x_m) = \mathbf{C}\left(\mathbf{G_1}(x_1), \ldots, \mathbf{G_m}(x_m)\right) \tag{2.27}$$
>
> 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 2.27 always yields a multivariate distribution with those marginals (p.40-41).

The copula function has since become more familiar to risk practitioners due to its great influence on the aggregation of risk broadening the search for more industrial copula functions such as the Normal, Clayton and/or Student. Our model is used to determine the aggregate distribution of a number of correlated random variables through the use of a Clayton copula.

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 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.

Jongh, De Wet, Raubenheimer, & Venter (2015) and Galloppo & Previati (2014) sought to address the shortcomings of Frachot et al. (2001) by finding possible ways to improve the problems of biases, such as "omitted variable bias" (OVB) and data delimitation in operational risk management. Furthermore, Opdyke (2014) advanced on this problem through a study intending on eliminating biases due to heavy tailed distributions i.e., overestimation of capital adequacy estimates in a time lag after realised losses due to extrapolation to the 99.9th percentile and an overstretched distribution.

Agostini, Talamo, & Vecchione (2010), Galloppo & Previati (2014), Opdyke (2014), Jongh et al. (2015), and others seem to explicate how greater accuracy, preci-

sion and robustness uphold a valid and reliable estimate for OpRisk capital as defined by Basel II/III. Transforming this basic knowledge into a "risk culture" or firm-wide knowledge for the effective management of OpRisk serves as a starting point forming a control function that provides attribution and accounting support within a framework, methodology and theory for understanding OpRisk. 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), Galloppo & Previati (2014), Opdyke (2014) & others follow along lines in the literature of recent attempts aimed at finding a statistical-based model for OpRisk capital calculation, which suggest integrating internal and external data as well as scenario assessments to endeavour on improving on accuracy for the estimates. In recent work, Badescu, Lan, Lin, & Tang (2015) reported that LDA modeling is found wanting due to the very complex characteristics of data sets required to establish OpVaR. Furthermore, continually emerging new found techniques are being built to deal with these issues that arise in LDA modeling; opening new and contentious areas of work and keeping practitioners and academics guessing at what revolutionary phase may follow w.r.t the latest research methods.

LDA integration methodologies

Agostini et al. (2010) also argued that banks should adopt an integrated model by combining a forward-looking component (scenario analysis) to the historical OpVaR, reinforcing foremost discussions in today's literature by involving subject matter expert analysis of the case, through an integration model which is based on the idea of estimating the parameters of the historical and subjective distributions and then combining them using advanced credibility theory (ACT).

The basis for the use of ACT is the idea that a better estimation of the OpRisk measure can be obtained by combining the two sources of information 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 the conventional weighted average approach chosen on the basis of qualitative judgements. Agostini et al. (2010) proposed the integration method be deemed as self-contained and independent of any arbitrary choice in the weights of the historical or subjective components of the model, which serves as a more compelling representation of facts.

Benefits and Limitations

The basic idea of integration methodologies discussed in the afore-mentioned section 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. These approaches are deemed to have significant advantages over conventional LDA methods proposing that an optimal mix of the two modeling components i.e., historical and subjective parts, could better predict OpVaR over traditional methods. Particularly in the work by 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.

These methods have the advantage of being completely self contained and independent of any arbitrary choice or weighting of the historical or subjective components in the model made by the analyst. These components weights are derived objectively, through robust means based on statistical uncertainties of information sources rather

than through risk managers choices based on qualitative motivations. However, they suffer from not explaining the prerequisite need for coherence between the historical and subjective distribution functions, required for the model to work; particularly when Chau (2014) and in a number of papers propose that using mixtures of (heavy tailed) distributions commonly used in the setting of OpRisk capital estimation cannot be avoided (Opdyke, 2014).

Modernising ORMF's: Historical development

The future regulatory environment requires OpRisk professionals who are not only intelligent, creative and motivated but also have the courage to uphold the OpRisk advisory service standards. Businesses that want to successfuly manage OpRisk 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.

Ongoing debates over the design of relevant OpRisk frameworks in the way of trade-offs between unnecessarily complex OpRisk measures who do not provide value towards achieving the intended business outcomes; attributed per force majeure of industry practitioners through their theoretical lens, versus regulators worry to improve simplicity and comparability, wherein OpRisk capital calculation is fully standardised according to a regulatory formula (Migueis, 2017), provide a backdrop to the OpRisk argument. On the one hand there is a element requiring risk sensitivity and on the other comparability across banks and countries.

Looking beyond current Oprisk modeling frameworks

Historical severity curves obtained from historical loss counts that are usually presented in conventional quantification techniques, such as in LDA modelling, have been widely considered to be the most reliable models when used in OpRisk loss estimation. However, they are not useful and have not been very successfull when used to predict future losses, particularly in the measurement of "predictive" OpRisk loss types, which capture forward-looking aspects of the BEICFs thereof. As stated in the industry position paper, see Group et al. (2013), these are OpRisk loss types with defined risk exposure and identifiable risk drivers, which are then incorporated as explanatory variables in "alternative" models whose aim is to replace the aforementioned LDA modelling techniques by measures using event frequencies based on actual exposures and available risk factors, instead of hisorical loss counts in the capital adequacy prediction problem (Einemann et al., 2018).

The aggregate event loss variable of the EBOR model derived from 2.5, with individual losses yields

$$\mathbf{Y} = \sum_{j}^{n} I_r \cdot L_r \cdot EI_r, \quad \text{where} \quad r \in i, \ldots, n \tag{2.28}$$

The EBOR model concept defines n potential loss events, where n is considered as the *frequency exposure* and no longer denotes $n = 56$ different components of the LDA model cells corresponding to BL/ET matrix combinations, but to individual loss events. At this stage, EBOR is not conidered as an option in the Basel II accord for OpRisk. Indeed, the regulatory framework proposes four approaches i.e., SA, BIA, IMA ans the SMA.

The EBOR model is a special case of the IMA conditional on the IMA being

allowed to depend only on aggregate number of events, and on the total loss amounts by BL/ET risk type cells but not on individual losses (Frachot et al., 2001). Frachot et al. (2001) demonstrates this when we find out the conditions under which both methods coincide i.e., $\mathcal{C}_{OpRisk}^{AMA} = \mathcal{C}_{OpRisk}^{LDA}$ and $\mathrm{EL}_k^{AMA} = \mathrm{EL}_l^{LDA}$ i.e., $Y_{ij} = S_{ij}^{T+\tau}$. It comes that the internal scaling factor γ_k:

$$\begin{aligned} \gamma_k &= \frac{\mathcal{C}_{OpRisk}^{AMA}}{\sum_r^n I_r \cdot L_r \cdot EI_r} = \frac{\mathcal{C}_{OpRisk}^{LDA}}{\sum_{l=1}^{N_k} X_{kl}} \\ &= \frac{G^{-1}(\alpha)}{\sum_{l=1}^{N_k} X_{kl}} \end{aligned} \tag{2.29}$$

$$\text{Since,} \quad \mathcal{C}_{OpRisk}^{AMA} = \mathcal{C}_{OpRisk}^{LDA} = G^{-1}(\alpha) \quad \text{it follows therefore}$$

$$\implies \sum_r^n I_r \cdot L_r \cdot EI_r = \sum_{l=1}^{N_k} X_{kl} \tag{2.30}$$

Where n is fixed to the number of observations in the internal database and $\mathbf{G}^{-1}(\alpha)$ is the quantile of ϑ for the level α. For illustrative purposes lets assume the Poisson/Log-normal compound distribution where the frequency parameter ($N \sim \mathcal{P}(\theta)$), and the severity parameters ($\zeta \sim \mathcal{LN}(\mu, \sigma)$): The IMA model can justifiably be developed as a proxy for the LDA model supposedly to capture the LDA model in a simplified way provided γ_k has the following functional form:

$$\gamma = \gamma(\theta, \mu, \sigma; \alpha) \tag{2.31}$$

In turn, see expression 2.7 of subsection 2.4.1; provided expression 2.31 holds, than it is justified that the EBOR model is also a special case of the LDA model (Einemann et al., 2018).

The aggregate event loss $\mathbf{Y}$'s relates to the sum of $(I_1, \ldots, I_n)$ denoting the event indicator; a vector of independent (bernoulli) rv's, whose joint event probabilities are specified through a bernoulli mixture model defined by: $\ni \mathbb{P}(I_j = 1|\Psi = \psi) = p_j(\psi)$

and $\psi = \mathbb{R}^m$, such that they have to attain values $y = (y_1, \ldots, y_n) \in \{0, 1\}^n$; whose sum is called the event **frequency variable**, taking the states 1 or 0 depending on whether there is a realised loss, or a pending loss/near miss. EI_j is the deterministic *severity exposure* of the j_{th} event and L_j is the (stochastic) severity ratio which specifies the loss ratio or loss-given-event (LGE) as a percentage of exposure.

Integration of EBOR and LDA models

The only missing piece for a sound Oprisk capital calculation exercise is left in merging the LDA and EBOR models in a fully integrated and diversified way (Einemann et al., 2018). This setup is achieved by specifying the dependence of the LDA frequency and EBOR frequency through an additional dimension of the copula, such that the EBOR model is considered as and additional cell, anagolously to the BL/ET matrix combinations in classical LDA model. Einemann et al. (2018) deduced a simple recursion formula which is used for a joint LDA and EBOR simulation algorithm smoothening the EBOR modelling integration into an LDA model. The output is a total number of EBOR events, $n_{r+1}(l)$ translated into a joint state of realisations $I_1(l), \ldots, I_n(l)$ for a specific scenario, such that

$$n_{r+1}(l) = \sum_{j=1}^{n} I_j(l) \tag{2.32}$$

The integration concept would also trigger changes of the LDA models input data to avoid double counting of loss potential, therefore it is assumed that the LDA and EBOR events are separated beforehand leaving the task of specifying the model.

Limitations of the EBOR model

In measuring and allocating OpRisk capital (Einemann et al., 2018)'s model 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 underperform for many other OpRisk event types who are not as intricately linked to the risk culture; wherein benchmark Oprisk loss severities in the bank are based (Ganegoda & Evans, 2013), since these EBOR models are typically designed to quantify specific aspects of OpRisk i.e., litigation risk have rather concentrated risk profiles and therefore most likely realised/settled over future coinciding periods. However, it cannot be stressed enough that EBOR models are important due to their wide applicability beyond capital calculation and its potential to evolve into an important tool for auditing process and early detection of potential losses.

EBOR methodology for capturing forward-looking aspects of ORM

In a theoretical paper, Einemann et al. (2018) 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. As mentioned in their paper (Einemann et al., 2018), they were the first to lay the groundwork for future development of their technique across industry, and to establish a common

language through a strategy for integrating EBOR models and LDA models. In the former EBOR model they incorporate "predictable" loss types e.g., they test their hypothesis on a portfolio of pending litigations, litigations being "predictable" as far as when given the event triggering the filing of the litigation case had already happened, and only the final outcome in court has to be modeled. In the latter LDA modeling case, they consider LDA components which cover risks that are well reflected through historical events.

What is exposure?

In the OpRisk context, the total OpRisk loss is captured by certain *exposure* measures, which are quantities that are thought to be roughly proportional to the overall risk associated with an operational event or loss (Parodi, 2014). The measure of exposure needed depends on what loss variable one is attempting at projecting which is dependent on a varied mix of factors. Specifically in relation to the LDA model for OpRisk the exposure measure is dependent on whether we are projecting the loss amounts (severity) from aggregate OpRisk loss events or the number of times (frequency) of OpRisk event failures. When carrying out this decision making expercise the following were considered:

* The availability of historical exposure data over the same period for which the losses are recorded
* The exposure estimate for future periods

In the EBOR ORMF context, exposure is defined as:

Definition 2.10.1.1. 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.

EBOR modeling techniques are specifically designed to quantify specific aspects

of OpRisk consisting of determining the aggregate event loss variable $\mathbf{Y}$ which may have rather concentrated risk profiles, obtained by linking OpRisk events to event types with defined exposures, in addition to "predictive" factors, through the introduction of a given set of *risk factors* who also sufficiently capture risk exposure to forward-looking aspects. As a conscequence, capital estimates adapt to real-time changes in the risk profile of a bank e.g., point-in-time changes in the portfolio mix, or the introduction of a new product. Context herein underlies proposition 3.7.1 referred to in Chapter 3, subsection 4.6.

A new class of EBOR models capturing forward-looking aspects

We are using GLM and GAMLSS methods to build a predictive OpRisk model in the sense that by including all the relevant risk factors that are responsible for PnL loss mechanisms, the model has a "comprehensive" effect, an all encompassing effect on the number and sizes of operational losses, and it's more expansive in it's uses to model a wider range of Oprisk loss types. The model incorporates an in-built offset feature, which is an intuitively significant addition and differentiates this modeling technique to the non-ideal actuarial model specified in Einemann et al. (2018). The *offset* is an additional model variable which is particularly useful in the modeling of growth rate phenomena in the data.

The afore-mentioned non-ideal nature in the actuarial technique for integrating EBOR models and LDA models is compounded by the real-world fact that OpRisk data is often difficult to parse into EBOR data and LDA data types as required by the integrated model; furthermore OpRisk data is often incomplete and many relevant variables are inconsistently coded and massively categorical (Yan et al., 2009). For these reasons set out in Yan et al. (2009)'s accounts, in most actuarial modeling situations modelers are forced to exclude variables that are relevant to predicting

frequency and severity of losses exacerbating the problem of OVB. In contrast, the offset option from GLM's offers it's classical uses of avoiding OVB amongst others, and is useful in predictive modelling.

GLM model specification for count data

In this thesis, we develop a data intensive GLM analysis of the *frequency* response variable viz. the loss ratio term anagolous to Einemann et al. (2018)'s frequency variable, called the LossIndicator; using an explanatory vector of p random variables (rv's) $\Psi = (\psi_1, \ldots, \psi_p)$, the risk factors, which are those casual factors that create losses with random uncertainty[10] and decidedly non-normal, and who introduce dependencies between variables including categorical outcomes and discrete counts; and an *offset* variable d_i which is discussed as a measure of exposure in the context of a Poisson regression. In the loss ratio modeling, Yan et al. (2009) directs the goal at building a model targeting the response $f(y;\theta)$:= the LossIndicator, which is to be layered on to the existing plan.

The *offset* is selected as a measure of trading risk exposure: i.e., the required correction for the period in days, d exposed to risk, and risk factors are the business environment and internal control factors (BEICF's) thereof e.g., information such as the trading times, trader identification, loss event capture personnel, trade status and instrument types, loss event description and reasons for the losses, loss event type categorisation, individual loss amounts, market variables which have an economic interpretation, trading desk and business line, beginning and ending date and time of the event, and settlement times, etc.

[10]Random uncertainties occur When a randomised algorithm returns an answer with slight variations from the "true" value of the quantity being measured. This random error is categoristic to Monte Carlo (MC) algorithms, who consume a random amount of resources (in the form of memory or time), as opposed to Las Vegas (LV) algorithms who always return precisely the correct answer or report their failure

Frequency model specification

As specified in the LDA model (Subsection 2.6.2), let $\mathbf{N}_{ij}$ be the number of times of OpRisk loss event failures over time $[T, T+\tau]$. The stochastic process $N_{ij} \leq n$ is called the frequency process. N_{ij} is equivalent to the r.h.s of equation 2.32, corresponding to Einemann et al. (2018)'s EBOR model *frequency exposure*, where n is the maximun number of events. Recall, the literature shows that Merton's EBOR model is a special case of the IMA, and the IMA is a special case of the LDA, therefore by induction the EBOR is a special case of the LDA hence compatible with LDA modeling, therefore can be seen as special cases of the proposed new EBOR framework, which encapsulates both EBOR And LDA modeling. The unit of exposure n now takes on the value of the upper bound of the rv $\mathbf{N}_{ij}$, the frequency variable in the current LDA model.

$$\mathbf{N}_k = \sum_{k=1}^{m} I_k \tag{2.33}$$

Where n is some terminal time $T+\tau$. John Ashworth Nelder & Wedderburn (1972), Ohlsson & Johansson (2010) and Covrig et al. (2015) show that this process is a Poisson process which follows a Poisson distribution with parameter $\theta = \lambda$, or otherwise the rate. Often the Poisson count λ needs to be described as a rate; More generally:

Definition 2.12.1.1. The **rate** is the mean count per unit exposure e.g., the OpRisk hazard rate can be specified as the rate per day.

i.e.,

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

The rate is specified in terms of units of *exposure*, for example, in OpRisk hazard rates, each potential OpRisk transaction event is "exposed" over the period $[T, T+\tau]$; it's detection life cycle period, and a P&L impact determined, So the rate may be defined in terms of transaction-days *at risk*.

Sequences of Bernoulli trails[11], form building blocks to generalise the Binomial variate $B : n, p$. Furthermore, as the number of trials tends to infinity the Binomial variate tends to the Poisson variate $P : \lambda$. Given that the Poisson variate is a limiting case of the Binomial variate, which in turn is a limiting case of the Bernoulli variate, this commonly positions the Poisson distribution as the favourable first approximation model to the frequency distribution.

Furthermore, the criteria of Evans, Hastings, & Peacock (2001) prescribe the following properties as a guide in choosing between distribution models: Binomial (variance $<$ mean); Negative Binomial (variance $>$ mean); & Poisson (variance $=$ mean). In this study, the intensity (rate) of occurrence of loss events is the fundamental unit of of analysis for estimating the number of loss events (frequency) used for OpRisk losses modeled as a Poisson variable with rate λ as the required limiting distribution. In the estimation of the number of Oprisk loss cases consisting of more failures ($x = 0$) rather than successes ($x = 1$), the likelihood of the Poisson distribution is favoured: Specifically, as it is known to be used more commonly for the the frequency approximation and presents the researcher with the opportunity for a intuitive interpretation of *exposure* as the rate for counting for all available observations over a time lag. Here we describe the *exponential dispersion model* (EDM) of the GLM, which generates the Poisson distribution by the model.

[11]The Bernoulli trial is a classification experiment that has one of two outcomes; success ($x = 1$) or failure ($x = 0$), denoted $B : n$, Where p is the number x of successes for a sample involving n trials, with replacement. Parameter p the Bernoulli probability parameter $0 < p < 1$. Other discrete distributions, such as Binomial and Poisson distributions are a special case of The Bernoulli trial (Evans, Hastings & Peacock, 2001)

$$f(y,\lambda) = \frac{\lambda^y e^{-\lambda}}{y!} \tag{2.34}$$

Modeling counts as realised operational hazard in an OpRisk group requires correction for the period d exposed to risk. The exposure measure is readily incorporated into the estimation procedure and is a quantity that is roughly proportional to the risk (Parodi, 2014). As this statement suggests and is further reinforced by Yan et al. (2009), the offset/exposure measure must be on the same scale as the linear predictor in the basic GLM framework. So the mean frequency will be estimated by the multiplicative model (Covrig et al. (2015) & Ohlsson & Johansson (2010)) corresponding to a logarithmic link function, a *log link*, where a new variable d_i appears

$$\begin{aligned} \lambda_i &= d_i \cdot e^{\beta_0} \cdot e^{\beta_1 x_{i1}} \cdot e^{\beta_2 x_{i2}} \ldots e^{\beta_p x_{ip}} \quad \text{Taking logs on both sides} \\ \ln\lambda_i &= \ln d_i + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \end{aligned} \tag{2.35}$$

$\ln d_i$ is the natural log of risk exposure, called the *offset variable* which affects the algorithm only directly before and after regression estimation, effectively replacing the rate λ with an adjusted rate (counts divided by exposures: $R = \frac{\lambda}{d}$) as the target variable; using exposure as weight; dispensing with the offset (Yan et al., 2009).

Yan et al. (2009) in their definition of exposure, the offset is most commonly discussed as a measure of exposure in the context of Poisson regression. For example when modeling rates in some observations from an OpRisk dataset, for set entries corresponding to a $d_i = 6$ month time lag between the moment the Oprisk event was conceived, T until the OpRisk event is realised at $T + \tau$; while another set of events correspond to a $d_j = 1$ year lag, then it is appropriate to use (log of) months of exposure as and offset (Yan et al., 2009). If not Yan et al. (2009) further deliberates, model variables correlated with months of exposure might possibly pick up some of the variation that should be explained by months of exposure, resulting in biased

parameter estimates.

Frequency model illustration

In their paper, the EBOR model pioneered by Einemann et al. (2018), a non-inflated successful claimed amount provides a plausible estimate for the capital charge for litigation risk, but mainly because it is particularly well-suited to the specific risk type dealt with i.e., due to better usage of extensive existing information (Boettrich & Starykh, 2017) and the more plausible model behavior over the litigation life cycle. This is important, it fits in with the required data availability prerequisite, such as the requirement for case-specific information for each litigation needed for accounting, as well as in the required identification of dependencies across the porfolio which makes it easier for outflow estimates in the provisioning process (Einemann et al., 2018).

Nevertheless, their EBOR model is bound to under-perform for many other OpRisk event types whose data types fall outside of the contraints, since these EBOR models designed for litigation risk are typically designed to quantify specific aspects of OpRisk. Litigation risk have rather concentrated risk profiles i.e., litigations can be grouped into clusters whereby a court ruling for one litigation impacts the likelihood of a payment for other litigations in the same cluster and, on the other hand does not influence the the outcome of litigations outside the cluster (Einemann et al., 2018). For example, in a chart of litigation settlement as a percentage of IPO issuing amounts, see Rosa (2012) which demonstrates the highest concentrations of payments toward the end of Q2 2012 on toward early Q3, tapering off in the early and latter parts of year. Context herein underlies proposition 4.6.1 referred to in Chapter 4, Section 4.6.

GAMLSS model specification for PnL impact data

As a way of addressing the OpRisk growth enigma and on how to go about thinking of VaR and the problems associated with the possibilities in 2.12.2 e.g., of extreme events, we re-examine HFLS/LFHS data severity modelling by introducing the GAMLSS framework. GAMLSS employs the power of modern statistical techniques and methods for performing a univariate regression analysis; more precisely the loss *severity variable*, where not only the location of the severity distribution but also the scale and shape of the distribution can be modelled by explanatory variables (M. D. Stasinopoulos, Rigby, & Bastiani, 2018). Much like how the linear model is a submodel of the GLM, the GLM and Generalized Additive Models (GAM) are submodels of the GAMLSS, therefore extends the features of these model types viz., assumed EDM for the response in GLM and GAM, moreover for the GAMLSS the assumed response distribution can belong to any parametric distribution.

Furthermore, GAMLSS are semi-parametric regression type models in the sense that all the parameters of the response variable distribution can be modelled using parametric and/or non-parametric smooth functions of explanatory variables, thus allowing for a very general and flexible system for modelling OpRisk loss severity variable. Due to this, they are often viewed as "beyond mean regression" (Kneib, 2013) models or distributional regression model (Fahrmeir, Kneib, Lang, & Marx, 2013) approaches. What this means is that the location, scale and shape of the response variable is allowed to change according to explanatory variables, which are expecially useful for OpRisk loss severity data modelling, whose attributes are continuous and positively skew, and/or leptokurtic.

Severity model specification

"The idea behind the semi-parametric characteristic of GAMLSS modelling is to let the data determine the relationship between the predictor $\eta = g(\mu)$[12], and the explanatory variables s_k, rather than enforcing a linear (or polynomial) relationship" (M. D. Stasinopoulos, Rigby, Heller, Voudouris, & De Bastiani, 2017, p. 16). As mentioned earlier, there are instances where the assumption of a constant scale parameter (σ), and shape parameters (ν, τ) is not appropriate such as for some highly skew and kurtotic distributions. On these occations modelling $\sigma, \nu, \&\tau$ as a function of explanatory variables solves the problem (M. D. Stasinopoulos et al., 2017).

A whole new industry of modelling emerges from initially modelling the dispersion parameter beginning with Harvey et al. (1976), followed on by Aitkin (1987), then furthered within the GLM by John A. Nelder & Pregibon (1987), Smyth (1989) & Verbyla (1993) in which the response may follow an expontial family distribution: M. D. Stasinopoulos et al. (2018)'s advanced mean and dispersion additive (MADAM) models is permitting a response variable following any simply known mean variance relationship whose formulation uses either GOV or marginal likelihood (approach to select the degree of penalisation) methods of fitting for the estimation of smooth function of predictor variables.

GAMLSS models allow the predictor to now predicts some $g_k(\cdot)$ for $k = 1, 2, 3, 4$, a smooth monotonic link function relating the distribution parameter θ_k to a linear or non-linear parametric functions e.g., the expected value of the response which may follow a EDM family, or in addition nonparametric smoothing functions h_{jk} of the explanatory variables x_{jk}, to the predictor η_k. Assuming $i = 1, \ldots, n$ independent observations $Y_1, Y_2, \ldots, Y_n$ with p.d.f $f_Y(y_i|\mu, \sigma, \nu, \tau)$ and

[12] η now referred to as the predictor and not linear predictor, since the smoothing terms introduce non-linearities in the model

$$Y_i = D(\mu_i, \sigma_i, \nu_i, \tau_i) \tag{2.36}$$

Where D is any distribution with (up to) four distribution parameters. In general the model has a structure something like

$$g_k(\theta_k) = \eta_k = \mathbf{X}_k\beta_k + \sum_{j=1}^{J_k} h_{kj}(\mathbf{x}_{kj}) \tag{2.37}$$

Where $\mathbf{X}_k$ is a known design matrix, and $(\beta)_k = (\beta_{k1}, \ldots, \beta_{kJ'_k})^{\mathsf{T}}$ is a parameter vector of length J'_k to be estimated and h_{jk} are non-parametric smooth functions of explanatory variables X_{kj} and $\mathbf{x}_{kj}$ are vectors of length n, for $k = 1, 2, 3, 4$ and $j = 1, \ldots, J_k$. The explanatory variables can be similar or different for each of the distributional parameters, which can be modelled as linear i.e., $\mathbf{X}_k\beta_k$ or smooth term functions $h_{kj}(x_{kj})$ for $k = 1, 2, 3, 4$ (M. D. Stasinopoulos et al., 2018).

Severity model illustration

More often than not, the commonly rare and low quality OpRisk data types exhibit lesser to lower concentrated risk profiles, which compounds the OpRisk data dilemnna when it comes to measuring OpRisk, as the less data available the less accurate VaR the estimation is likely to find. This means that for less concentrated risk type profiles, the more data available the more sensitive the deviance becomes which impacts on the variation of exposure in the model output. It's important to note that deviances may be intepreted as weighted sums of distances of estimated means from observations Y_i; Ohlsson & Johansson (2010) indicates that minimising the unscaled deviance plays a pivotal role in severity distribution modelling, as it is equivalent to maximizing the likelihood in the normal linear model and/or minimising of the the

sum of squares in methods used during frequency distribution model estimation.

Moreover, most OpRisk loss profiles are not necessarily concentrated and therefore the EBOR model is limited in it's current form. EBOR models are important nevertheless, due to their wide applicability beyond capital calculation and also due to the potential to evolve into an important tool for auditing process and early detection of potential losses, albiet the basis of moving toward a new EBOR model as more data becomes available. In this new framework the offset variable which eliminates OVB also introduces a sensitivity to the deviance and therefore caters for this variation of exposure. This paper advances toward a framework stringing along these challenges and benefits in mind. The context herein underlies proposition 5.4.1 referred to in Chapter 5, Section 5.4.

Review summary

According to Basel II Framework there are various ways that an AMA model ORMF can be constructed: Supervisors expect that the inputs (to the AMA model) are based on data which records a bank's actual loss experience as well as supervisory inputs to provide a foundation for the banks' scenarios within it's own risk and management profile. The Committee has observed that besides the historical loss component[13], in the absence of internally recorded high severity loss events the consideration of the impact of relevant external losses and/or factors reflecting the business environment and internal control systems. are often useful to inform the tail of the loss distributions for capital charge modeling (Cruz et al., 2015).

With ongoing reviews and refinements to Basel II and Basel III frameworks, banking supervision is looking to benefit banks through exploratory flexibility and the use of different modeling options to produce credible and robust estimates of OpRisk

[13]Internal loss data component that records a bank's actual loss experience

capital charge. As such, various aspects of the contribution to capital charge and the requirement for the use of certain key data features should be met as appropriate to the supervisory soundness standard set out in Committee et al. (2011). In this respect, there is a significant challenge that need to be addressed: Construction techniques of incorporating the different data elements that is highly desirable such that there is no disconnect between the measurement and the management of OpRisk within the bank.

In anticipation of aims to enhance risk sensitivity, ORMF's are tasked with combining the forward-looking components[14] with the historical components in a model based on the idea of estimating the parameters of the subjective and historical distributions separately, then integrating them by blending partial weighted capital requirement estimates representing the banks requisite capital requirements consummate with it's OpRisk profile (Committee et al., 2011). An arbitrarily chosen weighting of the backward-looking component and the forward-looking component is often imposed to represent a reasonable balance between conservatism and forward-looking perspective.

Instead of relying on assumptions, a promising approach in this respect are so-called exposure-based operational risk (EBOR) models which provide a transparent quantitative framework for combining forward-looking assessments of subject-matter experts, point-in-time data (e.g. current portfolios) and historical loss experience which aim at replacing unbounded historic severity curves by measures of current exposures (maximum possible loss) as well as using event frequencies based on actual exposures (maximum number of events) rather than historic loss counts. Regulators would no longer have to be suspicious of scenario analysis, BEICFs, or any other qualitative methods used to forecast exposure. Also, regulators no longer have to challenge

[14]Scenario data and BEICF indicators that provide forward looking views of potential OpRisk exposures and assessments of business risk factors as well as a bank's internal control environment

the specific statistical details of modeling frameworks or entertain the principles of credibility theory and optimal model averaging.

Conclusion

The underlying problem in OpRisk modeling is that capital estimates are based on backward looking decision making criteria which do not update on new information and therefore are slow to react to updated outcomes and changes in risk profiles. This weakens OpRisk management as models based on historical losses have an inherently backward looking character[15]; as opposed to artificially intelligent (machine) learning algorithms which use behavioral analytics to update on new information and therefore better linked to the loss generating mechanisms, and associated behavioral tendencies based on forward looking concepts.

In this study we are moving away from traditional "auto-regressive" theories constrained by statistical models assuming that means are a linear functions of covariates and normally distributed random errors, such as (general) linear regression models, to adopt leading edge (LDA & EBOR) theories with very flexible and unifying (AI) frameworks viz. GLM's and GAMLSS, into OpRisk processes and control systems and superimpose these on the OpRisk model, incorporating theories from behavioral finance to the learning of high dimensional, non-linear OpRisk loss generating mechanisms.

[15]Finance models depicting OpRisk theory describe behavioral phenomena, and therefore uncertainty as measured from situations modeled in the past are the best estimates for outcomes in future risk scenarios and are at best subjective approximations. In the OpRisk context, they are not as accurate as those in the modeling market risk and credit risks, as in the latter contexts the theory closely lends itself to scientific analysis versus the former case, which is further compounded by the fact that in models for OpRisks, accurate and reliable quantitative data is rarely available and often of low quality

Road map

This thesis therefore presents a systematic development of new ways to conceptualize, analyse and predict Operational Risks in Banking, using machine learning and artificial intelligence techniques showing how data intensive techniques can revolutionize entire Bank operational procedures. It is doing this by developing a new data collection methodology of loss data, new tools for data analysis and a new ORMF, while, in the meantime, analysing and rejecting the assumption of risk neutrality of bank and thus adapting risk management framework to a risk adverse banking sector. The contents of the chapter, Chapter 2, is a piece of critical review of the state of the art and its historical development.

Chapter 3 presents the GLM framework and the data used in the analysis. The dataset is an original collection of internal logs of a South African investment Bank for the first three months of 2013, for a total of more that fifty thousand unique trades. The dataset is really impressive and it is described and analysed with completeness. The high frequency characteristic of the data show patterns of operational risks and loss severity that would not be evident with more aggregate measures, where a lot of intrinsic variability is smoothed out. The descriptive analysis is, in itself, a great contribution in understanding the intra-day and intra-month patters of operational losses, which would be very useful to Bank management.

Chapter 4 presents a first model of loss frequency. The model is constructed via a model averaging of all possible combinations of covariates, where the weight in the average is given by an Akaike criteria. Once again, the patterns found in the data are large and different, as is customary in this kind of exercise, and it represents an interesting map of operational risk and its correlation with covariates.

Chapter 5 go further and present a more general model, that take in full con-

sideration the non-linear and multi-dimensional nature of the underlying problem. It is particularly important to give the right focus to large but infrequent losses (the heteroskedasticity in the data) that are the main operational risk to the stability of financial institutions. The current approach gives too much weight to these kind of operational losses and the methodology used in the thesis correct that by using the whole power of "learning" from the data.

Chapter 6 instead look at techniques to increase the amount of information available to perform the analysis. One problem of the analysis is that the short period considered limits the amount of operational losses in each loss categories from which to learn the underlying, non-linear, relationship. Therefore the thesis discuss and present methods to augment the data available by stretching the data to an annual dimension using the statistical properties of the data distribution and Monte Carlo methods. This "syntetic" annual data are than used to re-do the analysis of operational risk in Chapter 4. The results show the robustness of the methodology and show the possibility of using this approach to think about forward risks.

Chapter 7 uses the dataset to analyse the theoretical foundation of rationality and risk neutrality commonly assumed in traditional finance. This chapter gives a very interesting contribution to behavioral finance, showing how one can use the patterns in the data to explore how the risk aversion of a bank changes with past experience of operational losses. This implicit "learning" (explicit in the techniques) show a different interpretation of risk aversion parameter as an accumulation of institutional experience.

CHAPTER 3
CALIBRATION, DATA EXPLORATION AND EXPOSURE ANALYSIS

Introduction

When relationships are explored through a linear model[1], in context *linear* is interchangeable with "linearly combined" signifying a model used to assess and quantify the movements of a univariate response variable y by movements in one or more other explanatory variables x. Subtle conceptual changes are often applied to the response variable, such as in the use of a transformation $g(\mathbf{y})$, or the expected value $E(\mathbf{y})$ or $g\{E(\mathbf{y}|\mathbf{x})\}$[2]; describing the linear model linking a univariate dependent variable with usually several explanatory variables, to better explain the transformed response by the observed explanatory variables in x (De Jong, Heller, et al., 2008). Upon further scrutiny, the term linear actually refers to linearity in the β_j co-efficients of the linear predictor term written as $\mathbf{x_i}^T\beta$, not the x variables which incorporate related terms other than simple straight lines e.g., $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1 x_2$ is in a linear form.

Furthermore, the linear regression (LR) model is a special case of the generalized linear regression (GLR) model. The GLR model reduces to LR provided the error term (random component of the response variable y) has a normal distribution i.e., $\epsilon_i \sim \mathbf{N}(\mathbf{0}, \sigma^\mathbf{2})$ i.e., the normal assumption provides "ideal" least squares estimator properties that appeal to our understanding; which is essential from an intuitive perspective as non-normality violates the conditions of linearity (remembering the word "ordinary" in ordinary least squares (OLS) essentially refers to maximum likelihood estimators (m.l.e) when errors are normal). Many of the ideas and concepts (Dobson

[1] $E(\mathbf{Y_i}) = \mu_\mathbf{i} = \mathbf{x_i}^T\beta + \epsilon_i \qquad \epsilon_i \sim \mathbf{N}(\mathbf{0}, \sigma^\mathbf{2})$

[2] In this case the statistical average of $\mathbf{y}$ is modeled in terms of x, In other words, it is the expectation of that is being explained, and any observation on y will deviate from this statistical average.

& Barnett, 2008) of LR modeling carry over to GLR, however the "generalised" term is used to refer to (all) models in the linear forms.

GLR models are important in the analysis of the OpRisk dataset, in the sense that, the relationship between outcomes and drivers of risk are not simple "additive" case scenarios free from non-ideal complexities, rather they are often multiplicative posing asymptotic results with no definite solutions as one would encounter in real research. In fact, the real observations $y_1, \ldots, y_N$ regarded as realisations of the response variables (random variables) $Y_1, \ldots, Y_N$ often have distributions other than normal; and therefore since the introduction of GLR modeling for count data, OpRisk frequency modeling is ideally suited to a GLR model fit wherein response (probability) distribution is chosen from the exponential family.

Notation and coding for generalized linear models

Besides the afore-mentioned probability distribution, the model has a second component in the equation linking each response variable Y; modeled as a transformation of the mean $g(\mu)$, to a set of linearly related to explanatory variables $x_1, \ldots, x_p$ (De Jong et al., 2008). This linear combination has the form:

$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{3.1}$$

For responses $Y_1, \ldots, Y_N$ this can be written in matrix notation as

$$g[E(\mathbf{y})] = \mathbf{X}\beta \tag{3.2}$$

The *linear* term meaning that the linearly combined explanatory variables x_{ij} linking to each response variable y_i, $E(Y_i)$, their combination gives the **linear com-**

ponent written in equation 3.2. Defined and simplified using matrix notation:

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \text{ is a vector of responses,} \qquad g[E(\mathbf{y})] = \begin{bmatrix} g[E(\mathbf{Y_1})] \\ g[E(\mathbf{Y_2})] \\ \vdots \\ g[E(\mathbf{Y_N})] \end{bmatrix}$$

\end{equation*}

above $g[E(\mathbf{y})]$ a vector of functions of the terms $E(\mathbf{Y_i})$ (with the same g for every element)

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \text{ is the design matrix,} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \text{ is a vector of parameters.}$$

The design matrix $\mathbf{X}$, whose elements are represented by the ijth row/column pairs are measured values of continuous explanatory variables or levels of categorical explanatory variables, and a vector of parameters $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ are partial regression co-efficients. For continuous explanatory variables each unit increase in x_{1j}, y_1 increases by about β_j, holding all other variables constant (De Jong et al., 2008). For categorical explanatory variables, a similar logic applies whereby a modal class variable is chosen and corresponding elements of $\mathbf{X}$ are chosen to exclude and include the appropriate parameters for each variable levels, called dummy variables.

The generalized linear model

The OpRisk loss frequency model uses a nonhomogeneous exponential dispersion modeling (EDM) distribution process with offset function corresponding to a

generalized linear model (GLM), whose capabilities are of a general exponential distribution class unconstrained by the normal assumption and a choice of a monotonic transformation of the mean .viz, an offset variable (Ohlsson & Johansson, 2010). GLM's differ in that:

- The distribution of the target (dependent) variable is chosen from the exponential family, thus the distribution of the response can be heteroskedastic contrasting the homoskedastic assumption of normal regression.
- A transformation of the mean of the response is linearly related to the explanatory variables in the context of the previously defined linear sense i.e., a linear combination of β's.

Exponential dispersion models

As with the linear model, consider independent random variables (rv's) $\mathbf{Y_i}$ which are independent (Ohlsson & Johansson, 2010) but not identically distributed (Covrig et al., 2015). The probability distribution of an EDM whose probability depends on a parameter θ_i that is allowed to depend on i, is given by the following pdf (in the continuous case) or pmf (in the discrete case) of $\mathbf{Y_i}$:

$$f_{Y_i}(y_i; \theta_i; \phi) = \exp\left[\frac{a(y_i)b(\theta_i) - c(\theta_i)}{\phi} - d(y_i, \phi)\right], \quad y_i \in Y \tag{3.3}$$

where the *dispersion parameter* $\phi > 0$ is the same for all i, and a, b, c, & d are regarded as known functions. The dispersion parameter ϕ, is regarded to be fixed or unknown. When ϕ is a (fixed) scale factor in the variance, it defines a one parameter EDM, otherwise its' required to be estimated if unknown; in this case, we usually do not have a two-parameter exponential family, but we do have an EDM (Ohlsson & Johansson, 2010). Considering ϕ as fixed is significant as it maintains congruence

with hypothesis testing and confidence interval estimation consistency standards for GLM theory, as methods such as maximizing the likelihood is trivially equivalent to minimizing the unscaled deviance.

We develop a new ML assisted modeling approach based on Embrechts et al. (2018) and Chavez-Demoulin et al. (2006) dynamic POT-EVT method and Einemann et al. (2018) EBOR model and Ganegoda & Evans (2013) LDA model with exposure-based treatments and OpRisk covariant dependent loss modeling, whose internal data showing stylized features for OpRisk of any form.

Steps in modeling from EDM's to the generalized linear model for count data

Expanding the expression in equation 3.3 yields

$$\begin{aligned} f(y_i;\theta_i;\phi) &= \exp\left[\frac{a(y_i)b(\theta_i)-c(\theta_i)}{\phi}-d(y_i,\phi)\right] \\ &= \frac{1}{e^{d(y,\phi)}}\exp\left[\frac{a(y_i)b(\theta_i)-c(\theta_i)}{\phi}\right] \\ &= r(y,\phi)\frac{1}{e^{\frac{c(\theta_i)}{\phi}}}\exp\left[\frac{a(y_i)b(\theta_i)}{\phi}\right] \\ &= r(y,\phi)s(\theta,\phi)\exp\left[\frac{a(y_i)b(\theta_i)}{\phi}\right] \qquad (3.4) \\ \text{where} \quad r(y,\phi) &= \frac{1}{e^{d(y,\phi)}} \quad \text{and where} \quad s(\theta,\phi)=\frac{1}{e^{\frac{c(\theta_i)}{\phi}}} \end{aligned}$$

The choice of parameter θ_i determines the actual exponential response distribution characterising the statistical unit i, such as the Binomial, Normal or Poisson distribution. Thus, the exponential family representation depends on varying parameters θ_i, and a constant scale parameter ϕ. More precisely, in terms of $c(\theta)$ and $r(y,\phi)$,

$$E(y) = \dot{c}(\theta), \qquad \text{Var(y)} = \phi\ddot{c}(\theta), \tag{3.5}$$

where $\dot{c}(\theta)$ and $\ddot{c}(\theta)$ are the first and second derivatives of $c(\theta)$ w.r.t θ. Table 3.1 displays different choices of θ and $c(\theta)$ leading to several exponential family distributions.

Table 3.1
Exponential family distributions and their parameters (De Jong, Heller, et al., 2008a)

Distribution	Parameter(s)	θ	a(θ)	ϕ	E(y)	$V(\mu) = \frac{\text{Var}(y)}{\phi}$
Binomial	B(n,π)	$\ln\frac{\pi}{1-\pi}$	$\text{nln}(1+e^{\theta})$	1	nπ	n$\pi(1-\pi)$
Poisson	P(μ)	ln(μ)	e^{θ}	1	μ	μ
Normal	N(μ, σ^2)	μ	$\frac{1}{2}\theta^2$	σ^2	μ	1
Gamma	G(μ, ν)	$-\frac{1}{\nu}$	$-\ln(-\theta)$	$\frac{1}{\nu}$	μ	μ^2
Inverse gamma	IG(μ, σ^2)	$-\frac{1}{2\nu^2}$	$-\sqrt{-2\theta}$	σ^2	μ	μ^3
Negative binomial	NB(μ, κ)	$\ln\frac{\kappa\mu}{1+\kappa\mu}$	$-\frac{1}{\kappa}\ln(1-\kappa e^{\theta})$	1	μ	$\mu(1+\kappa\mu)$

The variance function

From equation 3.5, Var(y) for the exponential family distribution,

$$\ddot{c}(\theta) = \frac{\partial\dot{c}(\theta)}{\partial\theta} \equiv V(\mu), \tag{3.6}$$

can always be written as $\text{Var}(y) = \phi\text{Var}(\mu)$, where Var($\mu$) is called the variance function (De Jong et al., 2008); a critical construct unique to the GLM whereby the mean μ varies with the explanatory variables through Var(μ), further differentiating GLM's from the previously mentioned classical linear modeling case. Necessarily, modeling the mean as a function of explanatory variables amounts to modeling the relationship between the variance and the explanatory variables as well.

Since the scale parameter ϕ is constant, equation 3.4 can be written in the form

$$f(y;\theta) = r(y)s(\theta)\mathbf{e}^{a(y)b(\theta)} \tag{3.7}$$

If $a(y) = y$, the distribution is in canonical form and $b(\theta)$ is called the natural parameter of the response distribution (De Jong et al., 2008). Observations on y are assumed to be independent.

Links and canonical links

Once a response distribution $f(y)$ is tailored to the given situation, the link $g(\mu)$ must be chosen. The specific elements of a GLM are (Covrig et al., 2015; Dobson & Barnett, 2008):

1. The random component given by the independent random variables $Y_1, Y_2, \ldots, Y_n$ are not identically distributed. Note that the rv's $\mathbf{Y_i}$ for the Oprisk data, indexed by the subscript i, have different expected values μ_i. Sometimes there may be only one observation y_i for each Y_i, but there may be several observations $y_{ij}, (j = 1, \ldots, n_i)$ for each $\mathbf{Y_i}$. The pdf or probability mass function of $\mathbf{Y_i}$ is given in equation 3.7 for $f(y)$, which specifies that the distribution of the response is in the exponential family. The support set X of the rv Y_i is subset of $\mathbf{N}$ of $\mathbf{R}$.
2. The second advance is the extension of computational methods to estimate the model's systematic component, so called the "linear predictor" described in equation 3.1 built with $p + 1$ parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ and with p explanatory variables:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \qquad i = 1, 2, \ldots, n \tag{3.8}$$

3. The equation for η_i specifies to the situation that there is some non-linear function, a transformation of the mean, $g(\mu)$, that is linearly related to the explanatory variables contained on the r.h.s of equation 3.8, $\mathbf{X_i}^T\beta$, i.e.,

$$g(\mu_i) = \mathbf{X_i}^T\beta \tag{3.9}$$

The function $g(\mu_i)$ is called the link function.

Intepretation

Given a response variable y, for the initial formulation of glm's by John Ashworth Nelder & Wedderburn (1972), $b(\theta)$ determines the nature of the response distribution and the choice of link is suggested by the functional form of the relationship between the response and explanatory variables. In choosing these components extra steps are taken compared to ordinary regression modeling. Commonly used links functions are given in Table 3.2 which also presents the units produced for the various GLM links[3].

Offsets in GLM theory

When a specified additional GLM model variable ξ is appended to equation 3.9, whose co-efficient is constrained to 1, the expression derived (Yan et al., 2009) in equation 3.10 gives an offset feature.

[3]Table 3.2 below referenced to in previous subsection 3.6.1.

Table 3.2
The generalized linear model link functions with their associated units of intepretation. Note: This list is not exhaustive and there are likely more GLMs that are used within previous research

Link Function	$g(\mu)$	Target variable Effect	Canonical link for
Identity	μ	Original Continuous Unit	normal
Log	$\ln\mu$	count	poisson
Logit	$\ln \frac{\mu}{1-\mu}$	Risk	binomial
Probit	$\phi^{-1}(\theta)$	Risk	binomial
Power	μ^p	Count	$\Gamma(p=-1)$
		Count	inverse Gaussian(p=-2)

$$g(\mu) = \mathbf{X}^T\beta + \xi \tag{3.10}$$

The offset feature is thought of as using a residual from a prior "preliminary" regression model where ξ is the predicted value of the mean, $g(\mu)$, as the target variable of the regression of interest. This is the same thing as subtracting ξ from the target variable prior to running the regression, effectively removing the effects of nuisance variables from the target variable (Yan et al., 2009). In fact, Yan et al. (2009) mentions the classic use of an offset is to incorporate a measure of exposure when modeling rates, this is helpful when OVB threatens to distort the covariates.

Supposing the mean $g(\mu) = \lambda$ represents the mean probability for the likelihood of a realised loss over an OpRisk claim. Yan et al. (2009) notes this is mathematically equivalence to replacing the count with the occurrence rate/claim *frequency* (counts divided by exposure) $g(\mu) = \frac{\lambda}{d}$ as the target variable. Then equation 3.10 becomes:

$$\ln\left(\frac{\lambda}{d}\right) = \mathbf{x}^T\beta \quad \Rightarrow \quad \ln\lambda = \ln d + \mathbf{x}^T\beta \tag{3.11}$$

Where g is the log function and the variable d appears representing the risk *exposure* and ln$d$ is an "offset". As per definition 2.10.1.1, the lag represents exposure

i.e., in the OpRisk dataset these are the times to detection (exposure) of the realised losses. Equation 3.11 differs from the usual specification of the linear predictor due to the inclusion of the term ln$d$. With the offset, y has expected value directly proportional to exposure:

$$E(Y) = \mu = de^{\mathbf{X}^T\beta} \tag{3.12}$$

Proposition 3.7.1:

The exposure measure is a known initial condition (constant) which is readily incorporated into the mean estimation procedure and is a quantity that is roughly proportional to the risk (Parodi, 2014).

Proof. Suppose we wish to model the OpRisk loss frequency $y_{\text{rate}} = F$ as a GLM of covariates $\{X_1, \ldots, X_N\}$. Let $y_{\text{realised}} = C$ denote the number of realised OpRisk losses over a period, say years, and d (days) denote the exposure: Then $F = \frac{C}{d}$. The probability mass function of the Poisson distribution justified by the fact that it shows the number of relevant incidents over the specified time period, when i denotes the observation number is

$$P(C_i = c_i) = \frac{e^{-\lambda_i}\lambda_i^{c_i}}{c_i!}, \qquad c_i \in N, \quad \lambda > 0 \tag{3.13}$$
$$\text{where} \qquad \lambda_i = \exp\left[\ln d + \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right]$$

EDM for the Poisson variate

$$f_Y(y_i;\lambda_i) = \exp\{y_i \log\lambda_i - \lambda_i - \ln y!\}$$

The Log-likelihood is found by taking logs and summing over i (see derivation in

Appendix 8.1): Therefore the log-likelihood for the Poisson regression model with offset is:

$$\begin{aligned} l(c|\alpha,\beta) &= \sum_i log\{\frac{e^{-\lambda_i}\lambda_i^{c_i}}{c_i!}\} \\ &= \sum_i -\lambda_i + c_i log(\lambda_i) + \kappa \end{aligned}$$

For simplicity, $\kappa = -\log(\text{c!})$, (assuming that over-dispersion does not exist in the data, i.e., $\psi = 1$). We can recast the obove expression in terms of F and μ:

$$\begin{aligned} l(c|\alpha,\beta) &= \sum_i -d_i\mu_i + c_i log(d_i\mu_i) + \kappa = \sum_i -d_i\mu_i + c_i log(\mu_i) + \kappa' \\ &= \sum_i d_i(-\mu_i + \frac{c_i}{d_i} log(\mu_i) + \kappa') = \sum_i d_i(-\mu_i + f_i log(\mu_i) + \kappa') \\ &= \sum_i d_i \log\frac{e^{-\mu_i}\mu_i^{f_i}}{\kappa''} \end{aligned} \tag{3.14}$$

Since $\lambda_i = \mu_i \mathrm{d}_i$. In the above expressions $\kappa, \kappa', \kappa''$ denote constants that do not depend on the model parameters. the last expression is the log-likelihood function for the Poisson regression, cast in terms of $y_{\text{rate}} = F$. □

Offsets as a measure of exposure

Offsets are usually found in models constructed using the log link function as a measure of exposure in the context of Poisson regression (@ Yan et al., 2009). For example, the exposure measure shows up in epidemiological work as the number of people exposed to a pathogen; in which case @Yan et al. (2009) remarks, the response would be the number of people who contract the disease. Further examples

mentioned by @Yan et al. (2009) found in actuarial work might be the number of car-years insured; and the response would be the number of claims incurred. In both the actuarial and epidemiological settings above, the author agrees with @Yan et al. (2009)'s assertion, that the equation governing the value of the response is assumed to be roughly proportional to the value of the exposure. Other typical examples noted in Parodi (2014)'s accounts of exposure measures are the number of vehicle years, the wage payroll or number of employees for employers liability, the turnover for public liability, and the total sum insured for property and the like.

Furthermore, a useful property unique to the natural base e, when used in an exponential, is the fact that $\frac{d}{dx}e^{rx} = re^{rx}$. Put another way, if $x(t) = e^{rt}$, equivalence between exponential growth and the self-reinforcing growth phenomena viz.,

$$\frac{d}{dt}e^{rt} = re^{rt} \equiv \frac{d}{dt}x(t) = rx(t)$$

Which states: *The rate of change of a value is proportional to the value itself*, therefore extracting a relatable meaning to the constant r in this expression e.g., if $(d/dx)e^{rx}$ were the size of an investment, then the rate of change of the invested value is proportional to the value itself, and r the proportionality constant. By induction e^{rt} solves these systems.

Very often with viral transmission phenomena, the author's reasoning coincides to the relationship therein depicting a certain number of people exposed to a pathogen in the early stages of an epidemic when most of the population is susceptible, and the response would be the number of people who contract the disease as before-mentioned in Yan et al. (2009) and Parodi (2014); or restated in a similar sense, the number of people in the early stages of population growth before the effects of limited resources kick in, in which case when exposure (number of people) is high the more explosive (higher the slope of) the resulting population very much like in cases of exponential

growth. By the same token, the OpRisk loss for count data relationship captured by equation 3.12 depicts a scenario in the early stages of the investment banking deals life cycle of OpRisk loss events before the downstream effects of severe operational hazards are felt, kicking in altering managerial risk attitudes and operational risk controls, typical to scenarios involving exponential growth.

Exploratory data analysis

The main source of the analysis dataset is primary data collected from the internal logs of an investment bank and ethical waiver awarded under the auspices of adherence to anonymity clauses and compliance to non-disclosure agreements. Due to the sensitive nature of the detail, only a quarterly collection of internal OpRisk losses data was possible, and scrubbed onto a memory device for the period between 1 January 2013 and 31st March 2013 could be obtained from an investment bank in SA. The methods of data generation and collection was at the level of the individual transaction, involving unique trade detail 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) reports at the bank in question. 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 (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 (Parodi, 2014).

The loss data collection exercise (LDCE)

In this study, a sample dataset of internally generated OpRisk loss banking data collected from a bank undergoes preparation, to form a new pre-processed dataset consisting of an OpRisk loss frequency of $2,330$ loss amounts and their associated OpRisk loss severities amounting to $R61,534,745 $; consisting of realised, near misses, and pending loss impacts each with unique feature characteristics, developed using the official loss data collection exercise (LDCE) as defined by Committee et al. (2003) for internal data. The dataset in question is at the level of individual loss events which can therefore be modeled in a granular way, which facilitates the reflection of loss-generating mechanisms (Einemann et al., 2018).

A flowchart (detailed in figure 8.1 under Annex 1) adopted from Committee et al. (2001) is developed to for the mining exercise of classifying individual OpRisk data features to match business lines, event types etc., as per Committee et al. (2011): It is therefore also fundamental as part of this study to know when they happened, and be able to identify the root causes of losses arising from which OpRisk loss events. Similarly to the afore-mentioned, this study introduces an analogous mathematical framework for EBOR modeling, however the proposed OR framework is better suited

with a higher probability to determine the amount of capital necessary to absorb operational losses as it is applicable to a larger number of OpRisk types.

The LDCE 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 development, calibration and validation of EBOR models is challenging since new types of data and a higher degree of expert involvement across the institution is required, providing a transparent quantitative framework for combining forward-looking point-in-time data and historical loss experience (Einemann et al., 2018). The afore-mentioned 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. The dataset is then split into proportions and trained, validated and tested.

software package, the GLM function works with two types of covariates/explanatory variables: numeric (continuous) and categorical (factor) variables as depicted in Table 3.3. Multi-level categorical variables are recoded by building dummy variables, i.e., they are recoded by remapping the variables and transforming into numeric variables corresponding to each level. For example, the categorical variable *BusinessLineLevel1* consists of 8 distinct levels, 62% of the data observed belong to the 2$^{\text{nd}}$ business level (BL2) corresponding to Trading and Sales denoted by dummy variable 2; followed by BL3 (Retail Banking) denoted by dummy variable 3, then BL6 (Agency Services) denoted by dummy variable 6, BL5 (Payment and Settlement) denoted by dummy variable 5, then BL4 (Commercial Banking) denoted by dummy variable 4, then BL1 (Corporate Finance) denoted by dummy variable 1, then BL7 (Asset Management) denoted by dummy variable 7, and lastly BL9 (Corporate Level – Non-Business Line Specific) denoted by dummy variable 8 as detailed in Table 3.4.

Tables 8.5 & 8.6 in Appendix 8.6 comprise of the complete transformations for

Table 3.3
The contents of the traded transactions of the associated risk correction events.

Covariate	Levels	Storage Type
Trade		numeric
UpdateTime		numeric
UpdatedDay		numeric
UpdatedTime		numeric
TradeTime		numeric
TradedDay		numeric
TradedTime		numeric
Desk	10	categorical
CapturedBy[4]	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

all categorical variables. This is achieved through an implemented algorithm in R, through a transformation (see subsection 8.6.3 in Appendix 8.6) as recommended for the estimation of the GLM, particularly in the estimation of the poisson regression model for count data. Furthermore, for an equation between physical quantities to be valid, their nature must be homogeneous and such an equation must possess dimensional homogeneity. This condition is maintained by rendering both sides of

Table 3.4
Table generated by Rattle R script (8.6.3) transforming BL Variable to 'Dummy'

BusinessLineLevel1

n	missing	distinct	Info	Mean	Gmd			
1631	0	8	0,8	2,91	1,37			
lowest : 1 2 3 4 5	highest:	4	5	6	7	8		
Value	1	2	3	4	5	6	7	8
Frequency	17	1011	191	116	121	166	6	3
Proportion	0,01	0,62	0,1	0,07	0,07	0,1	0	0

the equation to counts i.e., numerical units, through the afore-said transformation to dummy variables.

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 et al. (2008), Frees & Sun (2010), Denuit, Maréchal, Pitrebois, & Walhin (2007), Cameron & 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 $= e^0 = 1$) the biggest absolute frequency.

Hypothesis 3.9.1:

By virtue of proportionality, a deductable of d% will achieve a reduction in the risk ceded (LGE) by the bank of exactly d%. This translates into a growth phenomena which can either be flat, or increasing or decreasing with λ_i, but it cannot be increasing for some value of λ_i decreasing for others.

Exposure analysis

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 appropriate action is taken. This measure is defined this way when specifically applied to projecting the number of loss events (frequencies). Offsets (defined as exposure) are used to correct for differing periods of observation (De Jong et al., 2008), failure to do so

would raise the specter of OVB. 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 adjust it to current exposure levels" (Parodi, 2014, p. 111). Yan et al. (2009) explains, model variables correlated with days of exposure might possibly pick up some of the variation that should be explained by days of exposure. For example, if some transaction records in the OpRisk loss dataset correspond to semi-annual (0.5*365-day) lags while other records correspond to annual lags, then it is appropriate to use (log of) days of exposure as an offset (Yan et al., 2009). Otherwise, this would result in biased parameter estimates.

The OpRisk loss frequency model uses a nonhomogeneous exponential dispersion modeling (EDM) distribution process (e.g., Poisson, Binomial, Gamma, etc..) with offset function corresponding to a multiplicative GLM process with p.d.f:

$$f_Y(y_i, \lambda_i) = \exp\{y_i \log\lambda_i - \lambda_i - \ln y!\}$$

Estimated by m.l.e with the log-likelihood found by taking logs and summing over i. That is to say, the weighted sums of squared distances of estimated means from observations y_i and deviance play a similar role as the is the cost function (Ohlsson & Johansson, 2010). Note that with an abundance of data, maximising the likelihood is trivially equivalent to minimising the unscaled deviance: Thus indicating that the difference to a squared distance is not significantly large if μ is close to y, suggesting that with a lot of data it may make little difference which deviance one uses.

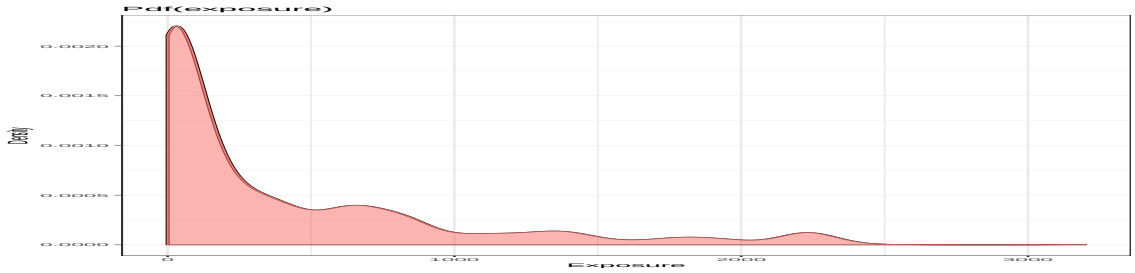
Characteristics of exposure

The exposure is depicted in graphs shown in Figure 3.1. The variable follows a logistic trend on $[0, 1]$, implying an FIs operational risk portfolio rises like a sigmoid

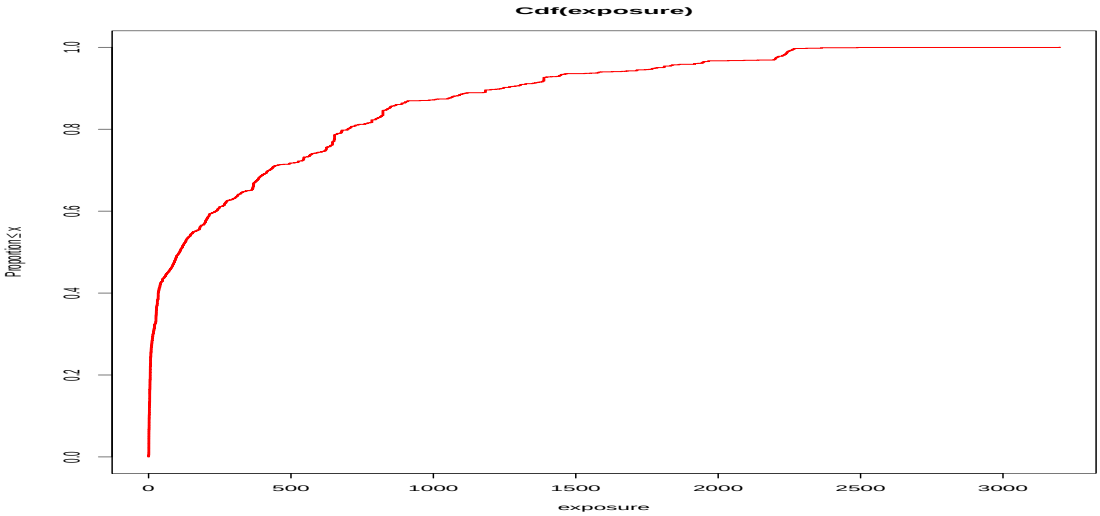
function throughout the period of observation, typically starting from 0, which then observes a plateau in growth. The average exposure is 389.99 or about 1 year, based on descriptive statistics from *OpRiskDataSet_exposure*. Additionally, the degree of concavity of the mid plot in Figure 3.1 is not so extreme to suggest a negligible total loss probability from a small adjustment in exposure, but nonetheless indicative of a situation of (moderate to far less effective) exposures whereby there is a significant probability of maximum possible loss occurring. Grid plots 3.1 portray the logistic function, together with a simple comparison 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.

Interestingly, the sigmoidal nature of the offset function unveiled in subsection 8.6.5 unveils analogous traits with the "activation" function at the center of neuro-scientific inspired artificial neural network (ANN's), which are often likened to biological neural systems. The sigmoidal makeup of the offset in the multiplicative (Poisson) model introduces a sensitivity or known "activation" analogous to the notion of backpropagation in ANN: Backpropagation is an ANN concept that helps minimise the cost of the network where some groups of neurons firing cause others to fire through a chain rule derivative expression describing how sensitive the cost is to a given weight, specifically, w.r.t dependence on how strongly positive or negative one network neuron layer's activation determines activation in the following layer, behind the significantly coined expression "neurons that fire together wire together".

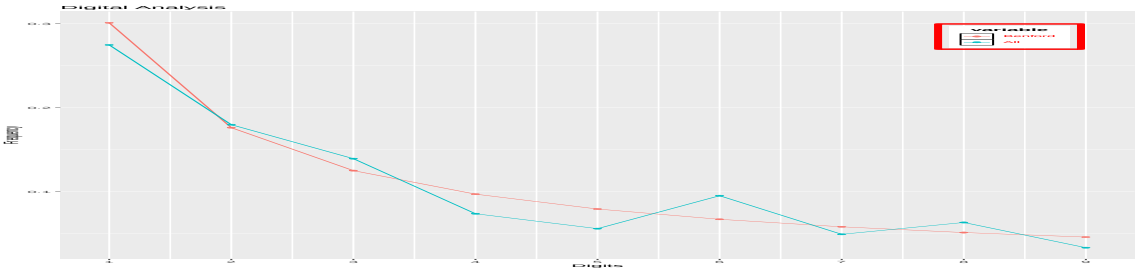
Once you know the effects on these activations on the cost function w.r.t the second to last iteration, the backpropagation process is repeated by applying a nudge to all the weights and biases feeding into that layer; or determined constants in the case of GLM/GAMLSS estimated parameters in an iterative "training" process, until the amounts of all weights and biases feeding into that iteration are added up to



(a) Frequency Distribution Function



(b) Probability Density Function



(b) Digital Analysis

Figure 3.1
A simple comparison of the Sigmoidal like features of the fat-tailed, right skewed distribution for the Exposure variable, and first-digit analysis of frequency distribution from the exposure data with the expected distribution according to Benford's Law

get the overall cost function and adjusted for through minimisation (validated). By repeatedly stepping downhill, the negative gradient of the cost function achieves the ultimate goal in changing the weights and biases in all connections, so as to most efficiently minimise the average cost of all training examples.

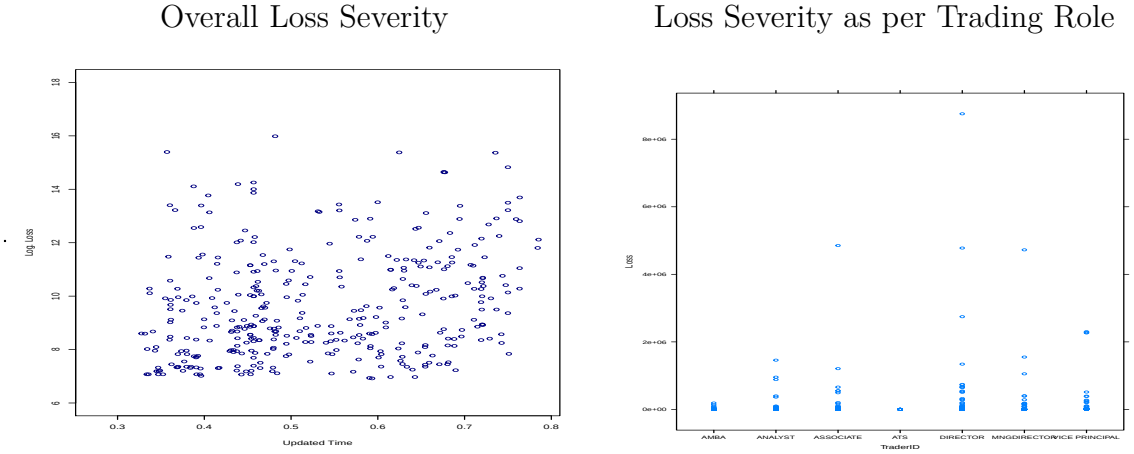
GLM's work with a general class of distribution unconstrained by the normal assumption which lets the data define the functional relationship but maintains an intepretable mathematical representation of OpRisk losses, and a monotomic transformation of the mean viz., the offset whose sigmoidal nature lays the foundation for associated advantages to generic optimisation algorithms such as ANN techniques, stringing along advantages of backpropagation and learning captured through the way the GLM adapts to novel environments, without the widely publicised drawback of uninterpretable networks.

Description of the dataset

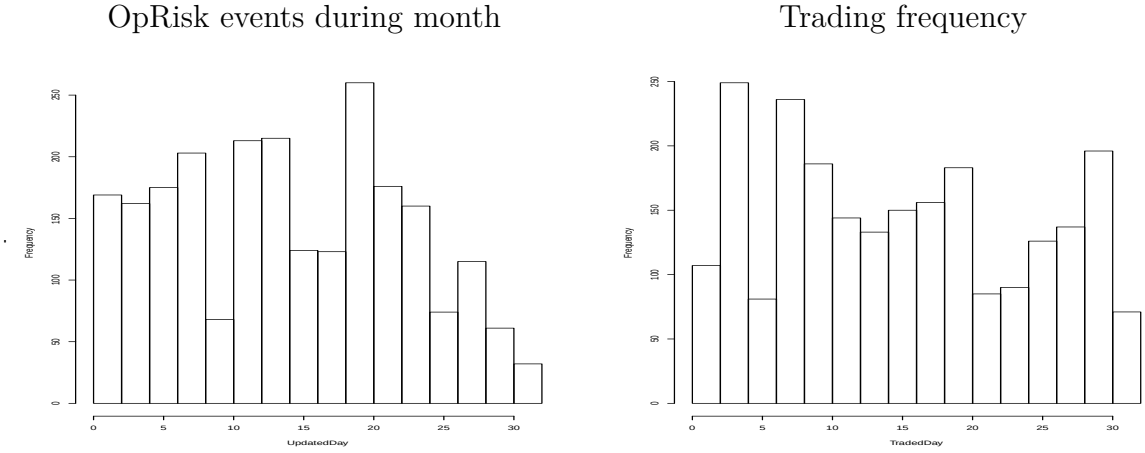
In this section, section 3.11, the dataset called *OpRiskDataSet_exposure*, provides data on the increase in the numbers of operational events over a three month period, beginning 01 January 2013 to end of 20 March 2013. For each transaction, there is information about: trading risk exposure, trading characteristics, causal factor characteristics and their cost.

Characteristics of 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



(a) Intra-day trend analysis of loss severities: overall and as per trading role



(b) Intra-month trends of OpRisk trading incidents compared to frequency of trading activity

Figure 3.2
(a) Scatterplots of intra-day trend analysis for logs of severities of operational events and for those identifying the trading role responsible/originating the loss incidents. (b) As for (a) but intra-month, and in the form of histograms depicting the frequency distrbution of the number daily operational indicents and the frequency of trades.

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 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 [**R**200, 00, **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 midday. 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.

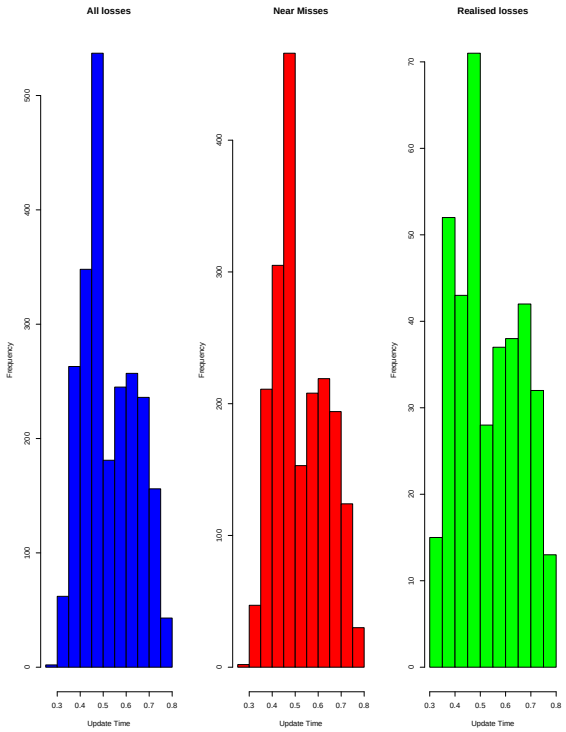
Similarly, the influence of trading desk's on the frequency of operational events can be analysed on the basis of the portfolio's bidimensional distribution by variables *Desk* and *LossIndicator*, which shows the proportions realised losses vs pending and/or near misses for each particular desk. The bidimensional distribution of *Desk* and *LossIndicator* is presented in a contingency table[5], Table 3.5, in which it's considered useful to calculate proportions for each desk category.

Table 3.5
Occurence of realised losses: proportions on desk categories

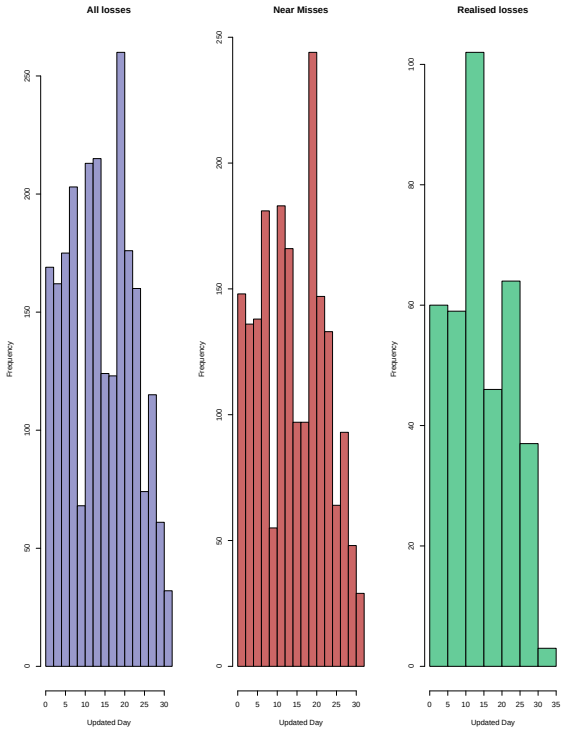
	No. of transactions		
Desk	no Loss	Loss	Total
Africa	49	10	59
Bonds/Repos	113	31	144
Commodities	282	45	327
Derivatives	205	24	229
Equity	269	66	335
Management/Other	41	2	43
Money Market	169	52	221
Prime Services	220	62	282
Rates	336	53	389
Structured Notes	275	26	301

Thus, as illustratred in figure 3.5, from 23,5%; the highest proportion of realised losses per desk is the Money Market desk, the figures are decreasing, followed by Prime Services (22%); Bonds/Repos (21,5%); Equity (19,7%); Africa (16,9%); Commodities

[5]A two-way frequency table, arranged in a matrix format, which is a mechanism for used in statistics to summarize the relationships between frequency distributions of categorical variables



(a) Incidents by the time in the day



(b) 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

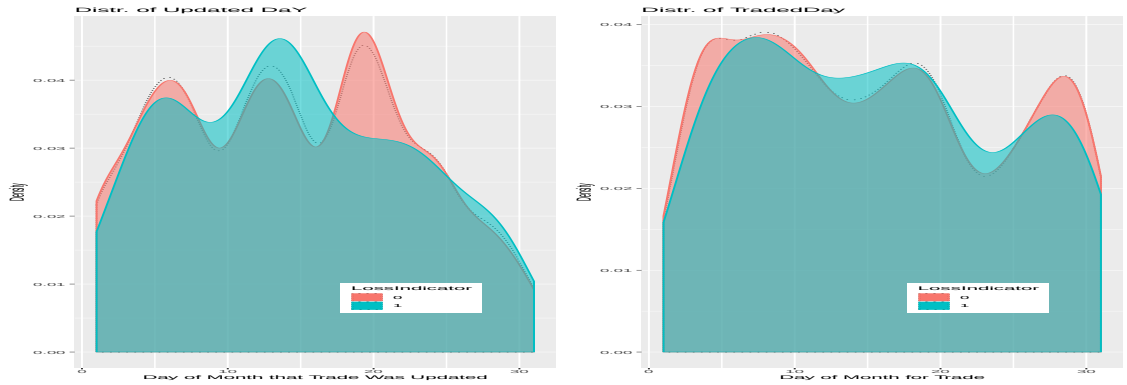


Figure 3.4
Density plots showing a comparison of realised vs pending losses and/near misses over a month for the day in the month the OpRisk incident was updated to the day in the month trades were traded/booked

(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 [6] 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 (algorithmic) trades, which are prone to the business disruption and system failure Oprisk category.

The bottom right mosaic plot in Figure 3.6 for technical support personnel shows the category 'at least one realised loss', which portrays a downward trend, slowing

[6] 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.

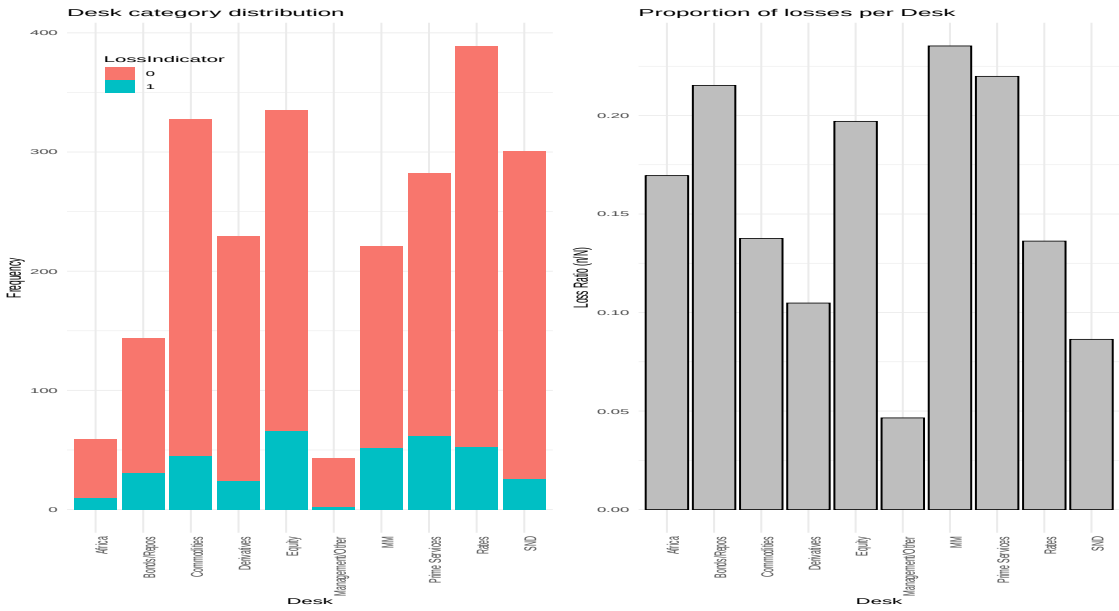


Figure 3.5
Histograms showing the proportions of realised losses vs all losses including pending and/or near misses by desk category

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 unauthorised users are often able to access systems due to lack of oversight, and 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 (unauthorised) user group shoul be prohibited from performing adjustments since don't have the mandate or may lack the skill to do so; notwithstanding being criminally minded insiders acting on their own or in unison to enable their underhanded practices and intentions without raising any suspicion.

Table 3.6
Summary statistics for all losses as per Instrument type

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Mean	23	34,603	46,007	306	7,697	44,157	192,513

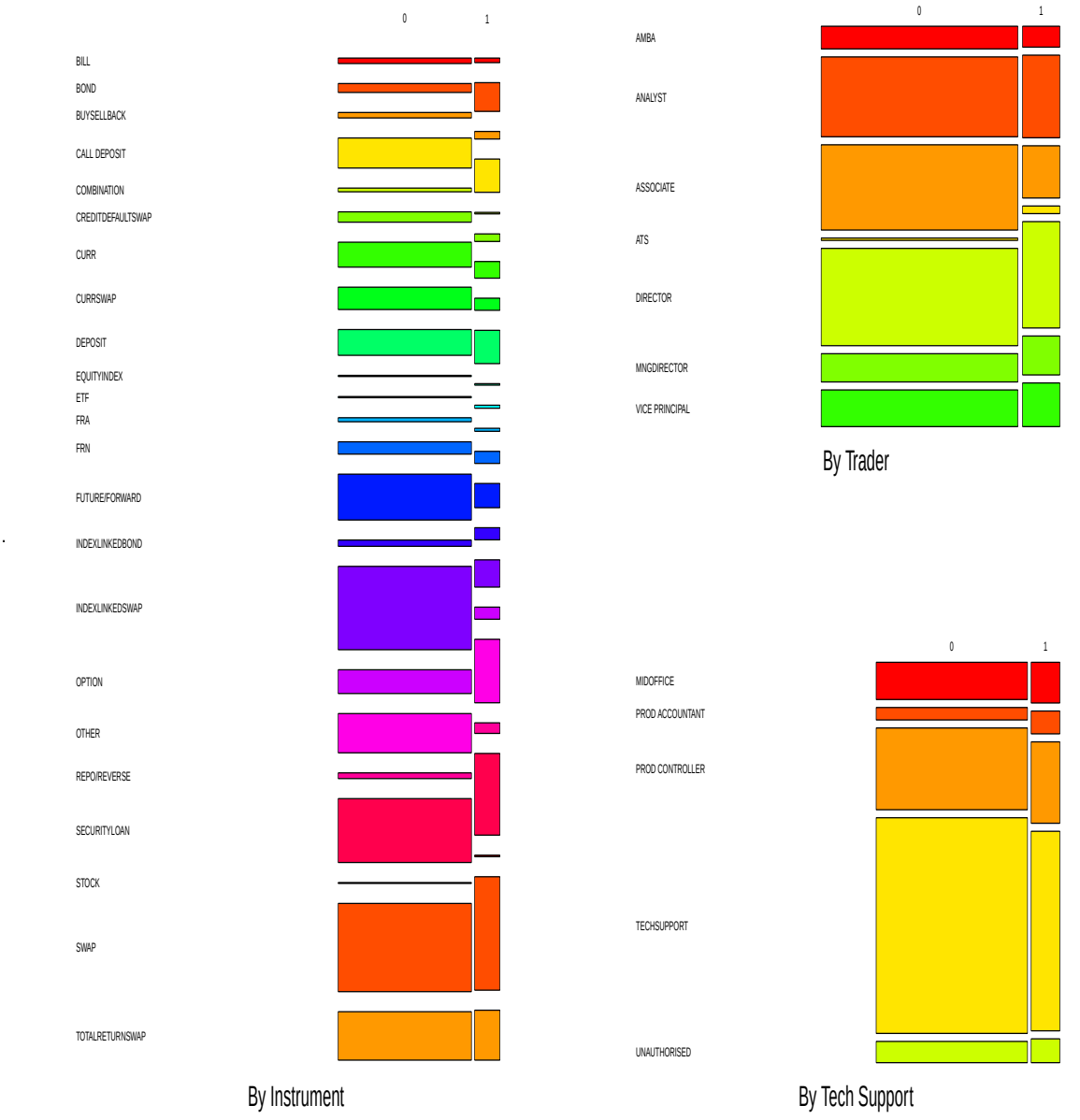


Figure 3.6
Mosaic grid plots for the bidimensional distribution by traded instrument, the trader originating the operational event, and by the technical support personnel involved in query resolution, against the target variable `LossIndicator` showing realised losses vs near misses/pending losses reported.

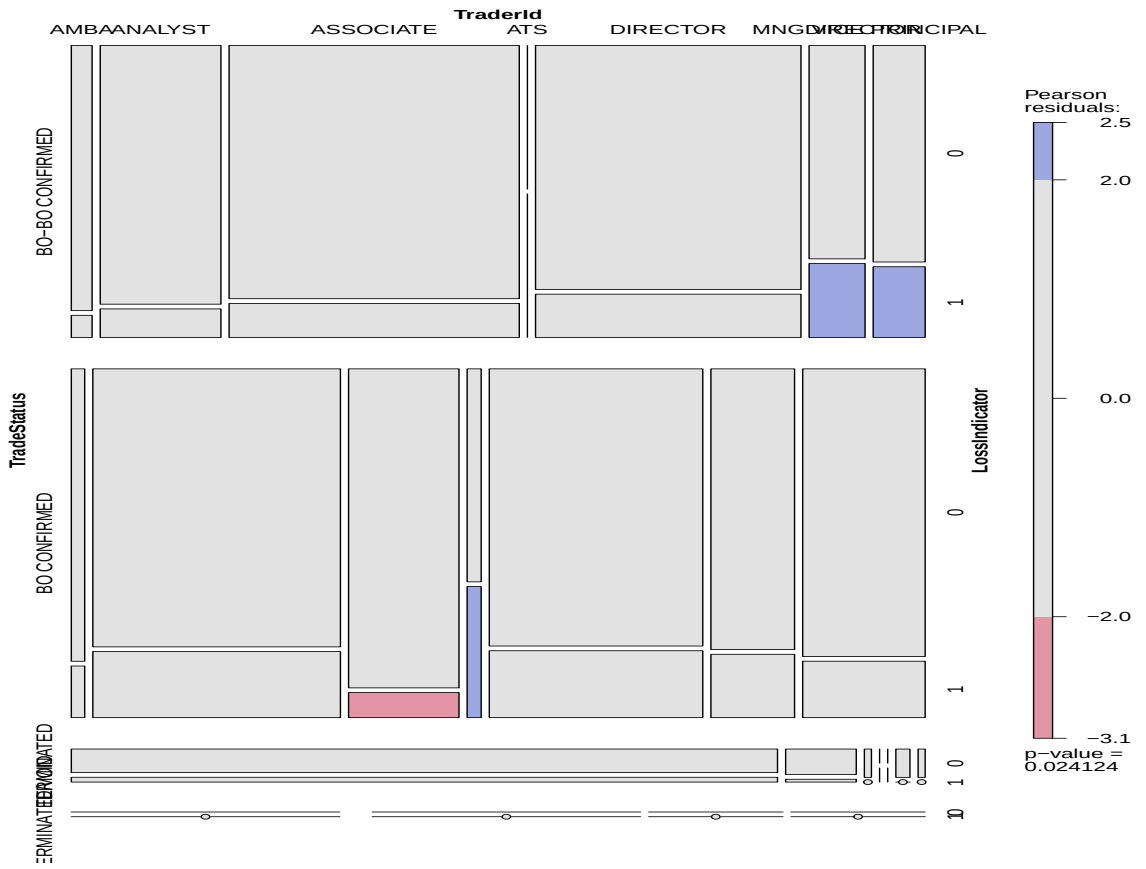


Figure 3.7
A mosaic plot representing the structure of the operational risk portfolio by trader identification (TraderId), the status ofthe trade (TradeStatus) and the number of realised losses vs pending or near misses

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 3.8, 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.

Table 3.7 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 3.7
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 3.8
A contingency table showing the bidimensional distribution of transactions by trader identification vs realised and/or pending losses, conditional on the trade status

Conclusion

Exposure response on risk influence

We have demonstrated equivalence between these two ways of modelling mean probability for the likelihood of a realised OpRisk loss over a specified time interval (by the Poisson regression assumption). As a by product of equivalence between the two specifications 3.11 and 3.10, we also uphold proposition 3.7.1 through the resultant relationship: $\lambda_i = \frac{d}{\mu_i}$: Demonstrating that time to detection (exposure d_i) is a quantity that is roughly proportional to the risk.

The exposure variable behaves as the proportionality constant i.e., (d_i) between the rate (λ_i and the magnitudes of the $p + 1$ parameters. Therefore, the exposure doesn't only represent a mechanism affecting the algorithm directly before and after regression estimation, but when it is doubled, it will indeed consequently double OpRisk. According to literature (Parodi, 2014), there should be an expectation when time to detection (exposure measure for each independent r.v Y_i) doubles whilst everything else (e.g. interest on an interest rate swap) remains the same, for the risk also doubling due to direct proportionality.

Visual exposure response analysis

The empirical finding determined through visual analysis illustrated in Figure 3.1, suggests a moderate exposure curve is a reasonable fit i.e., the exposure curve is neither strongly concave nor too close to linearity. A strongly concave exposure curve tells us that for even a small deductible (10% of the maximum probable loss), say due to an minor adjustment in the exposure (time to detection), whilst everything else

(e.g. interest on an interest rate swap) remains the same for small effects in predictor variable/s, will achieve a reduction in the risk ceded (LGE) by the bank, and the probability of a total loss is minimal (Parodi, 2014).

On the other hand in conjuction with Parodi (2014), if the exposure curve has no concavity at all, then the other extreme of a curve much nearer to the diagonal of a linearly shaped relationship, where a deductible of d% will achieve a reduction in the loss of exactly d%, and all losses are total losses. In this scenario deductibles are far less effective in reducing the risk ceded by the bank, as conceptualised in Hypothesis 3.9.1, exposure is directly proportional to a reduction in the risk ceded by the bank.

CHAPTER 4
AN EXPOSURE-BASED OPRISK ANALYSIS FOR FREQUENCY MODELING

Introduction

The fundamental research premise herein is to mechanise an operational risk management framework (ORMF) toward OpRisk loss with "learning" driving the modeling of capital adequacy requirements based on covariates, which also caters for exposure-based treatments and associated forward-looking aspects. This proves tricky for OpRisk due to the need for specific knowledge of the nature of potential OpRisk loss events viz., the required knowledge of the likelihood of loss-given-event (LGE) behind generating loss frequencies, or complexities due to a timing paradox (Einemann et al., 2018) which drive loss pnl impact generating mechanisms. Contrarily, in this thesis exposure is characterised by the significant lag that results between the moment the event is conceived to the point the loss materializes, thus the aforesaid paradoxical nature doesn't appear to be that all! but is rather there in recognition of an essential feature in timing of the dataset or phenomema studied.

For example, in the case of rogue trading, there is a frequency exposure associated with traders *going rogue*, due to a probability of rogue events happening between a specific group of traders over time, which is then modeled for each rogue trading event and the impact (severity based on the size of the position) of the loss when it is realised (at time of detection). Besides the implied dependence issues, concerns of multicollinearity are duly dealt with through the use of copulas (see chapter 2 Section 2.7), however, the aforementioned "paradox" often results in questionable capital estimates, especially for those near misses, pending and realised losses that need to be captured in the model.

A synthesis of research paradigms, the research design and concepts

A comparative style case study serves as a guide for bridging two competing (positivist and interpretivist) paradigms[1] synthesizing ontological, epistemological and methodological concepts. The researcher sets out on an exercise outlining a series of other distinct works that are complementary in nature to this one: Firstly through an exploratory interpretivist study as indicated in the preceeding chapter (chapter 3) suggesting hypotheses, propositions and mechanisms which avoid forcing plausible pre-examination theoretical structure on data, and thus serving as the basis for a secondary positivist approach, hence triangulating the two paradigms to sufficiently capture the chosen research approach/design.

This reasoning is propped up by advantages that come with combining the best features of both positivist and interpretivist works in the same research work, additionally brought forward by examining and interrogating antecedent models deliberated in Ganegoda & Evans (2013), Chavez-Demoulin et al. (2016), Embrechts et al. (2018), Einemann et al. (2018), and others. Furthermore, this study extends to their concepts through applications of ML technology made possible due to the availability of larger and better quality data sets which not only account for internal controls and risk culture but lead by example, as opposed to Ganegoda & Evans (2013)'s theoretical conceptualisations who don't sufficiently follow through and lack the crucial practical component. As suggested, consideration of literature from this list of authors and other works who use different and complementary approaches that address similar problems are outlined.

Ganegoda & Evans (2013) propose what exists as very general class of regression model so called GAMLSS framework introduced by Rigby & Stasinopoulos (2005), in

[1]A paradigm is a general perspective or way of thinking that reflects fundamental beliefs and assumptions about the nature of organisations

their study establishing relationship found between the size of a firm and OpRisk loss severities. Their model estimates the mean of the loss severity (response) variable but is also found to be appropriately suited to analyse its scaling properties. Not only do Chavez-Demoulin et al. (2016) build on this robustness through their methodological upgrades, but it's further extend to by Embrechts et al. (2018) in the same year with an offering of dynamic POT analysis that undergoes an EVT-based statistical approach for loss severity, and a nonhomogeneous Poisson loss frequency process with rate function, corresponding to a generalised additive model (GAM), also built depending on covariates: The GLM is a special case of the GAM in the sense that in the GAM you have two or more features interacting in your model some of those features are only being called when specific conditions are met, failing which the GAM defaults to GLM; inductively, the GLM being a special case of the GAM and itself a special case of the GAMLSS, therefore they are both special cases of the GAMLSS.

We extend to Ganegoda & Evans (2013)'s proposed methodology for modeling OpRisk loss severity and apply a special GLM case to the general methodology proposed by Chavez-Demoulin et al. (2016) and Embrechts et al. (2018) for loss frequency modeling depending on covariates. Such models avoid spurious results about the relationship between a series of covariates, e.g., exposure of the firm, BL/ET mix, etc, and the loss severity, such that all distributional parameters can be explicitly modeled (by modeling the tail index of the loss distribution) using both fixed and random effects. It suffices to say, if one were to simplify the study to a parametric linear model, you would be back mathematically in the context of Ganegoda & Evans (2013), Chavez-Demoulin et al. (2016), and Embrechts et al. (2018).

Furthermore, Einemann et al. (2018) introduces a general mathematical framework for exposure-based modeling whereby individual events are modeled in a more granular and comprehensive way than in Ganegoda & Evans (2013), Chavez-Demoulin et al. (2016) and Embrechts et al. (2018)'s models to avoid producing questionable

OpRisk loss results. Einemann et al. (2018)'s EBOR framework argue that dependence on covariates associated with internal control weaknesses is usually insufficient to fully capture forward looking aspects, due to the significant aforementioned lag between the time the loss event is triggered until the loss impact realisations.

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

OpRisk is characterised by a time delay τ, wherein the p&l impact lags behind the moment the OpRisk event is conceived up until the event is observed and accounted for. In this paper the author advances knowledge of the current ORMF's toward a new EBOR framework which aims to provide an exposure-based treatment of OpRisk losses catering for modeling capital estimates of forward-looking aspects of OpRisk. In addition, we define an original better fitting "exposure" feature , which differs only in definition to that postulated in Einemann et al. (2018)'s mathematical framework. Thus, a redefined "exposure-based" model is introduced to mitigate the unintended consequences of this "timing paradox"; that often results in disincentives to taking strong risk management steps to mitigate risk, and instead impose it as exposure or an offset whose classical uses eliminate OVB and incorporate it (exposure) as measure when modeling the mean (λ = rate).

In their pioneering study Einemann et al. (2018) unearth a modus operandi behind their EBOR model which is found to be anagolous to the BL/ET matrix combinations in the LDA model[2]. An additional cell is considered whose contributions in the classical LDA model build up a hybrid ORMF which integrates EBOR models with the LDA model, facilitating the migration of OpRisk types from a classical to

[2]Figure 6.4 in chapter 6, subsection 6.9.1 shows the formation of a $7 \times 8 = 56$ BL/ET combination of matrix risk-type cells

EBOR treatment through a quantitative framework (Einemann et al., 2018). Conceptually, the EBOR model component can be extended to include potential future events e.g., future litigations, based on some underlying property, capturing forward looking aspects of business environment and internal control factors (BEICF's) thereof.

In their assessment (Einemann et al., 2018), behind the LDA is the premise that each firm's OpRisk losses are a reflection of it's underlying OpRisk exposure, notwithstanding cases where you could have no loss or near misses although there was exposure. Dobson & Barnett (2008) relates OpRisk events to a varying or a constant degree of exposure, which needs to be taken into account when modeling counts or frequencies of occurance. In particular, the assumption behind the use of the Poisson distribution in the model to estimate the frequency of losses for all available observations, is that both the the intensity (or rate) of occurrence and the opportunity (or exposure) for counting can assume either of these two afore-mentioned forms (Dobson & Barnett, 2008). In the former case the varying degrees of exposure impact on the rate of events, whereas in the latter case the exposure is constant hence not relevant to the model.

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. The actual measure of exposure we need to use depends specifically on projecting the count of OpRisk events (frequency of realised losses) as the target (dependent) variable in the model as opposed to the measure of the target variable where the severity of the losses, e.g. in modeling rogue trading severity exposure of events is based on size of loss position at time to detection, or the acumen in skill level of the capturer (CapturedBy) as severity risk factors.

Sub-problem 1i.

The existing models in OpRisk measurement e.g.; Embrechts et al. (2018) and Chavez-Demoulin et al. (2006) dynamic POT-EVT method, Einemann et al. (2018) EBOR model and Ganegoda & Evans (2013) LDA model with exposure-based treatments, for which historical loss distributions are the best predictors of future losses, who by design lack the learning capabilities attributed to artificial intelligence (AI), hence assume that we do not learn from past losses and therefore suffer from this as a clear limitation. This is problematic for "predictable" risk types due to their model's practice of undercapitalising known[3] risks before they occur, and overcapitalising for risks after the losses materialise, creating inappropriate provisions for capital adequaecy (Group et al., 2013). These concerns motivate the development of an EBOR GLM modeling framework which not only captures past losses but also how exposures to forward-looking aspects affect risk attitudes.

Sub aim (i): Exposure-based OpRisk GLM models

To quantify OpRisk loss frequencies by introducing GLM modeling toward a new EBOR framework for OpRisk management, who are "predictive" due to learning capabilities, capturing exposures to forward-looking aspects in the OpRisk loss prediction problem of how past losses affect may risk attitudes. EBOR treatments effectively look into deep hierarchies in the features of covariates in the investment banking risk forecast business problem for frequencies, in the form of a transformation matrix, $e^{\beta_0 \mathbf{X}^T}$ providing a forward-looking measure as a solution, by using event frequencies based on actual operational risk (OpRisk) exposures in the business envi-

[3]to the extent of predetermined deadlines to outcomes of events with probable financial ramifications, e.g., litigation

ronment and internal control risk factors (BEICF) thereof.

Derivation of the poisson model from the exponential family of distributions

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; which is not all that obvious given that more and more trades are being automated. However, Where OpRisk loss categories such as execution, delivery and transaction processing would intuitively be reduced through automation due to lessened transactional riskiness, other loss categories such as business disruptions and/or losses during system downtime/failures are prone to amplify when they occur due to the sheer volumes. The recent global disruption of work due to the coronavirus (covid-19) pandemic induced transition to a remote working environment puntuates this point. Therefore, it is likely that the rate of operational hazard may well 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 in response to tightening OpRisk management controls. The afore-mentioned postulate provides a plausible model to start investigating this question. As a matter intrigue in this study, the researcher proposes that if they know the determinants of OpRisk losses via their model, they can then use the RHS variables (the independent variables) as controls so set the LHS (the OpRisk loss) to a certain target level, so they could invert the model and impose constraints on the most 'dangerous agents' (the largest contributors/drivers) of OpRisk.

In the above discussions, the question of increasing OpRisk hazard rates due to increasing transaction complexity arises, wherein μ_i, the expected number of new

cases on day t_i is modeled. As a starting point, and with reference to LDA model steps, one begins by using Poisson modeling for counts to estimate the rate of loss events frequencies. The Poisson model's flexibility permits the modeling of numerous operational loss counts and when the data are mostly zeroes and ones (when Poisson means are low). The exponential dispersion model assumes that the number of expected new OpRisk hazards often increase exponentially over time. Hence, if μ_i is the expected number of new cases over time $[T, T+\tau] = t_i$, then an appropriate model takes the form:

$$E(\mathbf{Y}_i) = \lambda_i = d_i \exp\left(\beta_{t_i} \mathbf{x}_i^T\right) \qquad \text{where} \quad \mu_i = \frac{\lambda_i}{d_i} \tag{4.1}$$

where the random variables $\mathbf{Y_i}$ are independent, $d_i = \text{exposure}_i$, and β's are a set of unknown parameters in β. For a list of N different OpRisk loss events, note that the random variables Y_i are the basis for the OpRisk hazard defined by a binary response variable *LossIndicator* which denotes the presence or absence loss. Define random variabels $Y_1, \ldots, Y_N$ as follows:

$$\mathbf{Y}_i = \begin{cases} 1 & \text{for } \textit{realised OpRisk losses} \\ 0 & \text{for } \textit{pending losses and near misses} \end{cases} \tag{4.2}$$

pending losses refers to events whose pnl impact is yet to be determined

$$\mathbf{Y}_i = y_{ij} * \text{LGE}(i) \tag{4.3}$$

indexed by the subscript i, who may have different expected values μ_i. It is important to note that sometimes there may be one observation y_i for each Y_i, but on other occasions there may be several observations $y_{ij} \quad (j = 1, \ldots, n_i)$ for each Y_i. Equation 4.1 can be turned into GLM form by using a log link so that

$$\ln\lambda_i = \ln d_i + \beta_{t_i}\mathbf{x}_i^T \tag{4.4}$$

Parameter λ_i, (the mean of the quantity being measured) will depend on risk factors e.g., the transaction population size, the period of observation, and various characteristics of the population (i.e., UpdatedTime, Instrument, TraderId, etc.) which are the causal factors that are associated with OpRisk hazards and therefore the basic units of measurement for estimating the resulting losses (with associated random errors . The transposed vector $\mathbf{x}_i^T$ represents the ith row of the design matrix $\mathbf{X}$, it takes the form; $t_i = x_{ij}^T, (j = 1, \ldots, p_i)$ for p explanatory variables (covariates or dummy variables).

The response variable is a series of OpRisk events $\mathbf{Y}$ where the probability of the event occuring in a very small time (or space) is low and the events occur independently. Again, since this is a count the Poisson distribution is probably a reasonable distribution to try. The Poisson distribution is denoted by $\mathbf{Y_i} \sim \mathbf{Poi}(\theta_i)$. Rewriting Equation 3.7 as

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)], \tag{4.5}$$

Substituting $a(y) = y$, $b(\theta) = \ln\theta$, $c(\theta) = -\theta$, and $d(y) = -\ln y!$; given ln is some monotone differentiable (link) function, so the GLM for this situaton uses a Poisson response distribution, log link: Equation 4.5 can be expressed as:

$$f(y_i;\theta) = \exp\left[y\ln\theta - \theta - \ln y!\right] \tag{4.6}$$

Equation 4.6 is the probability function for the discrete random variable $\mathbf{Y}$, it can be rewritten as

$$f(y,\theta) = \frac{\theta^y e^{-\theta}}{y!} \tag{4.7}$$

Where y takes the values $0, 1, 2, ...$ If a random variable has a poisson distribution, its expected value $E(Y)$ and variance $Var(Y)$ are equal i.e., $\theta = \lambda$.

The choice of the Poisson distribution for use on real world data becomes questionable since OpRisk loss concerns often result from high real world earnings volatilities, especially where the Poisson mean is low so that the data are mostly zeroes thus have a larger variance than the expected value i.e., *overdispersed*. It is always advisable to check the validity of the regression assumptions, this can be done using a procedure such as univariate least absolute deviation or rank regression (Lewis, 2004): A correctly specified model would be when its residuals have smaller standard errors than if one estimated each regression equation separately. A clear violation of the independence assumption shows up in a pattern when means of residuals are plotted against fitted values, probably resulting from an ommision of something important in the model (Wood, 2017).

As Lewis (2004) and De Jong et al. (2008) state, several procedural tests can be carried out viz., firstly evaluating the adequacy of a constant, then linear model, and if these aren't inadequate, add a qaudratic term, quintic term etc., in a sequential analysis to decide whether the addition of each term is justified. Suppose these data describes two main constructs for OpRisk loss modeling in and organisation viz., "internal controls" and "risk culture": the Poisson GLM which treats both as categorical variables is

$$y_{ij} \sim |\mathrm{P}(\lambda_{ij})|, \qquad \ln(\lambda_{ij}) = \ln d_{ij} + x_{ij}^T \beta \tag{4.8}$$

Here $i = 0, \ldots, l$ refers to internal controls, and $j = 1, \ldots, k$ to risk culture. Re-

sponse y_{ij} is the LGE count while $\ln d_{ij}$ is the exposure at risk. The explanatory variables are the usual intercept β_0, indicator variables corresponding to internal controls, and indicator variables corresponding to risk culture. The two main permutations to these procedures are to: [1] Test each expanatory variable sequentially, under the assumption that the previous explanatory variables are included in the model, [2] Test each expanatory variable under the assumption that all other variables are included in the model. The model has $(l-1)+(k-1)+1$ parameters.

$$\begin{array}{lll} \ln(\lambda_{ij}) & = & \ln d_{ij} + \beta_0 + \beta_1 i + \cdots + \beta_p i^p + \beta_{p+1} j + \cdots + \beta_{p+q} j^q. \\ where & (p,q) & \in \mathbb{N} \end{array} \tag{4.9}$$

Supposing as De Jong et al. (2008) notes, smoothness can be exploited by specifying polynomials in the regression for the explanatory variable in question by modifying the RHS of Equation 4.9 with higher term in the polynomial so that other situations other than the unrestricted spread of OpRisk hazards. In particular, adding a quadratic term $(\beta_2 t_i^2)$, now usefully approximates other situations which may influence the counts adapted to the Poisson case other than only those due to the unchecked prevalence of Oprisk hazards.

$$\mu = d_i \exp\left(\beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2\right) \tag{4.10}$$

A poisson regression operational hazard model

The random component is given by the independent random variables $Y_1, Y_2, \ldots, Y_n$, not i.i.d (Covrig et al., 2015; Wood, 2017). $\mathbf{Y}$ takes a (exponential) family argument, depending on parameters $\ln\lambda$, where λ represents the average frequency of the OpRisk transactions. The response data y_i is an observation of Y. The target variable

LossIndicator defined as per definition 4.2 is the basis for the poisson distribution as a reasonable model of choice. As per equation 4.7, it's probability mass function (pdf) is:

$$\begin{aligned} Y \quad &\sim \quad \text{Poi}(\lambda), \quad f(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!} \\ &\text{where} \quad y \in \mathbb{N}, \text{and} \quad \lambda > 0. \end{aligned} \tag{4.11}$$

Again, the expectation and variance $E[Y] = \text{VaR}[Y] = \lambda$[4], are both equal to parameter λ simultaneously. The model's systematic component, equation 3.8 specifies the linear predictor and is built with $p+1$ parameters $\beta = (\beta_0 \ldots, \beta_p)^t$, with p explanatory variables:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \qquad \text{where} \quad j = 1, \ldots, p_i \tag{4.12}$$

If sample variables $Y_i \sim \text{Poi}(\lambda_i)$, then $\mu_i = E[Y_i] = \frac{\lambda_i}{d_i}$; the link function between the random and systematic components, viz. a transformation by the model by some function $g()$, which does not change features essential to to fitting, but rather a scaling in magnitude: i.e., the link between natural canonical parameter θ in equation 3.3 and parameter λ, the mean frequency of poisson distribution $\theta = \ln\lambda$, or otherwise the rate, will be predicted by the model…

$$\begin{aligned} \lambda_i &= d_i \text{exp}(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) \quad \text{or} \\ \lambda_i &= d_i \cdot e^{\beta_0} \cdot e^{\beta_1 x_{i1}} \cdot e^{\beta_2 x_{i2}} \ldots e^{\beta_p x_{ip}} \end{aligned} \tag{4.13}$$

Where d_i represents the risk exposure for transaction i. Taking logs on both

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

sides of equation 4.13, the regression model for the estimation of loss frequency is:

$$\ln\lambda_i = \ln d_i + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \tag{4.14}$$

where $\ln d_i$ is the natural log of risk exposure, called the "offset variable".

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, Wood (2017) explains, 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 thesis.

Proposition 4.6.1:

EBOR OpRisk treatments through GLM modeling frameworks "learn" from data to provide a forward-looking LGE of the risk matrix solution to LDA loss frequency modeling problem.

Logistic regression and GLM's: Loss frequency, Indicator variables

A GLM is introduced starting by estimating the expected number of OpRisk events (the mean OpRisk frequency) by a poisson model given by equation 4.11: Followed up by estimating the target variable using the binomial/bernoulli distribution

complementing the numerical counts with factor (categorical) outcomes, thereby covering the range of probable model fits using the glm function. In calling the GLM we specify the target variable *LossIndicator*; the explanatory variables are also composed of numeric, continuous and categorical variables. Where the variable in the argument of a GLM is categorical, we choose to specify the modal class as the reference level. A user defined function "getmode" accomplishes the following as specified see chapter 3 section 3.9 on page 83: It selects the modal class of observation in each factor, and the dataset is reordered using the *relevel* function in RStudio.

Estimation of some poisson regression model

Let us consider the OpRisk loss model for frequency of losses, whereby the *LossIndicator* is the target variable: We shall estimate the mean quarterly rate in the OpRisk hazard portfolio through poisson regression models i.e., the target variable is *LossIndicator*, the mean daily loss frequency in the risk correction statistics is estimated through the poisson regression model. Other GLM arguments are: The aforementioned link function poisson(link="log"); a data frame containing the OpRisk dataset, data=crs$training; 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 introducing 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. The following fits the model (the log link is canonical for the poisson distribution, and hence the R default) and checks it:

```
Call:
glm(formula = LossesIndicator ~ TradedDay + Desk, family = poisson(link="log"),
```

```
data = crs$training, offset = log(Exposure))
```

The output in the appendices Appendix 8.7, see basic model build summary presented in subsection 8.8.1, where significant[5] variables are depicted. The coefficients of the categorical variable *Desk* are reordered and weighted against the modal class: *DeskRates*. Interestingly the modal class is ommitted in the results section since the coefficient of the modal class is $e^0 = 1$, indicating that the remaining classes are weighted against it.

Using this bivariate model wherein the author describes the OpRisk loss frequency data through parametisation, the estimated quarterly OpRisk (LossIndicators) frequency of realised losses for each *Desk* category (excluding the insignificant ones) are:

* $0,0013 = e^{-8.053221} \cdot e^{-0.014087} \cdot e^{1.457695}$, for the combination of the TradedDay and DeskAfrica category, which implies that frequency of realised losses for this combination of predictor variables is 4.3 $(= \cdot e^{1.457695})$ fold (times) higher than the realised loss frequency of OpRisk causes in the reference desk category, viz. the Rates desk.
* $0,0018 = e^{-8.053221} \cdot e^{-0.014087} \cdot e^{1.764230}$, for the combination of the TradedDay and DeskBonds/Repos category, which implies that frequency of realised losses for this combination of preditor variables is 5.83 $(= \cdot e^{1.764230})$ times higher than causes in the reference desk category.
* $0,0008 = e^{-8.053221} \cdot e^{-0.014087} \cdot e^{0.924033}$, for the combination of the TradedDay and DeskCommodities, which implies that frequency of realised losses for this combination of preditor variables is $2,52$ $(= \cdot e^{0.0.924033})$ fold higher than the causes in the reference desk category.
* $0,0012 = e^{-8.053221} \cdot e^{-0.014087} \cdot e^{1.365152}$, for the combination of the TradedDay and

[5] A variable is significant if its p value < 0.05

DeskEquity, which implies that frequency of realised losses for this combination of preditor variables is $3,92 \quad (= \cdot e^{1.365152})$ fold higher than the causes in the reference desk category.

* $0,0026 = e^{-8.053221} \cdot e^{-0.014087} \cdot cdote^{2.129594}$, for the combination of the TradedDay and DeskPrime Services,an increase of $8,4 \quad (= \cdot e^{2.129594})$ fold times higher w.r.t the baseline (the Rates desk)
* about $0.00015 = e^{-8.053221} \cdot e^{-0.014087} \cdot e^{-0.716361}$ of the last desk category DeskSND, which means a decrease of about 51

The predicted mean frequency (λ) of OpRisk losses for operational losses i, for the bivariate model **freqfit1**, is given by:

$$\begin{aligned} \mu_i &= \text{exposure}_i \cdot e^{-8.053221 \cdot \text{Intercept}_i} \cdot e^{-0.014087 \cdot \text{TradedDay}_i} \cdot e^{1.457695 \cdot \text{DeskAfrica}_i} \\ &\cdot e^{1.764230 \cdot \text{DeskBonds/Repos}_i} \cdot e^{0.924033 \cdot \text{DeskCommodities}_i} \cdot e^{1.365152 \cdot \text{DeskEquity}_i} \\ &\cdot e^{2.129594 \cdot \text{DeskPrime Services}_i} \cdot e^{-0.716361 \cdot \text{DeskSND}_i} \end{aligned} \tag{4.15}$$

We now fit a more comprehensive model wherein all 13 explanatory variables are introduced into the global model; this shows realised losses for quarterly OpRisk incidents in the all varibles inclusive case. Again, we use *LossIndicator* as the target variable, while the other 13 variables are predictor (explanatory) variables. Fitting the glm global model yields output seen in subsection 8.8.2 of the Appendix 8.7 on page 246.

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 (bivariate) model the AIC value is 2253.4 and 1907.6 for the second (multivariate) model, which suggests that the second model,

freqfit, the model in which we considered an all inclusive list of 13 predictor variables is a better fit since there is a marked reduction/improvement in AIC magnitudes compared to the first value, hence **freqfit** is prefered over the bivariate (first) model.

Estimation of some binomial regression model

Another approach to the estimation of the mean frequency is to assume that the variable that shows the mean frequency follows a binomial distribution. Consider the predictors of such a regression model are given by the global model and presented by calling the glm model yielding output results seen subsection 8.8.2, section 8.8 in Appendix 8.7.

```
Call:
glm(formula = LossesIndicator ~ UpdatedDay + UpdatedTime + TradedDay +
    TradedTime + Desk + CapturedBy + TradeStatus + TraderId +
    Instrument + Reason + EventTypeCategoryLevel1 + BusinessLineLevel1,
    family = binomial(link = "logit"), data = crs$training,
    offset = log(Exposure))

    Null deviance: 2380.8  on 1630  degrees of freedom
Residual deviance: 1270.6  on 1553  degrees of freedom

AIC: 1426.6

Number of Fisher Scoring iterations: 18
```

The AIC value for the binomial model of 1426.6 is less than that of the poisson model 1907.6, therefore by the same token to that fashioned in section 4.7.1, an estimation of the models by a comparison of information criteria (AIC's) which enables the choice the most appropriate or "best" fit model is carried out: First through establishing significance viz., if the residual deviance and the corresponding number of degrees of freedom don't have value significantly bigger than 1 i.e., the multivariate model freqfit $\frac{1270.6}{1553} = 0.8$, and therefore retaining the binomial model i.e., the model

with the smaller AIC value.

Model selection and multimodel inference

Burnham & Anderson (2002)'s introduction of the information-theoretic approach permits a data-based selection for the "best-fit" model in the analysis of the OpRisk dataset *OpRiskDataSet_exposure.csv*, and a ranking and weighting of what remains. This approach allows traditional (formal) statistical inference to be based on the selected "best-fit" model, which is now based on more than one model (multimodel inference). As a requirment the R package to load is the multimodel inference (*MuMIn*) Rstudio package.

Data dredging

We then use the *dredge* function to generate models using combinations of the terms in the global model. This function also calculates 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".

The function "MuMLn::dredge" returns a list of 4097 models, which is every combination of predictor variable in the global model freqfit. Model number 894 is the best-fit: All predictor variables included in this model have a positive effect on the target variable except for the preditor TrddD (**TradedDay**) which has a negative effect on the likelihood of a realised loss (target variable *LossIndicator*) i.e., the later

in the month of the transaction, the less likely a loss is realised. Additionally, from the delta (=delta AIC) one cannot distinguish between models 894, 382, 1918 and 1406 since (using the common rule of thumb) they have AIC < 2.

Of the top seven models (listed below); 1918 & 2942 each hold nine; 894, 1406 & 1854 hold eight; 382 & 830 hold seven; and lastly 318 hold six predictor variables respectively. Where a variable doesn't have a value associated with it does not mean no effect, but rather that it was not included in the model. For example, model 894 returns a combination of the eight variables 1/2/3/4/5/6/7/8, corresponding to top most model in the following output predictor variables (abbreviated in the header row), see figure 8.2, subsection 8.9.1, section 8.9 in the Appendix 8.7.

Information from the AICc's values suggest, that of the top eight 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 & 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 & 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, if we want to get only the models that have delta AICc < 2; we therefore use "subset=TRUE"

Now we have AICc values for our models and we have the average (mean) model, we denote this **Amodel** sumamrized below. For full comprehensive results, see figures

8.3, 8.3 & 8.3 in Appendix 8.9.2.

```
Call:
model.avg(object = get.models(freqfits, subset = TRUE))

Component model call:
glm(formula = LossesIndicator ~ <4096 unique rhs>, family =
poisson(link = "log"), data = crs$training, offset = log(Exposure))

Component models:
                           df   logLik    AICc  delta weight
1/3/4/5/6/9/10/12          71  -901.58 1951.72   0.00   0.08
1/2/3/4/5/6/9/10           74  -898.47 1952.07   0.36   0.07
1/3/4/5/6/9/10             70  -902.97 1952.32   0.60   0.06
1/2/3/4/5/6/9/10/12        75  -897.57 1952.47   0.75   0.06
1/3/4/5/6/8/9/10           71  -902.12 1952.80   1.08   0.05
1/2/3/4/5/6/8/9/10         75  -897.89 1953.12   1.40   0.04
1/3/4/5/6/9/10/11/12       72  -901.23 1953.20   1.48   0.04
 [ reached getOption("max.print") -- omitted 4888 rows ]
```

Discussion on single "best model" vs ensemble

With the best model for this dataset, the traditional thinking of how to go about VaR and the question around how often an extreme event will happen for how much of a buffer to subsidise against OpRisk losses, was answered by using distributions and looking at percentiles of extreme events as estimated by one "best model". The practical implications of more accurate predictive analytics on this question, say in a case of determining the capital buffer to ward against OpRisk losses may require a rethink, as with more accuracy there is no single best model, rather an ensemble wins.

This may shift going about the historical thinking of OpRisk VaR due to this more accurate predictive analytical measure i.e., when there is no single "best model" when going about the thinking of mitigating against the effects of the OpRisk problem:

With an ensemble we see less subsidies between risks within the pool of independent risks, and better subsidies can be designed as protection against the possibilities of extreme events, resulting in guarding against over/under compensated buffers which in turn results from the reduction of variance that arises due to aggregating/pooling independent risks.

Consequently, higher risk extreme events who would normally be charged more than they are prepared to as per proportionate OpRisk loss allocation, now benefit equitably along with the risk categories depending on their allocated risk threshold aligning risk takers within their respective trading limits.

Model performance evaluation

We have gained initial insights through data exploration in Section 3.9 and then built models. The next critical step is to evaluate our model using data mining techniques. For this we need to split our data sample into three subsets. We use a 70/15/15 sampling strategy; the 70% subset sample for the *training dataset*; for the sake of model building we defined and used the training dataset, 15% for the *validation dataset* used to test different parameter setings or different choices of variables whilst we are data mining, and another 15% for the *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 this we use the training and validation datasets (Williams, 2011).

The resulting error estimates come out in the process of testing the performance of the models we build, by first using the validation dataset in preliminary and intermediate stages, putting in adjustments where required and re-evaluating the error rate, until the final modelling phase where the refined model is evaluated using the testing dataset to provide an unbiased error of the final results. In this way the fitted

model is used for ex post prediction and accuracy is determined by looking at the discrepancy between actual and forecasted values.

Confusion Matrix and Statistics

To measure the level of accuracy of the decisions made by the model compared with the actual decisions, a *confusion matrix* is used. The confusion matrix, otherwise known as an *error matrix* is a mechanism used to provide an understanding of how well the model will perform on new previously unseen data i.e., used to evaluate the model.

The confusion matrices for the training, validation and testing datasets for the poisson **Amodel** that we have previously seen are shown in Figure 4.1. Two tables per dataset are displayed, the first list the actual counts of observations and the second the percentages. We can observe using the testing dataset, that for 69% of the predictions the model correctly predicts pending or near misses i.e., no pnl loss impact (called the true negatives). That is, 243 out of the 350 loss events are correctly predicted as pending or near misses. Similarly we see that the model correctly predicts realised OpRisk losses (called the true positives) on 6% of the events. These two combined determines the (above average) performance level of the model which has a prediction accuracy of 75% versus what the actual OpRisk losses were.

In terms of how correct the model is, we observe that it correctly predicts near misses and pending losses 243 out of the 287 events on which losses no not actually materialise. This is $\sim 85\%$ accuracy for predicting near misses and pending Oprisk loss events. This is called the true positive rate, also know as the recall of a model. It is a measure of how many of the actual positives the model can identify, or how much the model can *recall*. This recall is also known as the *Sensitivity*. Similarly, the true negative rate is another measure which also arises and is known as the *Specificity*.

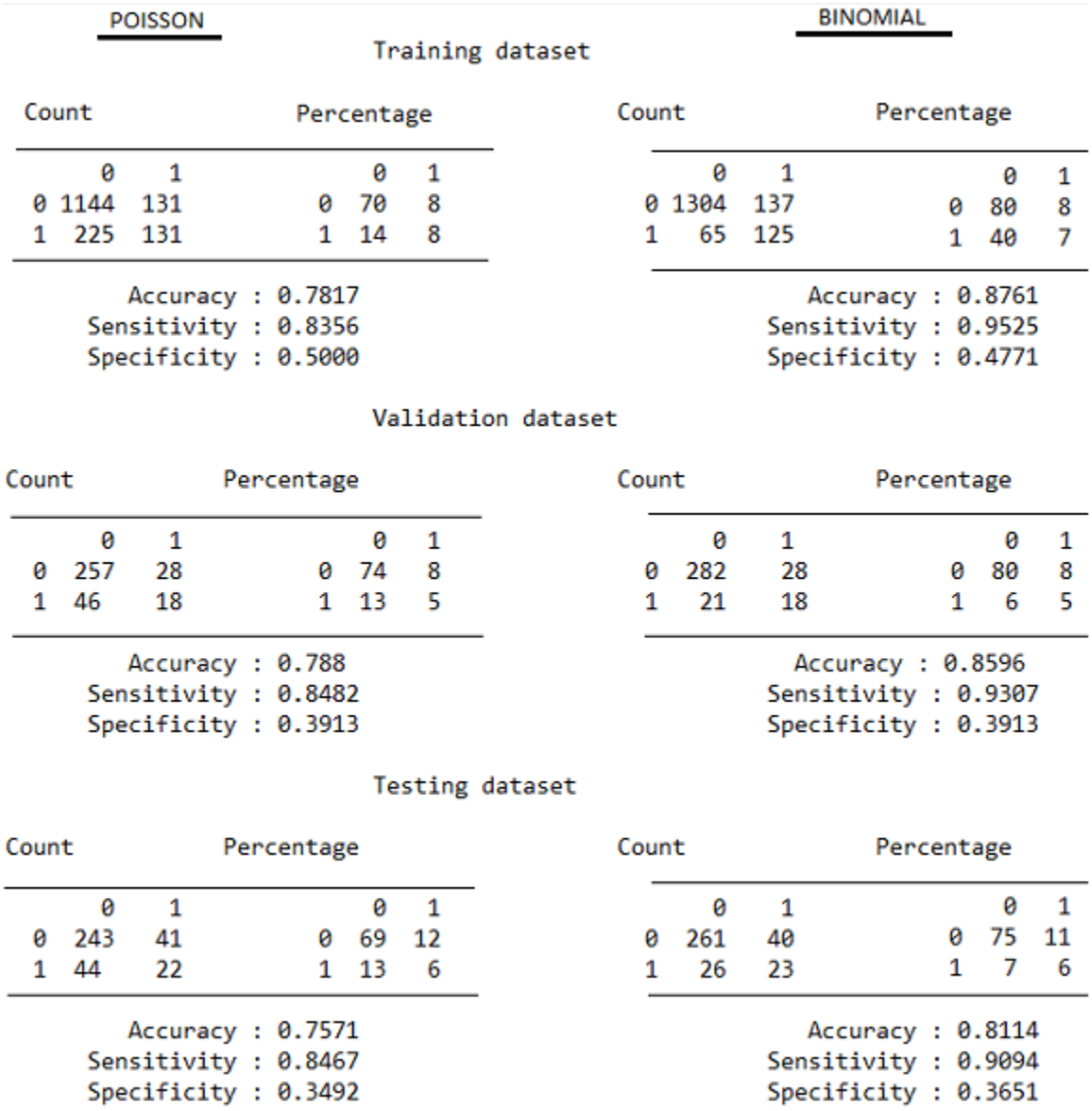


Figure 4.1
Basic comparison of prediction and actuals as confusion matrices using both counts and calculated as a percentage. Summary statistics are computed and displayed

It correctly predicts realised OpRisk losses 22 out of the 63 events on which losses actually materialise. This is $\sim 35\%$ accuracy for predicting realised Oprisk loss events.

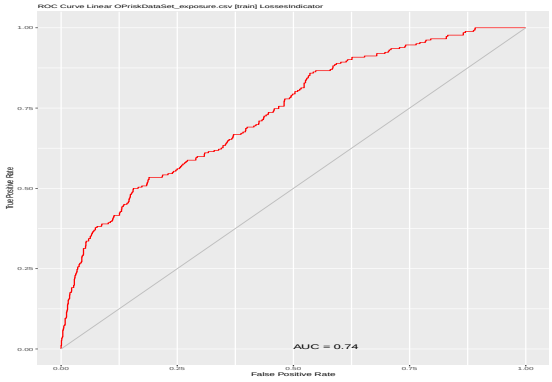
We also see 41 potential events when we expect pnl loss impacts and none occur (called the false positives). If we were using this model to help us decide whether or not to be wary and closely monitor future daily trading activity, then there aren't probably serious conscequences in this instance, we heightened our risk threshold without need. More serious though is that there are 44 actual loss events when our model tells us there will be no pnl impact yet losses occur (called the false negatives). We are inconveniently incurring OpRisk losses without moderating for the risk. Notice that the overall accuracy of the training dataset is 78% which is not surprising as the estimated accuracy of the resulting learned model leads to overoptimistic estimates during the validation state, due to systemic algorithmic training viz., overspecialisation of the learning algorithm to the data.

ROC Curve (receiver operating characteristic curve)

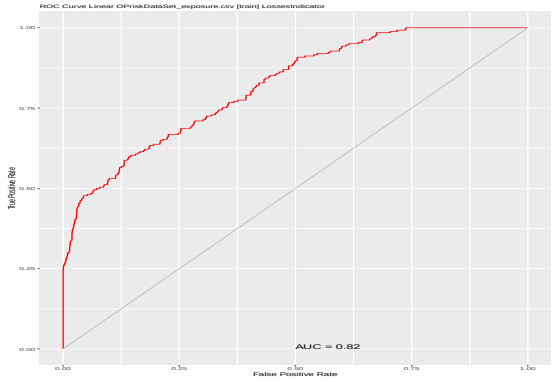
An ROC chart plots the true positives against the false positive rate, essentially to compare the performance of the model against known outcomes and is used to identify a suitable trade-off between effort and outcomes. Generally the larger area-under-curve (auc) also the probability that the classifier scores a randomly drawn positive sample higher than a randomly drawn negative sample. For a more comprehensive comparison, see figure 8.7 in Appendix 8.7.

Discussion on multimodel inference (MuMIn) model performance

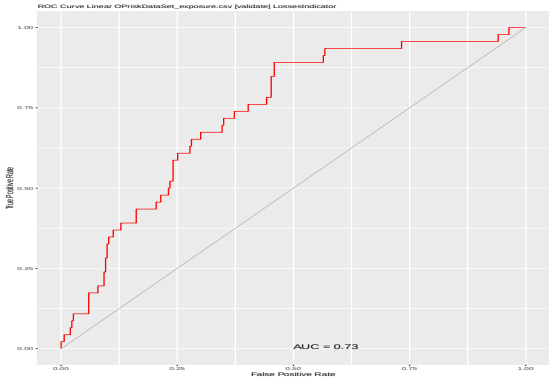
Conceptually the "best model" **Amodel** represents the phenomena hypothesized from the information in the observed data, which then forms the basis for



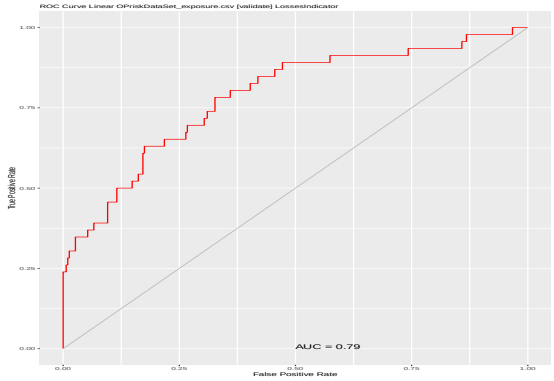
(a) AUC for poisson parameter depicting model's performance on the training set



(b) AUC for binomial parameters depicting model's performance on the training set



(c) As in (a) but depicting model's performance on the validation set

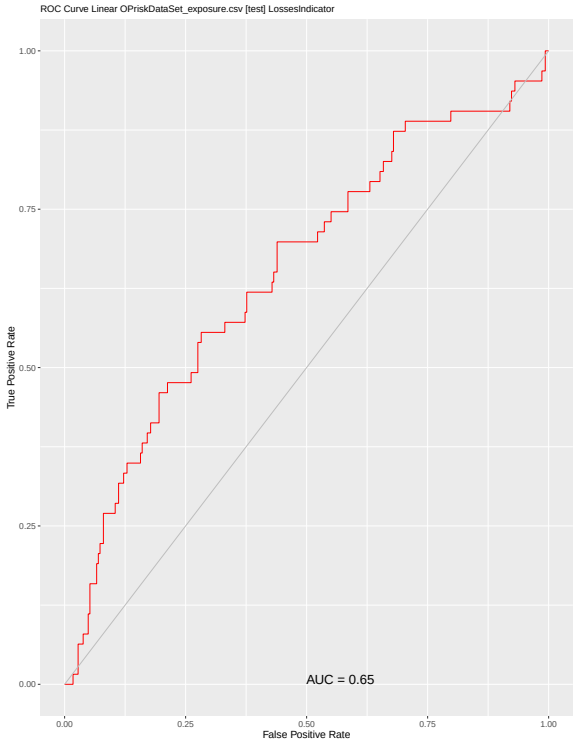


(d) As in (b) but depicting model's performance on the validation set

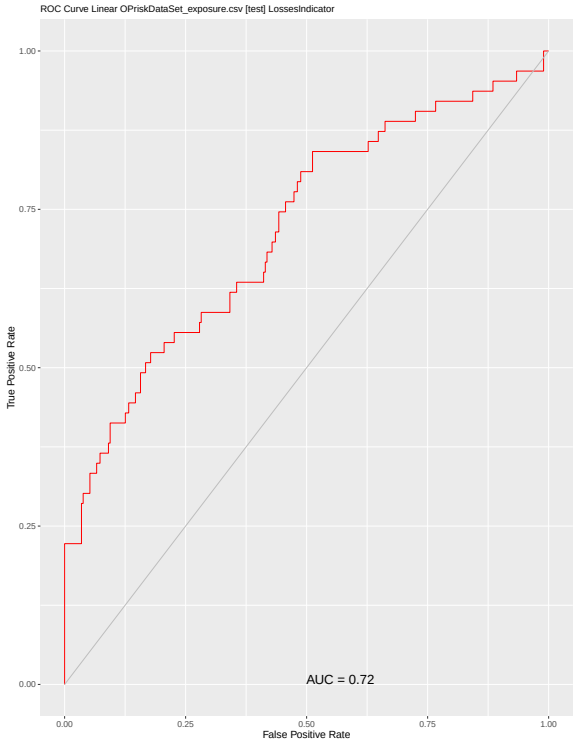
Figure 4.2
Comparison of area under the curve (AUC) for the Poisson, Binomial GLM ROC curves for the Training and Validation datasets

making inferences about the OpRisk frequency processes or system that generated the data. Multimodel inference leads to even more robust inferences, especially from the point of view that the selection of the model used to estimate the mean frequency must serve the ultimate root cause analysis objective of OpRisk control: When calculating the capital requirement using a robust OpVaR measurement technique, at the same time decide take into account as many characteristics of the trading OpRisk dataset as possible, as well consider how the variables interact with each other.

The performance measures here can either tell us that we are going to experience realised losses from OpRisk events more often than we would like, or overcompensate



(a)



(b)

Figure 4.3
(a) ROC Curve for a binomial GLM on the validation dataset (b) As for (a) but on the testing dataset. The area under the curve `auc` is a measure of the performance of the model. A perfect model would have 100% of the area under the curve

for losses that do not materialise. This is an important issue i.e., the fact that the different types of errors have different conscequences for us. Higher risks (extremal events) are normally compensated more than necessary and often than not cause social division within management structures when sectors cannot reconcile or afford the protection.

Conclusion

The selected models, parameter estimates and appropriate measures of precision, reporting and intepretation of ranks of the models and Akaike weights, model-averaged parameter estimates used in prediction, and the ranking of the relative importance of predictor variables and their usefulness have been discussed earlier under subsection 4.8.1. This is primary evidence acting as a pre-cursor to the representation of the scientific proposition through goodness-of-fit tests viz., via confusion matrices and ROC curve analysis.

Verily, the results from the confusion matrix 8.6 for the better fitting Binomial model vs the Poisson case, of a greatly improved Sensitivity of 90.94% and Specificity of 36.51%, and 8.6 showing a slightly more concave ROC curve infer exceptional goodness-of-fit whereby model accuracy has been substantially improved by about 5%. The selected model "AModel", (given a priori issues modeling the scientific proposition 4.6.1) supports the notion that "learning" from looking at evidence in the deep hierarchies of features in the transformation matrix, $e^{\beta_i \mathbf{X}^T}$, acting on the exposure, achieves a "forward-looking" solution to the approximation of the LGE and therefore the random mechanism which generates the loss frequency distribution.

A better estimation model than the traditional "best model" approach focuses several competing models in the estimation process of relationships of variables, and

on the estimation of effect size and measures of precision allows for inferences made in Proposition 4.6.1 to be valid. The minimal K-L information loss using information-theoretic criteria i.e., AIC, AIC_c,QAIC is easily computed using model averaging based on Akaike weights.

CHAPTER 5
GAMLSS: A RATHER GENERAL AND FLEXIBLE FRAMEWORK FOR SEVERITY MODELING OF OPRISK

Introduction

Besides exposure-based issues, an emerging contentious area in developing an OpRisk loss severity model is evidence of heteroscedasticity in the data. One possible explanation for this; alluded to in Committee et al. (2011) LDCE reports in 2002 and 2008, is the consideration of differing OpRisk losses seen across different OpRisk event types and business lines. They reveal that certain OpRisk event types and business lines have distributions with longer tails than others; notwithstanding the common occurrence of using an inappropriate assumption of a constant scale parameter (M. D. Stasinopoulos et al., 2017) for different ET/BL categories of OpRisk that do not capture this tail behaviour properly, hence the need for models accounting for variations around the means of severities (Ganegoda & Evans, 2013). Consistent with earlier reports included in the work by Ganegoda & Evans (2013), and as a solution to account for the problem heteroscedasticity, modeling the variation in location ,scale and shape of the distribution across different business lines and event types should also be allowed in the estimation model.

In this study, an important new algorithm for ORMF's is laid out coupled with data intensive estimation techniques first introduced by Rigby & Stasinopoulos (2005) viz. Generalised Additive Models for location Scale & Shape (GAMLSS) with capabilities to tease out 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. Through a GAMLSS framework, OpRisk loss severities are estimated by a very general class of regression model that allows any distribution for the response variable and all the parameters of the distribution can be modeled as functions of explanatory variables: GAMLSS not only trys to fit the mean of the response variable but is also more appropriately suited to analyse the scaling properties of the severity of losses. We stress that the internal OpRisk data at our disposal provided the motivation for developing the new methodology.

The 2013 OpRisk loss financial data

The OpRiskDataSet_exposure data resulting from the LDCE contains intra-day pnl impacts (`Losses`) of amendment activity during trading over a three month period, beginning 02 January 2013 to 28 March 2013 across primary and secondary market platforms in an investment bank in South Africa. It is collected to be able to explore and analyze whether the tightening of OpRisk controls would result in the lessening of the degree of pnl (`Loss`) impacts. As a plausible approach to exploring this question, the researcher suggests it's best uncovered provided OpRisk loss severities can be predicted accurately and reliably through a model; which is notably difficult to obtain given the heaviness of distribution tails ordered when choosing an appropriate distribution for the three month information provided.

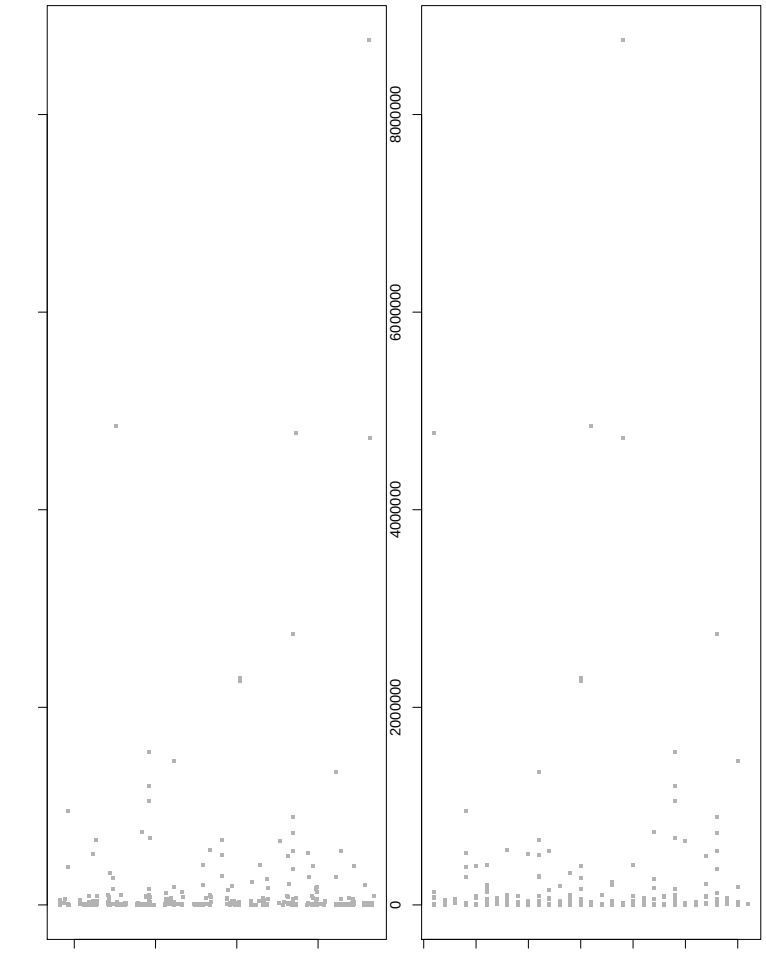
The 3-month dataset in question is found to ideally suited to explore the phenomena in question, as the dataset holds true to the industry standard characteristics displayed by extreme events that disproportionately affect losses, as illustrated in descriptive scatter plots 5.1a and box plots 5.1b, displaying loss severities throughout the day and the month respectively. Figures 5.1a and 5.1b show plots of the pnl

impact, `Loss`, against three appropriately selected explanatory variables chosen for purposes of demonstrating the complexity of the relationship between `Loss` and the explanatory variables. If the dependence of the median of the response variable `Loss` on `UpdatedTime`, `UpdatedDay`, etc.. is nonlinear, the need for a nonparametric smoothing functions, and/or linear or nonlinear interactions in the statistical model is essential for the analysis of the OpRisk loss data viz., the new GAMLSS technique.

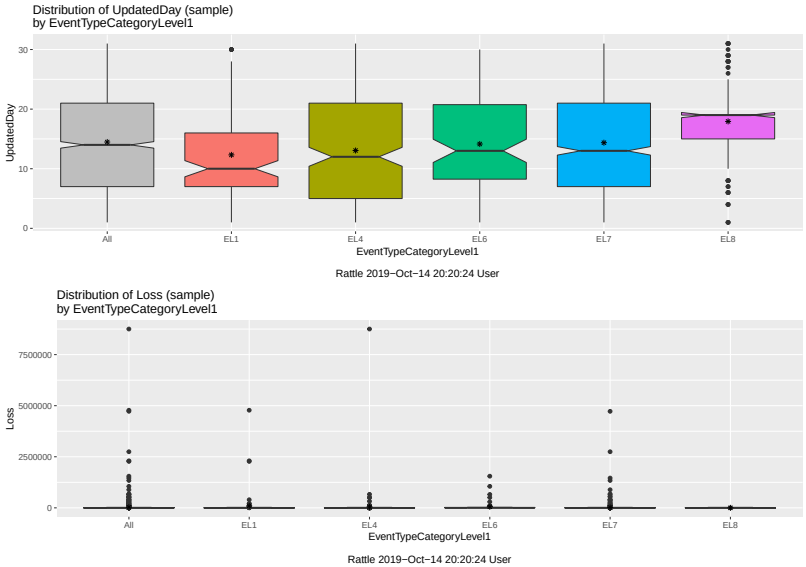
In the first two plots labeled figure 5.1a for the two explanatory variables in the bivariate plots, `Loss` vs `UpdatedTime` and `Loss` vs `UpdatedDay`, there is obvious nonlinear dependence between the mean of the response variable `Loss` and the `UpdatedTime` and `UpdatedDay`, here nonparametric smoothing functions may be needed. There is also a clear indication of the non-homogeneity of the variance of loss, therefore modelling the variance of `Loss` requires a statistical model which caters for the dependency of the variance of its's mean and/or explanatory variables. The first boxplot Figure 5.1b displays how the day in the month `UpdatedDay` varies according to the OpRisk event, and the second how the `Loss` varies according to same. There is clear indication of positive skewness in the distribution of `Loss` depending on the explanatory variable `EventTypeCategoryLevel`, and asymmetrical boxes about the median and long upper and lower whiskers, especially in the first boxplot, emphasizing the need to explicitly model for this.

Modeling selection for severity of OpRisk losses using GAMLSS

A very general class of regression model viz., GAMLSS, offer a solution to the problem of heteroscedasticity due the introduction of a convenient framework to test various distributional assumptions; which doesn't limit the modeling of the location of the distribution as in GLM and other similar frameworks (Cruz et al.,



(a) Plots of `Loss` against explanatory variables



(b) Boxplots of `UpdatedDay` and `Loss` against explanatory variable `EventtypeCategoryLevel1`

Figure 5.1
(a) Plot of the Loss (pnl impact) against explanatory variables UpdatedTime and UpdatedDate. (b) As for (a) against EventTypeCategory.

2015). Essentially, standard general additive (GAM) models emulate scenarios where one can have two or more features interacting, calling some of those features only when specified conditions are met, which leads to a more accurate estimate of location, scale and shape parameters.

In the statistical software package `R`, for example, we can fit *log* `losses` assuming the're normally distributed modelling the standard deviation with the function `gamlss()`, using event type business line combinations as an explanatory variables. We can obtain a model that allows greatly varying losses in one BL/ET combination (i.e., BL_i/ET_j to have a larger standard deviation) compared to the losses in a BL_k/ET_l combination, but both having the same mean (or medial) *log* `Loss`. Stated differently, different quantiles of the distribution scale differently, furthermore if we assume losses follow a 3 or 4 parameter distribution, with shape parameter/s related to some covariates, then we can allow for negative, positive, symmetric skewness, and leptokurtic or platykurtic kurtosis depending on the values of the covariates.

In the GAMLSS framework variables can be modeled as linear and/or non-linear, parametric and/or smooth non-parametric functions of the covariates. On a given set of plausible distributions (to be identified in the following section), we use regression modeling to determine the parameters using the functions pb(), ps(), cy(), tp(), or pvc() introduced by Eilers, Marx, et al. (1996), which are all based on penalised B-splines. Penalised splines are piecewise polynomials defined by B-spline basis functions in the explanatory variable used to let the data determine how the location scale and shape of the distribution of the OpRisk loss severities and ultimately how the losses are estimated. In the process (of defining a close approximation, and by selection of a plausible distribution for the GAMLSS, discovering the effect of OpRisk pnl losses) the coefficients of the basis functions are penalised to guarantee sufficient smoothness (Eilers et al., 1996) i.e., determining how smooth the fitted curve is effectively striking a balance between bias and variance in the estimation of

the curve (M. D. Stasinopoulos et al., 2017).

Sub-problem 1ii.

The existing models in OpRisk measurement for which historical loss distributions are the best predictors of future losses, who by design lack the learning capabilities attributed to artificial intelligence (AI) knowledge systems, hence assume that we do not learn from past losses and therefore suffer from this as a clear limitation. Notably, there are AI systems such as the Kalman filter whose methods 'learn' but are restricted as the best possible estimators in the *linear* dynamical systems case, compromising their suitability to cases involving more complex dynamical systems. This is problematic for "predictable" risk types due to their model's practice of undercapitalising known risks before occur, and overcapitalising for risks after the losses materialise, creating inappropriate capital estimates (Group et al., 2013).

Furthermore, and besides extensions and generalisations viz., extended/unscented Kalman filter developed to account for the nonlinear case, these concerns motivate for the development of GAMLSS models which account for variations around the mean of severities in order to account for heteroscedasticity and inappropriate assumptions; evidenced by M. D. Stasinopoulos et al. (2017), Rigby & Stasinopoulos (2010), M. Stasinopoulos, Rigby, & Akantziliotou (2008) and others, of a constant scale and shape parameters for different ET/BL categories of OpRisk that do not capture tail behaviour properly.

Sub aim (ii): Exposure-based OpRisk GAMLSS models

To quantify OpRisk loss severity through GAMLSS modeling, thus further contributing toward a framework for OpRisk which is "forward-looking" due to it's learn-

ing capabilities, effectively replacing historical loss severity curves obtained from historical loss counts by models accounting for variations around mean severities (μ), dispersion about the mean (σ), and the shapes of the distribution induced by explanatory variables. Engle (1982) was the first to model σ offering this as a solution to the problem of heteroscedasticity and the inappropriate assumption of a constant variance implicit in traditional modeling uses up to now, which is inconsistent with the wide spectrum of OpRisk loss events in the dataset i.e., due to the different ET/BL loss categories of OpRisk. This has since created a whole industry of related models in finance proposing models for σ that capture tail behaviour properly, especially for those more flexible options for distribution fits. Heavier tailed distribution fits with appropriate goodness of fit ideally reveal information for the risk practitioner of whether to go against the odds of BL/ET loss combinations known to be more or less variable comparatively, and are therefore desirable to assist in taking preemptive preventative measures.

Model selection strategy in GAMLSS framework

Employing M. D. Stasinopoulos et al. (2018)'s model: Let $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \Lambda\}$ represent an expansion of equation 2.37. The form and components of $\mathcal{M}$ is general and implies that the distribution be in parametric form; Voudouris, Gilchrist, Rigby, Sedgwick, & Stasinopoulos (2012) and M. D. Stasinopoulos et al. (2017) define each component respectively as follows:

1. $\mathcal{D}$ specifies the distribution of the response variable,
2. $\mathcal{G}$ specifies the set of link functions,
3. $\mathcal{T}$ specifies the terms appearing in all the predictors for μ, σ, ν and τ,
4. Λ specifies the smoothing hyper-parameters which determine the amount of smoothing in the $h_{kj}()$ functions of equation 2.37.

Component $\mathcal{D}$: Selection of the distribution

An appropriate distribution fit based on residual diagnostics is suggestive of estimating and potential inferences regarding casual relationships between the response variable (this case being the OpRisk losses) conditional on explanatory variables. Given a plethora of competing plausible distribution fits that potentially fit the data adequately, the ultimate choice of a particular distribution to go with is non-trivial. From the marginal analysis of the OpRisk Loss dataset a set of plausible distributions can be identified; specifically using methods proven to satisfy the condition that the continuous distributions in which the range R_Y has been expanded to include some discrete values with non-zero probabilities; notably $R_Y = [0, \infty)$, which indicates zero losses are included.

Additionally:-

* The variate takes values within to the appropriate range viz., $[0, \infty)$;
* The distribution is relevant because in has an explicit p.d.f, c.d.f and inverse c.d.f, explicit moment based measures of location, scale, skewness and kurtosis (i.e. population mean, standard deviation, γ_1, γ_2 (Rigby, Stasinopoulos, Heller, Gillian, De Bastiani, & Fernanda, 2017));
* Explicit centiles and centile based measures viz., median, semi-interquartile range, skewness and kurtosis ($\gamma, \text{st}_{0.49}$ resp.);
* Continuity of $f(y|\mu, \sigma, \nu, \tau)$ w.r.t y and it's derivatives w.r.t μ, σ, ν, τ;
* Allows for flexibility in specifying the distribution of severity and also allowing for the modelling of distribution parameters as function of explanatory variables

3-parameter distributions on $[0, \infty)$

We begin model selection using a three parameter distribution otherwise known as the Lambda, Nu and Sigma (LMS) method (M. D. Stasinopoulos et al., 2018) selected from the family of zero adjusted distributions i.e., a distribution adjusted to fit on zero and the positive real line $[0, \infty)$, and then move to its extensions viz., four distribution parameters. Of the zero adjusted distributions, only the Zero adjusted gamma ZAGA(μ, σ, ν) and the zero adjusted inverse gamma ZAIG(μ, σ, ν) can be fitted explicitly[1] in GAMLSS, therefore we begin with this in mind. The parameters μ, σ and ν in this case, are the approximate mean, approximate coefficient of variation and skewness parameters.

Furthermore, the ZAIG and ZAGA distributions are befitting as most OpRisk loss events do not draw an associated realised pnl (loss) impact, similar to phenomena commonly associated with these distributions such as the amount of insurance claim over a period; therefore there being a good chance of zero pnl impact; nevertheless, also spanning situations for OpRisk loss events that do have associated pnl impact such that losses are defined on zero and the positive real line $[0, \infty)$. The aforementioned "mixed" distribution are a special case of finite mixtures of continuous and discrete distribution specific to the `gamlss.family`, but there are extra mixed distributions that can be defined (using the package `gamlss.inf`) if need be, or even new families of distributions created at will.

Generalized beta type 2 distribution (GB2)

Introducing the fourth parameter τ for flexibility in modelling for leptokurtosis or platykurtosis, targeting explicitly defined continuous mixture distributions results in a choice of the only creation of it's kind viz., the generalized beta type 2 distri-

[1] Measures of functions that are obtained through closed form mathematical functions i.e., not requiring numerical methods

bution, denoted by GB2(μ, σ, ν, τ). The GB2 distribution is the only four parameter distribution generated by univariate transformation satifying all the desirable properties above for fitting a GAMLSS to estimate the nature of OpRisk loss severity. The GB2, sometimes called the generalised beta of the second kind, or the generalised beta-prime distribution is defined by assuming the tranformed variable $(Y = g(Z))$ Z, has an F distribution with degrees of freedom 2ν and 2τ, i.e., $F \sim F_{2\nu,2\mu}$, where $Z = (\tau/\nu)(Y/\mu)^{\sigma}$.

Given $X = x$, Y (the `Loss` severity is the target variable) is modelled here by a generalized beta type 2 GB2(μ, σ, ν, τ)[2] distribution is defined by:

$$\begin{aligned} f(y|\mu,\sigma,\nu,\tau) &= |\sigma| y^{\sigma\nu-1}\{\mu^{\sigma\nu}\mathrm{B}(\nu,\tau)[1+(y/\mu)^{\sigma}]^{\nu+\tau}\}^{-1} \\ &= \frac{\Gamma(\nu+\tau)}{\Gamma(\nu)\Gamma(\tau)}\frac{|\sigma|(y/\mu)^{\sigma\nu}}{y[1+(y/\mu)^{\sigma}]^{\mu+\tau}} \end{aligned} \tag{5.1}$$

The choice of distribution for the OpRisk severity `Loss` variable is based on it's simplicity as it also qualifies for being fitted explicitly in GAMLSS for the mean and median of the response variable. Some scenarios reveal important special cases and limiting scenarios, such as the `Pareto2`(ν, τ) obtained by setting $\sigma = 1$ and $\nu = 1$ in equation 5.1. "Paretian type" tail distributions: A four way category split of distributions based on the heaviness of their tails is an important method for choosing the approprite tail in the development and selection of the distribution with tails correspond to mild, slow, wild (Paretian) and extreme (heavier than Paretian) randomness of Mandelbrot (1997), for observations on a variable, particularly true with emphasis of the need for a better understanding of measures of OpRisk such as VaR estimates, which are heavily used in this thesis.

[2]The GB2 adjusts the obove density $f(y|\mu, \sigma, \nu, \tau)$, resulting from the condition $y > 0$, where $\mu > 0$, $-\infty < \sigma < \infty$, $\nu > 0$ and $\tau > 0$. The mean and the variance of Y are given by $E(Y) = \mu\mathrm{B}(\nu + \frac{1}{\sigma}, \tau - \frac{1}{\sigma})/\mathrm{B}(\nu, \tau)$ for $-\nu < \frac{1}{\sigma} < \tau$ and $E(Y^2) = \mu^2\mathrm{B}(\nu + \frac{2}{\sigma})/\mathrm{B}(\nu, \tau)$ for $-\nu < \frac{2}{\sigma} < \tau$, See Stasinopoulos, Rigby, and Akantziliotou (2008).

Also setting $\sigma = 1$ in equation 5.1 gives a form of Pearson type VI distribution, $\tau = 1$ gives the `Burr`III (or `Dagum`) distribution, $\nu = 1$ gives the `Burr`XII (or `Singh-Maddala`) distribution, and lastly setting $\nu = 1$ and $\tau = 1$ gives the `log-logistic` distribution.

Component $\mathcal{G}$: selection of the link functions

Also for $k = 1, \ldots, 4$, let $g_k(\cdot)$ be known monotonic link functions relating the parameters θ_k to explanatory variables (predictor η_k) through extended semi-parametric additive models given by:

$$\begin{aligned} g_1(\mu) &= \eta_1 = \mathbf{X}_1\beta_1 + \sum_{j=1}^{J_1} h_{1j}(\mathbf{x}_{1j}) \\ g_2(\sigma) &= \eta_2 = \mathbf{X}_2\beta_2 + \sum_{j=1}^{J_2} h_{2j}(\mathbf{x}_{2j}) \\ g_3(\nu) &= \eta_3 = \mathbf{X}_3\beta_3 + \sum_{j=1}^{J_3} h_{3j}(\mathbf{x}_{3j}) \\ g_4(\theta) &= \eta_4 = \mathbf{X}_4\beta_4 + \sum_{j=1}^{J_4} h_{4j}(\mathbf{x}_{4j}) \end{aligned} \tag{5.2}$$

where for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, J_k$; $\mathbf{X}_k$ is a matrix of explanatory variables and $\beta_k^T = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{J_k'k})$ is a parametric vector of length J_k' to be estimated, $\mathbf{x}_{ik}$ a fixed known design vector of length J_k'' and h_k are non-linear smooth functions of explanatory variables of $\mathbf{x}_{ik}$ (Eilers et al., 1996). The explanatory values x_{jk} are assumed to be fixed and known and the univariate function h_{jk} is an additive smooth parametric function assumed to have continuous first and second order derivatives. If for $k = 1, 2, 3, 4$, $J_k = 0$, then the GAMLSS model 5.2 reduces to a non-linear parametric model. If in addition, $h_k(x_{ik}, \beta_k) = \mathbf{x}_{ik}^T\beta_k$ for $i = 1, 2, \ldots, n$ and $k = 1, 2, 3, 4$, then equation 5.2 reduces to a linear parametric model (Rigby, Stasinopoulos,

Heller, & De Bastiani, 2017).

Component $\mathcal{T}$ and Λ: selection of the terms and smoothing parameters in the model

Given that a set of plausible distributions have been identified, let χ be the selection of all terms available for consideration. Their parameters are modelled as regression models. In particular the non-linear parameter vectors β_k and the non-parametric functions h_{jk} for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, 3, 4$ in equation 5.2 are estimatied by maximizing the penalised log-likelihood as a way of understanding how the location, scale, skewness and kurtosis parameters of the loss severity distribution are affected by the explanatory variables.

Proposition 5.4.1:

EBOR OpRisk treatments through GAMLSS modeling frameworks are a complementary flexible model designed to "learn" from data[3] *to provide a forward-looking solution to LDA loss severity modeling problem and account for variations around the mean of loss severities, notwithstanding favouring more flexible heavy tailed distributions with appropriate goodness-of-fit, thereby affirming combinations of "odds against" one's BL/ET mix loss variations against another.*

Modeling of the GB2 distribution parameters

The parameters μ, σ, ν, and τ of the GB2 distribution are modelled as regression models using semi-parametric penalised B-spline functions, extended to incorporate non-linear parametric and/or non-parametric smooth functions x.

[3]By virtue of Rigby & Stasinopoulos (2010)'s assertion of the inclusive GLM (a machine learning technique) the GAMLSS also incorporates prior weights.

Step-wise selection of explanatory variables

M. D. Stasinopoulos et al. (2017) proposes several different strategies that could be applied for model selection of the terms used to model the location, scale and shape parameters (μ, σ, ν, and τ); however, the procedure in the analysis as outlined in M. D. Stasinopoulos et al. (2018) and Voudouris et al. (2012), is comprised of the function `stepGAICAll.A`. This strategy is chosen by selecting all terms using the GAIC for all distribution parameters by a forward, backward or stepwise procedure, assuming the particular response distribution function stated in M. D. Stasinopoulos et al. (2017) and Rigby et al. (2017). The final model may contain different combinations for μ, σ, ν, and τ.

This is essentially conducted in R, using first a GAIC selection procedure to select an appropriate model for μ_i with σ_i, ν_i, & τ_i fitted as constants, dropping unnecessary terms, selecting and adding additive terms and smoothing terms in an iterative forward and backward selection procedure (M. D. Stasinopoulos et al., 2018). In order to do this a formula is created containing all the linear main effects and second-order interactions plus smooth function of explanatory variables. This is a model building function followed by arguments appropriate to the task. In the software package R, the specific argument is achieved through a `scope` statement which defines the range of models or "ensemble" for examining the step-wise search, whereby the function `FORM` consists of a list containing the upper and lower formulae, as demonstrated in see Appendix 8.11.4. The final model will consist of choice of possibly different subsets of terms for μ_i, σ_i, ν_i, & τ_i.

Model selection

Using the `stepGAICAll.A`() function in GAMLSS. We compare the models that best fit the 2013Q1 dataset of OpRisk losses, pending and near misses, conditional on the available data. From the listed table of plausible distributions; Table 5.3, the GB2() distribution is the best fit for the data based on the model selection strategy discussed earlier in section 5.5.1. Figure 5.2 shows summary statistics of the GB2() distribution model.

The estimated GAMLSS-based "best model"

The empirical GAMLSS-based model $Y \sim \text{GB2}(\mu, \sigma, \nu, \tau)$ where $Y = \log(\text{Losses due to OpRisk events})$ and the fitted models for μ, σ, ν and τ, given by equation 5.2 for the chosen model is:

$$\begin{aligned}
\log(\mu) &= \texttt{UpdatedDay} + \texttt{Desk} + \texttt{CapturedBy} + \texttt{TradeStatus} + \texttt{TraderId} + \texttt{Instrument} \\
&+ \texttt{Reason} + \texttt{EventTypeCategoryLevel1} + \texttt{Exposure} + \text{pb}(\texttt{CapturedBy}) + \text{pb}(\texttt{TraderId}) \\
&+ \text{pb}(\texttt{TradeStatus}) + \texttt{TraderId:Reason} + \texttt{UpdatedDay:Desk} + \texttt{Desk:Exposure} \\
&+ \texttt{Reason:EventTypeCategoryLevel1} + \texttt{TradeStatus:EventTypeCategoryLevel1}, \\
\sigma &= \texttt{BusinessLineLevel1}, \\
\log(\nu) &= 1, \\
\log(\tau) &= \texttt{Instrument}
\end{aligned} \tag{5.3}$$

Model diagnostics and checking

The model is built in a way of understanding how the location, scale, skewness and kurtosis i.e., $(\mu_i, \sigma_i, \nu_i, \ \& \ \tau_i)$ parameters of the distribution of severity of OpRisk losses are affected by the explanatory variables viz., the `UpdatedDay`,

```
Family:  c("GB2", "Generalized beta 2 (i.e. of the second kind)")
Fitting method: RS()

Call:  gamlss(formula = Losses ~ UpdatedDay + Desk + CapturedBy +  TradeStatus +
TraderId + Instrument + Reason + EventTypeCategoryLevel1 + Exposure + pb(CapturedBy)
+ pb(TraderId) + pb(TradeStatus) + TraderId:Reason +  UpdatedDay:Desk +
Desk:Exposure + Reason:EventTypeCategoryLevel1 + TradeStatus:EventTypeCategoryLevel1,
sigma.formula = ~BusinessLineLevel1,
nu.formula = ~1,
tau.formula = ~Instrument,
family = GB2, data = crs$training,
mu.start = NULL, sigma.start = NULL, nu.start = NULL, tau.start = NULL, n.cyc = 1700,
trace = FALSE)

Mu Coefficients:
(Intercept)                          UpdatedDay                                Desk
11.3274048                           0.0769694                           0.2639089
CapturedBy                         TradeStatus                             TraderId
-0.1666376                          -1.9155606                           0.2107559
Instrument                              Reason          EventTypeCategoryLevel1
0.0323856                           -0.1054440                          -1.7661357
Exposure                     pb(CapturedBy)                      pb(TraderId)
0.0018252                                NA                                NA
pb(TradeStatus)                TraderId:Reason                   UpdatedDay:Desk
NA                                  -0.0309107                          -0.0118501
Desk:Exposure       Reason:EventTypeCategoryLevel1  TradeStatus:EventTypeCategoryLevel1
-0.0002197                           0.0619507                           0.4030053
Sigma Coefficients:
(Intercept)  BusinessLineLevel1
     0.7302              0.1251
Nu Coefficients:
(Intercept)
        2.7
Tau Coefficients:
(Intercept)   Instrument
-2.11476      0.02488

Degrees of Freedom for the fit: 22.28 Residual Deg. of Freedom   236.7
Global Deviance:     5910.89
            AIC:     5955.44
            SBC:     6034.67
```

Figure 5.2
Summary statistics of fitted distribution GB2(μ, σ, ν, τ), the "Generalized beta type 2 (i.e. of the second kind)" using fitting method: RS()

distribution	df	AIC
Generalized Beta type 2 GB2	22.27608	5955.438
Zero Adjusted Inverse Gamma (ZAIG)	13.00028	5964.959
Zero Adjusted Gamma (ZAGA)	25.13312	6066.610

Figure 5.3
Summary of the fitted models for the OpRisk 2013Q1 data, showing the effective degrees of freedom (df) used in the model and the AIC(k =length of loss variable).

`UpdatedTime`, `CapturedBy`, `TraderId`, `Reason`, `Desk`, `Instrument`, `TradeStatus`, `EventTypeCategoryLevel1`, `BusinessLineLevel1`, `Exposure`.

Diagnostic plots

Figure 5.4 displays the (normalized quantile) residuals carried in the diagnostics for the fitted model GB2(). They are computed by finding the equivalent standard deviate for each response observation routinely plotted in four realisations of the quantile residuals. This aspect of the data can also be shown through the use of a worm plot (Voudouris et al., 2012).

Normalized quantile residuals

The top two panels (Panel (a) and (b)) plot the residuals against the fitted values of μ and against the index respectively. Note in panel a the plotting precedure plots two observations differently in the x axis from the rest; an unnecessary complication that is brought about by due to the nature of the distribution (asymmetrical and bounded) introducing spurious patterns causing difficulty to distinguish between large residuals and outright outliers. The realisations in panel (b) are ignored due to the inconsistent patterns, while the fitted values in panel (b) show a random scatter around the horizontal line at 0.

The bottom two panels (Panel (c) and (d)) provide a kernel density estimate and

normal QQ plot of normalised quantile residuals respectively. Looking a the summary statistics of the quantile residuals below, their mean is close to zero, their variance close to one, their co-efficient of skewness near zero and kurtosis is close to 3. The residuals statistics suggest an approximate normal distributional shape as portrayed by the kernel density plot, furthermore the Filleben correlation coefficient is near 1, as expected for an adequate model.

```
*******************************************************************
          Summary of the Quantile Residuals
                               mean   =  0.02081528
                           variance   =  0.9648917
                   coef. of skewness  =  -0.2544155
                   coef. of kurtosis  =  2.485503
Filliben correlation coefficient  =  0.9945067
*******************************************************************
```

The resulting QQ plot is displayed in the bottom right hand corner of Figure 5.4. The largest positive residuals are in the upper tails of the plot corrsponding to the $(2, 2)$ cell: Losses almost always lessen in severity if control mechanisms are continually tightened. Figure 5.4 shows evidence of large negative residuals as well as large positive residuals. Very few of both negative and positive residuals are individually insignificant, and the actual contingency table cells represented in the left tail of the QQ plot remains relatively contant with each realisation of the quantile residuals, preserving the overall pattern across realisations. The quantile residual plot shows that there are not enough small counts in the contingency table to conclude incompatibility with the prescribed model.

Worm plot

Figure 5.5 shows the results of a single worm plot of the entire residuals: The worm plot allows the detection of inadequacies in the model globally or within a specified range of an explanatory variable (M. D. Stasinopoulos et al., 2017). In this

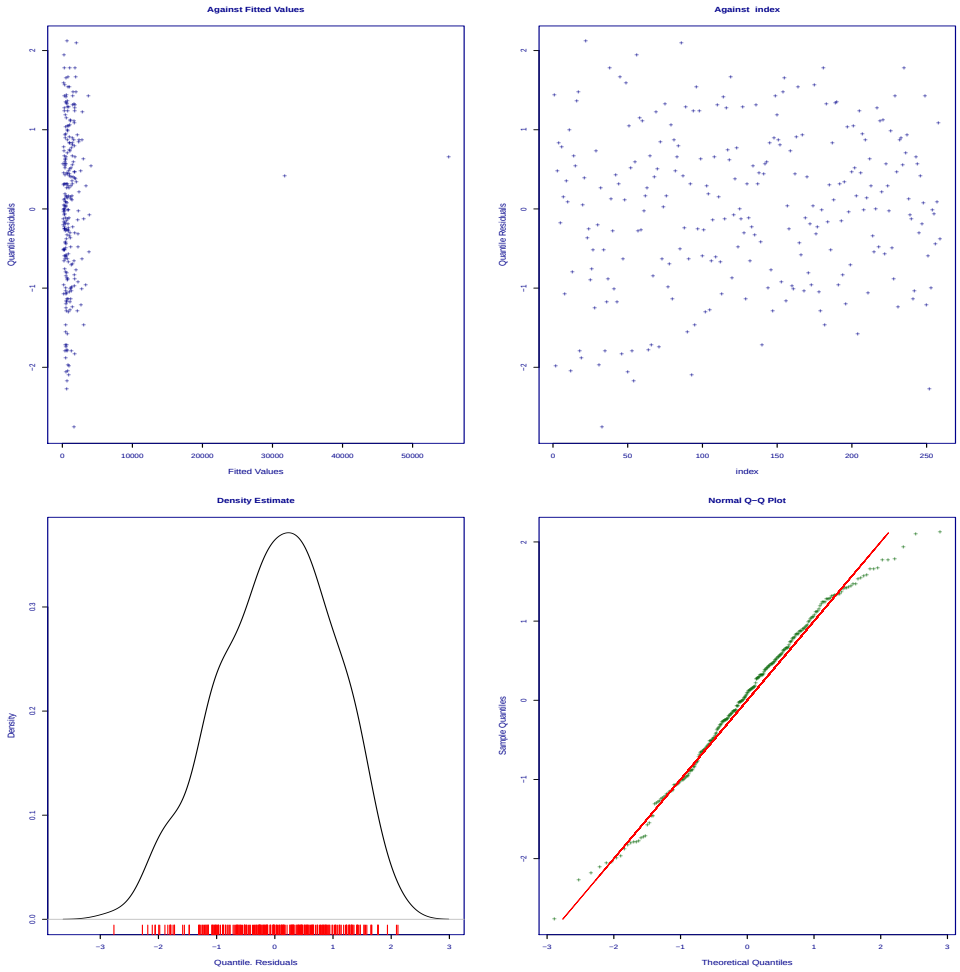


Figure 5.4
The result of the plot displays (normalized quantile) residuals from model GB2(μ, σ, ν, τ), the top-left panel plots the residuals against the fitted values of μ, the bottom-left panel provides a kernel density estimate and the normal QQ plot for the residuals in bottom-right panel. The last panel results are meaningless

.

case it is a detrended version normal $Q-Q$ plot (i.e. where the line in a $Q-Q$ plot has been transformed horizontally) shown in the bottom right hand panel of Figure 5.4, used as a diagnostic tool for checking for the presence of skewness or kurtosis, or any indication of nonnormality in the residuals (within the corresponding range of the explanatory variable, i.e.,UpdatedTime, exposure).

All observations in the worm plot of the residuals of the fitted model lie within the confidence bands (i.e., within the "acceptance region" inside the two elliptic curves) and are close to the horizontal line, indicating that the $GB2$ distribution is adequate in modeling the distribution, and therefore provides an appropriate fit to the data. The $Q-Q$ plot in the right bottom corner of Figure 5.4 and the worm plot in Figure 5.5 indicate that there is no reason to dispute the adequacy of the fit.

Intepretation of the fitted model

The fitted functions for all the terms for μ are shown in Figure 5.6 and the fitted terms for σ and ν in figure 5.7 for distribution parameter of Equation 5.3. The fitted nonparametric functions cannot be described simply in mathematical form (M. D. Stasinopoulos et al., 2017), however one of the benefits is their representation in the form of the displayed plots: Note that $\log(\tau) = 1$ which indicates that the distribution of the loss severities appear to be highly positively skewed (fat tailed).

Dynamics of μ, σ and τ distribution parameters

Dynamic plots in figure 5.6 show that the predictor $\eta = \log(\mu)$ for median pnl impact rises linearly with `UpdatedDay`, `Desk`, `TraderId`, `Instrument`, & `Exposure`, it also falls linearly with `CapturedBy`, `TradeStatus`, `Reason`, & `EventTypeCategoryLevel1`. This effectively means more greater Oprisk losses are associated with operational

Diagnostic plot showing adequacy in modeling the $GB2$ distribution

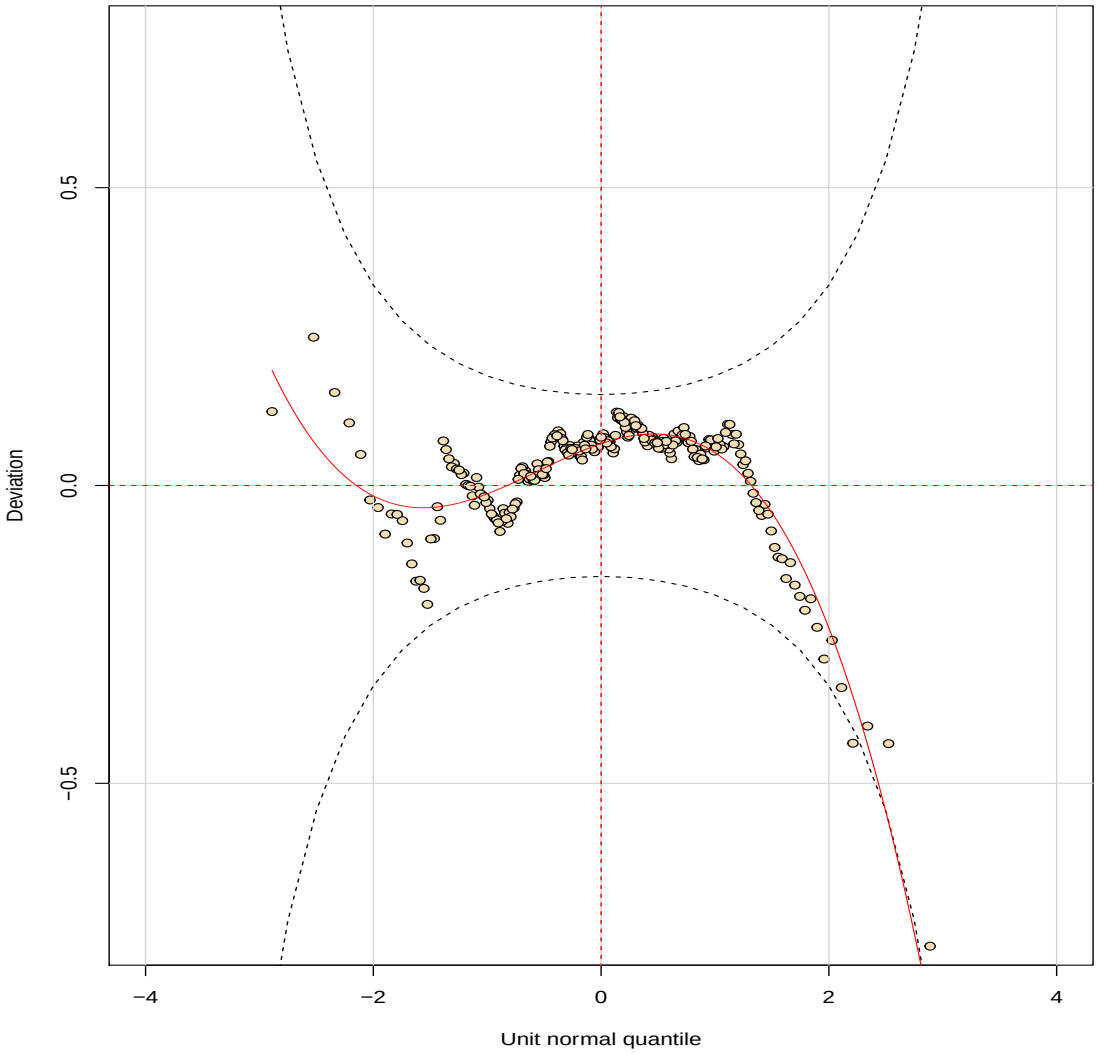


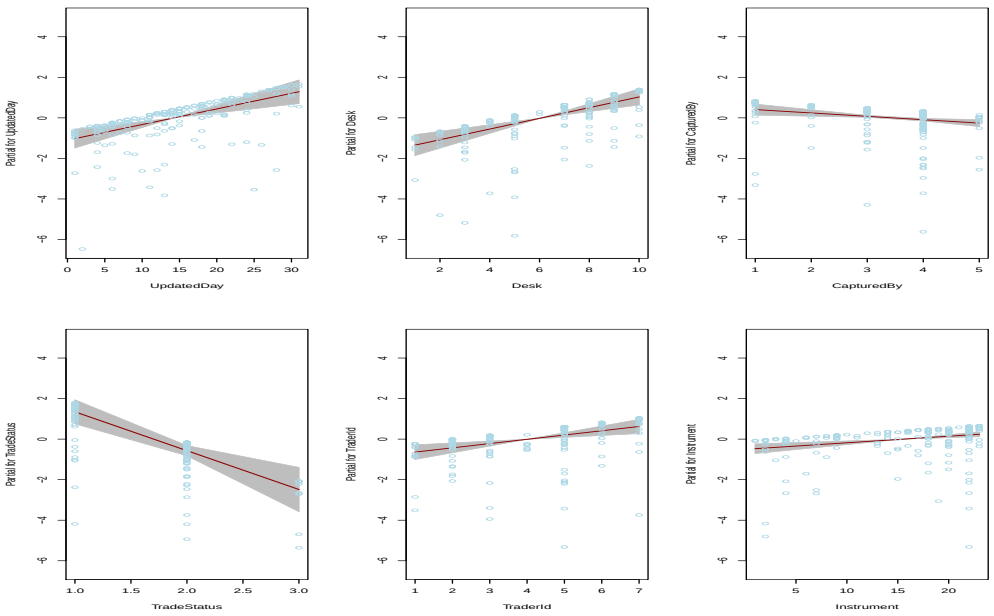
Figure 5.5
Worm plot from the generalised beta type 2 fitted model $GB2$

transactions carried out in the latter part of the month as well as those with higher degrees of exposure. FO and MO/BO Personnel higher up in responsibility also incurr more severe losses as expected.

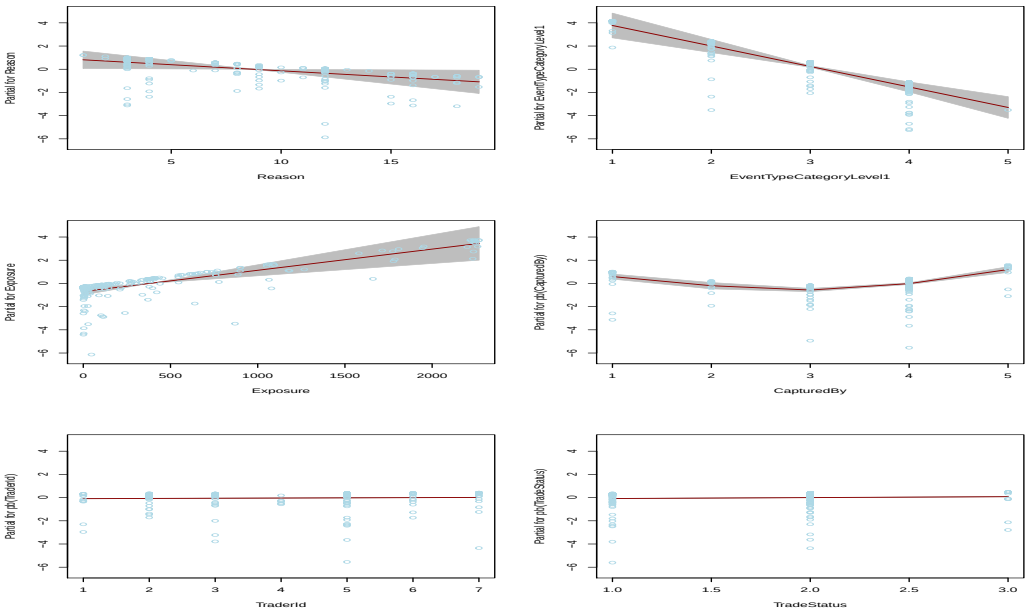
Nonlinearity of the characteristics of MO/BO personnel with the mean pnl impact can be observed in the plot displaying pb(`CapturedBy`). The quadratic shape alludes to the finding in Chapter 3, in the bottom right mosaic plot of Figure 3.6 in subsection 3.11.2; that most losses are "captured" by or attributed to MO personnel who are mandated to perform trade amendments as an integral part of their designated duties, followed by the dip in the curve as you would expect as less responsible MO/BO personnel get involved. Henceforth and not surprising the least experienced and mostly "unauthorised" MO/BO personnel are noted to inflict the more impactful of the loss severities as demonstrated in the upward trend toward the extreme end of the curve. This rising trend indicates that less experienced BO personnel are prone to higher OpRisk hazard as would be expected.

In the assessment of the regression model it is important to note how due to the flexibility of the GAMLSS model, the approximate coefficient of variation, σ, of the $GB2$ distribution is affected by the `BusinessLineLevel1`. The σ of the GB2 distribution might be intepreted as the measure of the variation introduced due to the sheer scale of transactions inherent in one business line over another known to scale differently. This is illustruted in the introduction of section 5.2.1, and is consistent with subject-specific expertise.

The left-most graphic in Figure 5.6 shows the partial for the τ.



(a) Partials for the μ (median) distribution parameters of Equation 5.3



(b) As above

Figure 5.6
Dynamic plots for the GB2 distribution parameters of the fitted terms for the model.

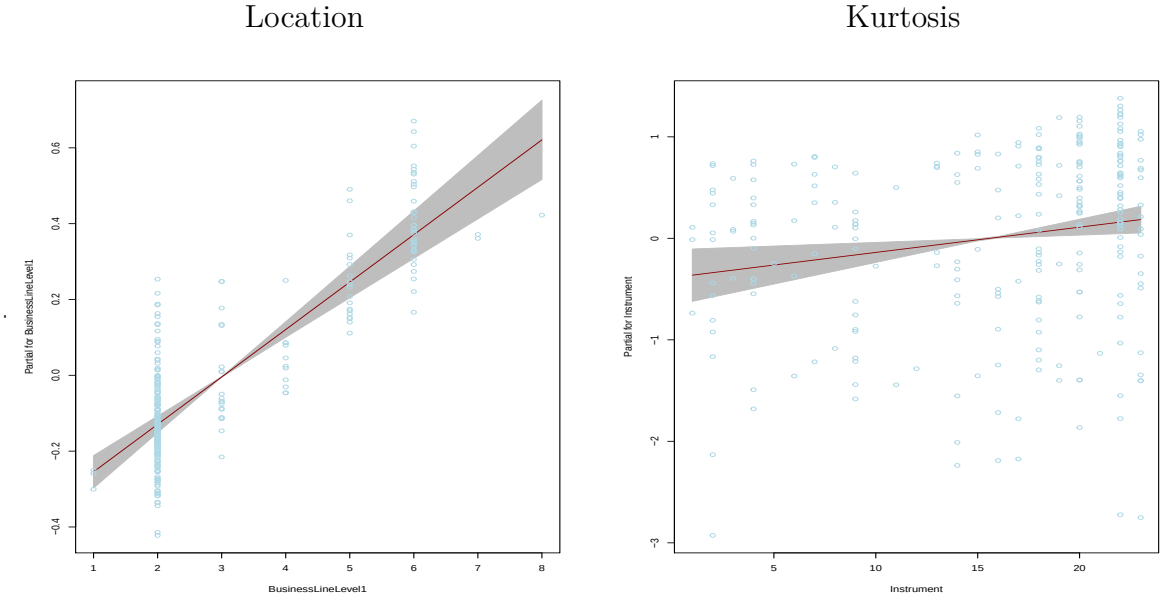


Figure 5.7
simple plots of the additive contribution the `BusinessLineLevel1` term and the `Instrument` term in modelling the variation of exposure (σ), and the kurtosis (τ)

Pdfs and the fitted GB2 distribution

Figure 5.8a gives the fitted distributions of the severity of OpRisk Losses for: Observations 2,consisting of HFLS OpRisk type associated with system related updates[4] with exposure of $239\frac{1}{4}$ days and $(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}) = (978.3, 3.88, 14.88, 0.13)$; observation 136 a moderate pnl loss associated with early terminations, closed out deals or early delivery on trades with 286 day exposure and $(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}) = (1500.12, 3.42, 14.88, 0.16)$; observation 86, a LFHS Oprisk loss type of which exhibiting the extreme OpRisk loss severities, noticing the trend in differing vertical scales for the plots, which resulted due to economic adjustments (exposure of 12 days after the amendment) performed on a deal following a request from a client, and $(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}) = (2045.97, 2.67, 14.88, 0.21)$.

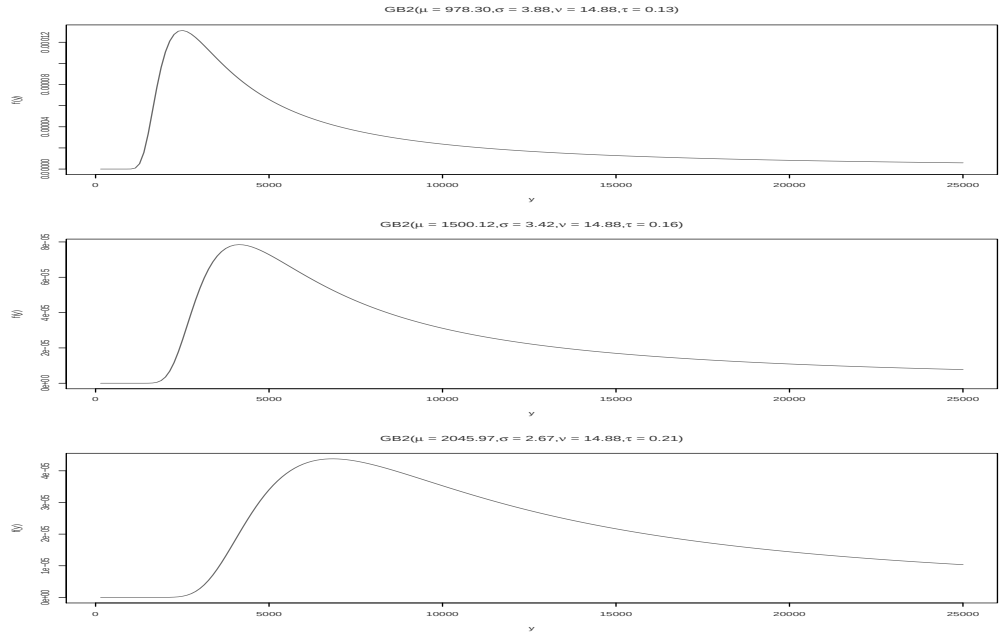
[4]This finding is consistent with the analysis in Chapter 3, Subsection 3.11.2, of the mosaic plot Figure 3.7 and in the contingency table, Table 3.8

Henceforth, due to the uses of GAMLSS the shape of distribution of the severity of OpRisk losses can be estimated early on in the lifecycle of a deal, implied in probabilistic statements that can be made through the application of the GAMLSS model using the GB2 distribution. That is to say, more accurately determine future risk appetite indexed on the deal's OpRisk loss type and adjust dynamically the loss severity mitigation controls. Figure 5.8b also shows the fitted model for the selected OpRisk loss severities superimposed on a smoothed scatterplot diagram as a way of demonstrating the fexibility of the distribution. These empirical findings, as implied by the intepretability of the GB2 distribution which now enters the literature in modelling of OpRisk Loss severities, reveal "odds against" knowing if one banks losses are more or less variable than another.

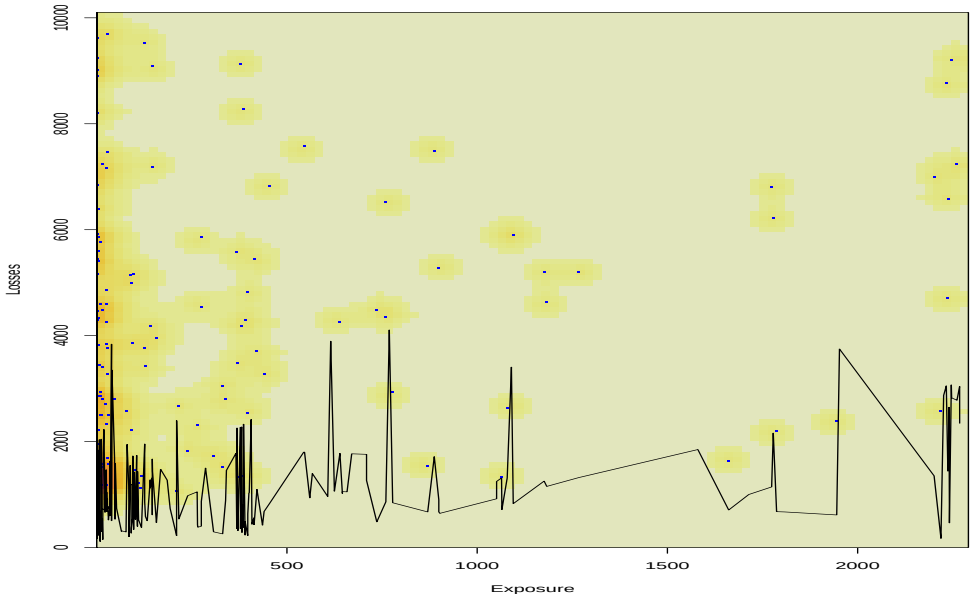
Conclusion

Using the flexibility of the GAMLSS framework, the four components for model selection: $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \Lambda\}$, for distributions, *link functions, additive terms*, and *smoothing parameters* respectively, explores an iterative process based on a *forward, backward*, or *stepwise* procedure for finding the coefficients of a type of non-parametric regression model using the `stepGAIC()` function; alternatively, one where no assumption of the functional relationship is made but by rather letting the data define the functional form i.e., let the data speak for itself. This is the basis for it's power and the clear advantage presented is the better exploration of the statistical properties of the OpRisk loss severity and revision to the dominant modeling paradigm of the traditional "ARMA" (auto regressive moving average) statistical model.

The afore-mentioned (AIC) model selection criteria coupled with model diagnostics tools in the residuals of the fitted models for the GAMLSS i.e., normalised quantile residuals and the *worm plot* verify the $GB2()$ as an adequate GAMLSS



(a) The pdfs for HFLS, LFHS OpRisk loss types



(b) The fitted GB2 distribution to the data

Figure 5.8
Plots of the generalised beta type 2 fitted model $GB2$ with (a) HFLS and LFHS OpRisk losses in increasing orders of magnitude (b) The fitted model in the deviation range $(0, \text{R}10,000)$ include all the points in the scatterplot.

model fit which outperforms its rivals given the reported information criteria, standard normal true residuals and the regions identified for an appropriate fit by the worm plot. The GB2() distribution provides a flexible model for OpRisk loss severity exhibiting variations around the mean severities and appropriate estimations of scale and shape parameters due to the wide spectrum of loss events to be modeled, and therefore accounts for heterogeneity in the model.

The GB2() distribution supports Proposition 5.4.1; identifying a number of statistical techniques who for the first time eliminate the variance-bias dilemma wherein generality always came at a high cost with respect to stability. That is to say, by applying the semi-parametric GAMLSS model $\mathcal{M}$, we eliminated the theoretical concern involving highly unstable parameter fitting and consequently OpVaR estimation, and re-examined the the argument of of infinite variance (Voudouris et al., 2012).

CHAPTER 6
DATA AUGMENTATION AND CAPITAL ADEQUACY

Introduction

After monitoring, profiling and integration, the known HFLS & LFHS data management dilemma needs to be overcome as a last stage, which requires some innovation. In the OpRisk context, problems due to data sensitivity concerns and cost constraints have limited the study to only three months of available data, which by implication means increasing the number of data points in some way. Moreover, ML algorithms are data driven therefore the more data the better the model. One way of adding to new data points to the three month data is by *data augmentation* derived through heuristics from the internal sources of an institution using an extrapolation technique i.e., based on heuristics the relevant fields are updated or provided with values.

Overview of OpRisk generation

It is important to note how OpRisk manifests itself: The causes and sources of operational loss events are observed (risky) phenomena associated with operational errors and are wide ranging (King, 2001). By definition, the occurrence of a loss event is due to PnL volatility from a payment, settlement or by being on the wrong end of some legislative ruling within the capital horizon, over a time period of usually one year. (Einemann et al., 2018). These operational loss events are almost always initiated at the dealing phase of the investment banking process; which more often than not implicates front office (FO) personnel who bear the brunt of responsibility

for the loss. As such, PnL volatility is not only related to the way firms finance their business, but also in the way they *operate*.

In operating practice, with all controls in place one assumes that in observing or following instructions, one is consciously analysing and accurately executing tasks based on information made available or at their disposal. Their assumption is that measures can be trusted to ward off undesirable or risky operating practices, however persistent operational risk losses from undesirable sub-conscious events indicate possible (human or non-human) faults in information processing, which we are not consciously aware of, or faults transpiring sub-consciously but ultimately lead to OpRisk losses e.g., during the trading process in cases involving a lapse in attention and a OpRisk events occurs as a result of a mismatch between the trade booked (booking in trade feed) and the details agreed to by the trader.

The middle office (MO) and back offices (BO) conduct OpRisk managements' task of building mathematical models to be used to predict OpRisk losses and ultimately determine capital adequacy required to absorb these losses. The implications from modeling can be used to better understand the broad view of the overall company's OpRisk through PnL attribution carried out from deal origination to settlement. For instance, the results of the model can be used to better understand the interrelationships between risk factors and potential dependencies on various mitigation and management strategies (Acharyya, 2012) e.g., human error is a potential risk factor resulting in (positive or negative) PnL impacts: An efficient strategy (to also deal with trade breaks) would be to use this fact and counteract negative impacts which can be mitigated by an efficient trade amendment policy offsetting the outflow of PnL with an equal and opposite inflow or cash injection.

Other fundamental aspects of OpRisk generation

The argument for operational riskiness goes beyond lack of awareness and deficiencies in skill due to human failings of the conscious/sub-conscious forms. Hemrit & Arab (2012) discuss a fundamental rethinking of an old problem attributed to the birth of OpRisk pointing to those stemming from other fundamental aspects associated with natural catastrophes e.g., disease, environmental degradation, etc., with serious widespread consequences due to discontinuity of services, giving rise to many forms of risk (market, liquidity, credit or operational risk), which could lead to systemic risk. As Hemrit & Arab (2012) determine in their concluding remarks: It suffices to day, the case for financial institutions in defending their most profitable activities is non-trivial, particularly through an exercise using a calculation estimation of a minimum and adequate capital charge for OpRisk, independent of other risks.

Furthermore, according to (Acharyya, 2012) 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 regulatory capital (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.

Research objective

1. To circumvent the known HFLS & LFHS problem that comes from a combination of small sample sizes with a wide range of loss types, and associated difficulty of ”upward bias” from estimating errors due to m.l.e fitting of heavy-tailed

distributions, along with the resulting OpVaR model instability to ultimately realise the move toward an EBOR framework for OpRisk in the banking industry, balancing out both the existence of a heavy-tailed distribution fit on the one hand, and the stability, or variance, of the model derived from a small-data environment as it is applied to new OpRisk loss data on the other.

Sub-aim (iii): model OpRisk losses depending on augmented covariates

To deploy GLM and GAMLSS models forecasting OpRisk loss target variables: loss frequency and loss severity, a dataset derived through a data augmentation algorithm wherein approximate weights of input features consisting of a comprehensive list of user defined input variables (the risk factors) are modeled through heuristics derived in the underlying value-adding processs, with associated root causes in sequence. Consequently, the new framework overcomes systemic OpVaR model instability issues, and small-size data difficulties replacing capital estimates' "upward bias" of overcapitalizing realised losses after they occur. Therefore the additional features should be able to adapt to changes in the risk profile of the bank[1] providing sufficient incentives for to mitigate OpRisk.

Augmenting data for near misses, pending and realised OpRisk losses

Every year as standard operating practice, amongst banks captured key risk indicators (KRI's) are a number K_{annual} of events: A discrete random variable N of OpRisk loss events that either have no associated pnl impacts viz., near misses, pending impacts i.e., losses yet to make their mark, or actual realised OpRisk losses with pnl impact. In our case, ex post estimation as per subsection 4.7.2, Chapter 4, we have

[1]Upon the addition of new products or varying the business mix of the bank

modeled the number or frequencies of OpRisk losses following a binomial distribution, whereby each loss X_i follows this random variable with an associated severity distribution. Although The Poisson distribution is commonly favoured over the Binomial distribution in the estimation, our logic for the best fitting (binomial) distribution is owed to sequential conversion (as indicated in Chapter 2, section 2.12.1), wherein the moment generating function (m.g.f) of the binomial distribution is found to be one and the same as the m.g.f of the poisson distribution, which implies the discrete random variable N can also be modeled as a Poisson variable with rate λ as the required limiting distribution as well.

We have a 3 month dataset over the first quarter (Q12013), with a population of $K_{Q12013} = 2330$ OpRisk loss events collected internally i.e., between 1 January 2013 and 31 March 2013, and of these events we have a number $N = 371$ of realised losses. The researcher's task remains to relate how the size K_{Q12013} of the population enter the risk model; in response it doesn't appear explicitly in the model (Parodi, 2014), however, it is taken into account during the creation of the model. Our challenge as posed by the problem with regard to extrapolating data points is that we do not know what the future population K_{annual} is, and what associated realised losses N_{annual} hold except it is logical to assume that the two are larger in the annual case in comparison to the Q12013 scenario.

Firstly, one must determine the frequency distribution parameter λ: More precisely, the constant of proportionality between the current OpRisk loss size i.e., population K_{Q12013}, and the Poisson rate λ becomes relevant when taken into account for determining the annual population K_{annual} to be determined and its associated predicted losses. A numerical method using the software package Mathematica is used on the test set; see Appendix 8.9.3 this can be used to evaluate the parameter λ. This is carried out through the use of the R script (`r Rate`) in Appendix 8.12.1, which yields a daily rate of $\lambda = 0.20739163$ per day quarter. Computing for this

mean (λ) yields a cut-off probability of 0.18009498. By a simple growth formula, one years of data (4 quarters) i.e., 3 months * 4 = 1 year:

$$\begin{aligned} \text{1yr Population} &= \text{Initial Population} + \text{Initial Population} \times (1+\lambda)^n \\ &= 2330 + 2330 \times (1 + 0.18009498) + 2330 \times (1 + 0.18009498)^2 \\ &+ 2330 \times (1 + 0.18009498)^3 \\ &= 11,791 \quad \text{observations} \end{aligned} \tag{6.1}$$

This corresponds to a 1yr population of $11,791$ observations. In the following next step an extrapolation algorithm script is used to generate the $k_{annual} = 11,791$ observations determined for the augmented dataset.

Algorithm for population

The extrapolation algorithm which augments the existing data to increase the size of the population by the heuristics approach, effectively increasing the range of observations to span a 1 year period is best done in the Matlab code (due to it's matrix based foundations), see section 8.13 on page 266 in the Appendix 8.12. It's worth noting that this is only made possible as the solution to the density estimation problem has already been resolved.

Algorithm 5.1*: The dataset augmentation meta-algorithm for building the best amount of data observations required to train.*

1. Generating complicated high-dimensional input $\mathbf{x}$ to create new data and add to the training set
    - Import data and in excel and code to understand its nature i.e., Understand covariates through common vertices

- In order to simulate data we need to construct a random generator i.e., run a Monte Carlo simulation for a large number of iterations
- Established `TradedTime` $\in [0, 0, 79)$ is root fundamental to all the covariates, where $0\times$ days, is the start-of-business (SOB) time, & $0, 79\times$ days is the close-of-business time.

2. Simulation of observations with large range of factors of variation. This employs a powerful regularization strategy, seen as a process of constructing new inputs by injecting noise into the input of training data
   - Sample of n_{RANDOM} uniformly distributed $U[0, 1]$ frequency entries determined between SOB and COB
   - $\forall$ unique numbers $\in [0, 0.79)$ multiply by noise on replicated observations in the training set by careful tuning the noise which translate the training observations into altered magnitudes by small variations
   - Rotating or scaling operations are also useful
3. We can now generate more observations for this observation built from the root `TradedTime` covariate.
4. Separate mixed observations into numerical and categorical type variables in preparation for random sequencing and positioning
5. Numerical variable generation - generate a random sequence of new observations by replicating the family of distributions from the original training data
   - For each variable e.g., `exposure` find range of values constrained in [min, max], then find mean. Split his interval down the mean to obtain 2 intervals of equal length.
   - Generate random samples $\in [0, 0.79)$, split the two intervals into 4 and sort them in ascending order. compute the median within each interval
   - Within each range [min, max] captures the range, mean captures commonalities in the data due to its central positioning, & median captures

irregularities due to outliers (in the tail)

- Reverse the ordered numbers which now come from the family of distributions from the training data by a sorting process to complete the randomisation of numeric variables

6. Categorical variable generation - determined by row and column positioning of categorical (factor) variables as situated in the training data
   - generate uniformly distributed $U[0, 1)$ random frequency entries and multiply by the length of the original sample population 2330
   - Taking say $100,000$ data points, employ the regularization strategy mentioned above, generated by multiplying with any generated random number and using the floor (rounded down) function on the output to ensure the number interval $[0, 2330]$, and to avoid duplicated random numbers so each unique position in the original dataset is associated with a specific observation
   - As in numerical case, reverse the ordered numbers by use of a sorting process to complete the randomisation of categorical variables
7. Go through steps 1 through 6 repetitively, appending the output to build a matrix containing all $11,761$ observation entries

The generated output matrix, consisting of $i = 11,761$ observations in \$j = \$ can be observed in the excel spreadsheet labeled *Augmentation_Dataset_Model.xlsx* in Appendix 8.13.1.

Research question 6.5.1:

To what extent does an improvised (AI aided) augmentation dataset eliminate the HFLS & LFHS dilemma in OpRisk viz., loss data types replete with heteroskedasticity, resulting from combinations of small data sample sizes with wide ranges of event type losses to be modeled (Larsen, 2019), often rendering bias in maximum likelihood estimators (m.l.e's). If so, how do the intended implications on OpVaR model stability

aid in the estimation of frequency and severity distributions of OpRisk losses?

Deploying the R model

Model deployment is an important issue that is often overlooked in data mining projects, yet translates the benefitting of modeling once one has been developed, evaluated and deemed suitable. Following the preceding processes of running the model over some observations and generating performance measures, the model is "scored" through applying it to a new dataset using a form of a `predict()` function. Scoring allows the forecasts to be analysed by class, probability or forecated estimate, then stored and saved to a file for later use in model deployment.

This is the most effective approach to model deployment and is practiced regularly as new data entries become available i.e., particularly in the R concept, model outcomes are saved for later use then at a later time a new dataset scored using the saved model and applying it to some new data enhancing the data mining capabilities within an organisation. That is to say, following the heuristics approach in the build toward a dataset through data augmentation, point-in-time estimates of the target variables; LossIndicator and Losses, are modeled through deploying the selected GLM and the GAMLSS models to generate forecasts.

Deploying the estimated GLM frequency model

A classification approach is applied when generating target values to generate data from the GLM model. In what follows, the afore-mentioned task is enabled by the solution to the density estimation problem (**Amodel**), which is readily available from the loss frequency distribution. A classifier needs to take a complicated, high-dimensional input $\mathbf{x}$ and summarise it with a single category identity y. Un-

der transformational invariance, the classifier transforms $\mathbf{x}$ inputs in the *Augmentation_Dataset_Model.xlsx* matrix to easily generate new $(\mathbf{x}, y)$ pairs.

Recall the formula for the estimated model **Amodel** whose target variable we are modeling:

$$\begin{aligned} \mathrm{Y}_i^t &= \beta_0 + \beta_1 \mathrm{BusinessLineLevel1}_{i1}^t + \beta_2 \mathrm{CapturedBy}_{i2}^t + \beta_3 \mathrm{EventTypeCategoryLevel1}_{i3}^t \\ &+ \beta_4 \mathrm{Desk}_{i4}^t + \beta_5 \mathrm{Instrument}_{i5}^t \beta_6 + \mathrm{Reason}_{i6}^t + \beta_7 \mathrm{TradedDay}_{i7}^t + \beta_8 \mathrm{TradedTime}_{i8}^t \\ &+ \beta_9 \mathrm{TraderId}_{i9}^t + \beta_{10} \mathrm{TradeStatus}_{i10}^t + \beta_{11} \mathrm{UpdatedDay}_{i11}^t + \beta_{12} \mathrm{UpdatedTime}_{i12}^t + \epsilon_i^t \end{aligned} \tag{6.2}$$

where, (matrix slices) $\mathrm{t} = \{1, \ldots, 78\}$ and (numrows) $\mathrm{i} = 1, \ldots, 4096$
and, target variable $\mathrm{Y} = \mathrm{LossesIndicator}$

The results are simulated by R script `r GLM Forecast` written in subsection 8.13.2 referred to in Appendix 8.12. A portion of the output results are displayed in Appendix ref{} and can be observed in the excel spreadsheet labeled *Augmentation_GLM_Model_Deployed.xlsx* in Appendix 8.13.1.

Deploying the estimated GAMLSS severity model

ML algorithms that inject noise, such as adding Gaussian noise to the inputs are performing a form of data augmentation. In this paradigm, this can be referred to as the process of regularisation, which is a generally applicable process of constructing modeled outputs from new inputs, i.e., (predicting $\mathbf{y}$)'s from given $\mathbf{x}$. Goodfellow, Bengio, Courville, & Bengio (2016) prescribes, the main task facing a regressor is to be invariant to wide variety of transformations e.g., multiplying those new $\mathbf{x}$ inputs by random noise, hence generate new $(\mathbf{x}|\mathbf{y})$ pairs.

When making predictions from $P(\mathbf{x})$ used to estimate $P(\mathbf{y}|\mathbf{x})$, by virtue of the afore-mentioned context statistical inference must also be accommodated. This

description requires the model to be probabilistic which by implication, includes members in the GAMLSS model. One can construct models in which the generative model of $P(\mathbf{x})$ shares parameters with a model of $P(\mathbf{y}|\mathbf{x})$. This also means, with reference to the structure of the GAMLSS model that $P(\mathbf{x})$ is connected in a way that is captured by the shared parametisation (Goodfellow et al., 2016).

Much like the GLM case, however in the case for the GAMLSS scenario, the regressor transforms $P(\mathbf{x})$ inputs in the same *Augmentation_Dataset_Model.xlsx* matrix, with the exception of pending and/or near misses as determined from the GLM scenario. Recall the estimated "best model "equation 5.3: Each model l outputs a probability distribution $p^{(l)}(y|\mathbf{x})$. The prediction of the ensemble is given by the arithmetic mean of all these distributions:

$$\frac{1}{k}\sum_{l=1}^{k} p^{(l)}(y|\mathbf{x}) \tag{6.3}$$

The results are simulated in an R script `r GAMLSS Forecast` written in subsection 8.13.3 referred to in Appendix 8.12. A portion of the The output results are displayed in Appendix 8.12and can be observed in the excel spreadsheet labeled *Augmentation_GAMLSS_Model_Deployed.xlsx* in Appendix 8.13.1.

Further analysis and intepretation of augmented dataset

The augmented OpRisk risk data consists of 1,345 realised losses out of a total of 11,760, simulated over a 1 year period beginning from 1$^{\text{st}}$ January 2013 to 31 December 2013, unadjusted for inflation. As described these loss events were not actual recorded events obtained through any OpRisk database, hence not actual losses in real time but a sequence of a wide range of modelled losses with a minimum of

less than R100 and the maximum of over R850,000. The losses are simulated at the most granular level (on a trade by trade basis) then aggregated on a daily level eliminating bias and non–representativeness in the dataset. The tractability in the data conveniently permits tracing back to uncover root causes, therefore enabling heuristic tests for accuracy and completeness, as well as creatively put forward the timely establishment of tests for reliability and whether the loss data process flow is valid. This LDCE electronic solution benefits OpRisk management as the data can easily be analysed in various formats, define follow up action items timeously, and assign responsible persons to attend to unexplained losses.

Event loss (EL) type findings

Figure 6.1 illustrates a histogram of the full augmented dataset referred to in Appendix ref{App:Appendix D}. The proportions of OpRisk event types (EL's) pertaining to trade statuses over the afore-mentioned given period are displayed. A further drilling down is portrayed in the analysis illustrated in Figure 6.2: Realised OpRisk loss data is extracted from total augmented dataset sample and a similar proportioning of original root causes is portrayed.

Of the 11,760 OpRisk losses, histogram 6.1 details that most trade amendments are performed in BO confirmed status, followed closely by BO-BO confirmed trades, then a drastic decline (down $\sim 1000\%$) to trades in Terminated states, then a further decline down $\sim 2000\%$ in documented trades in Voided status.

Additionally, root cause analysis portrays a decreasing trend from EL7(7,448) - Execution, Delivery and Process Management (EDPM); EL4(2,298) - Clients, Products and Business Practices (CPBP); EL1(1,109) - Internal Fraud (IF); & lastly EL6(905) - Business Disruption and System Failures (BDSF). A interestingly comparable meta-analysis is unveiled in the picture portrayed in Histogram 6.2. Though

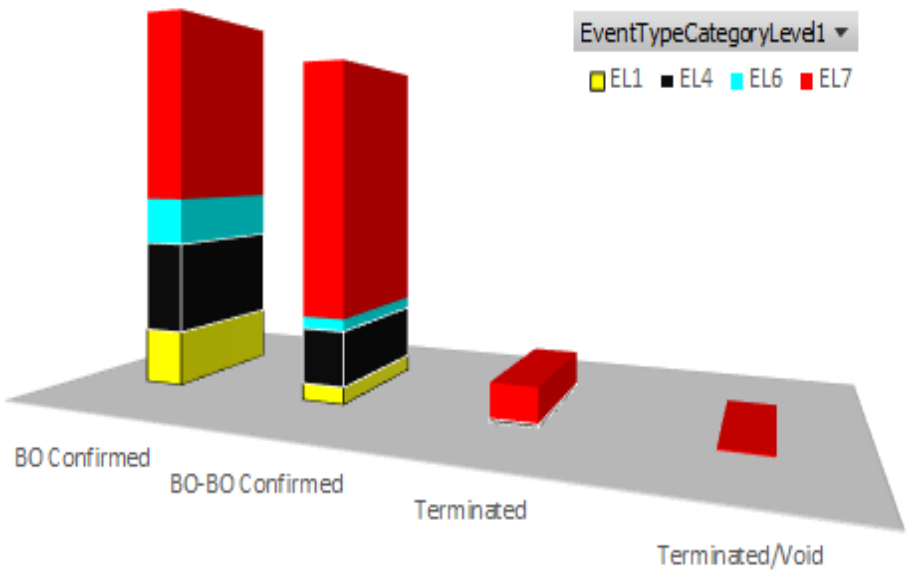


Figure 6.1
Histogram of frequencies for all losses per event type for amended, terminated, & Voided deals

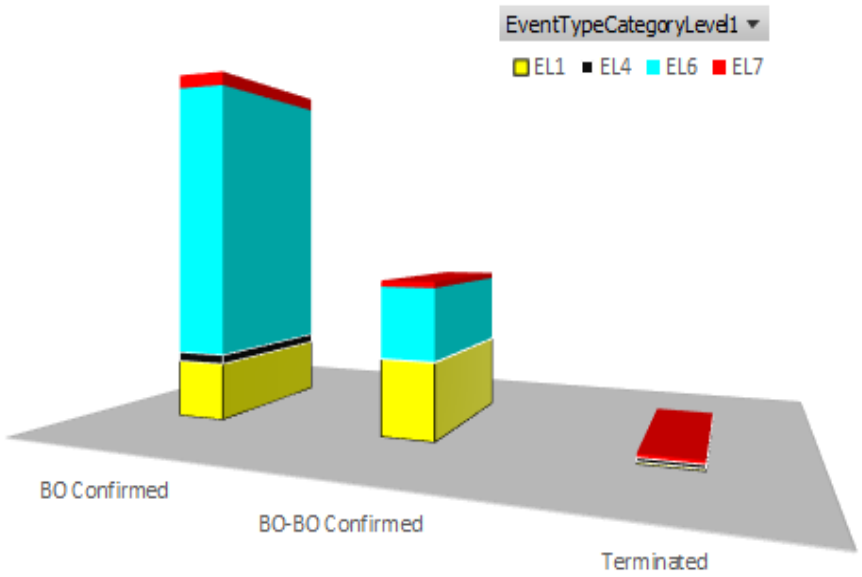


Figure 6.2
Histogram of frequencies for realised losses per event type for amended, terminated, & Voided deals

the trend in amendment detail versus trade status prevails, the root cause analysis reveals the prominent loss bearing causes are due to BDSF(905); followed by IF(346); then CPBP(60); & lastly EDPM(34).

Business line (BL) findings

In the meta-analysis, noteworthy similarities in trends pertaining to OpRisk losses across business lines, paints a contrasting landscape to the afore-mentioned event type trends. Firstly, w.r.t the trade status and root cause trend analysis, both are consistent between the two scenarios i.e., the trade amendment detail characterised by ordered decreasing magnitudes remains consistent throughout for business line root cause analysis; Leading with Investment Banking (BL1 - Corporate Finance (CF) & BL2 - Trading and Sales (TS)); followed by Banking (BL3 - Retail Banking (RB), BL4 - Commercial Banking (CB), BL5 - Payment and Settlement (PS), & BL6 - Agency Services (AS)); and lastly Other (BL9 - Corporate Level – Non-Business Line Specific).

At the level of individual loss events it is fundamental that the business line owner in the bank knows when they happen, be able to identify the event type/root causes of losses arising from operational events and attribute the losses appropriately. For example, the TS/BDSF maximum loss may raise a red flag: It is currently flagged to have occurred as a result of processing errors on that day, notably due to the reason, "Capture Errors-Direction/Amount/Rate/CP", resulting in a single loss of R866,670. The intrinsic benefit becomes clear here, as OpRisk teams become equipped to proactively manage `OpRisk`, perhaps moving this OpRisk type out of BDSF events to where it's better suited, perhaps in a different business line and more appropriately covered for. Moreover, experience from the meta-analysis wherein business lines were noted to impart predictability over event types provides incentive for these actions,

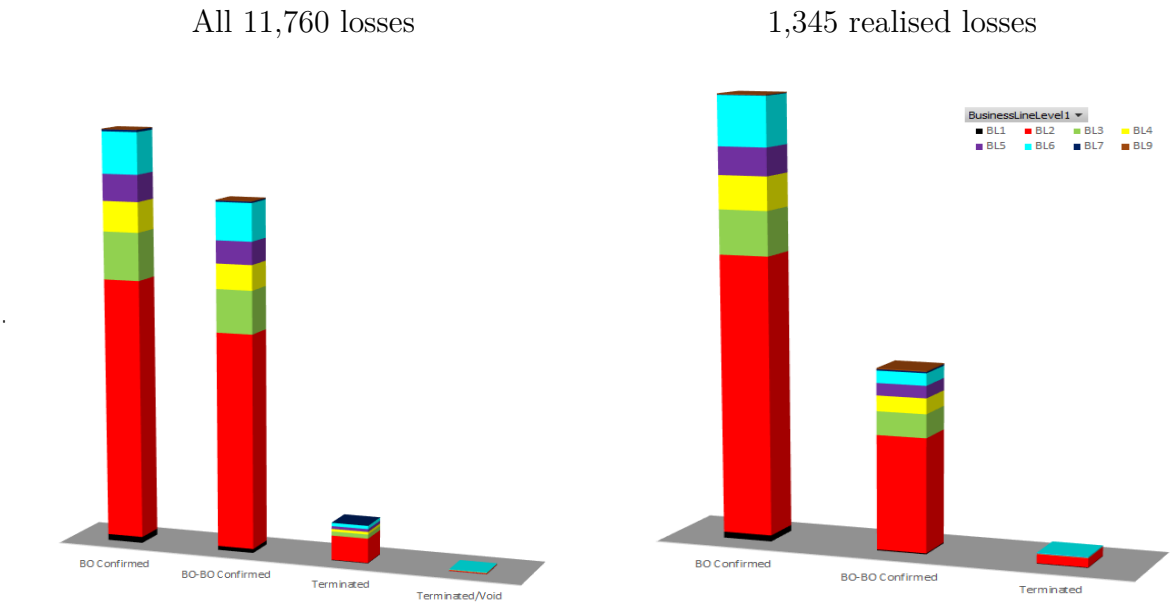


Figure 6.3
Histograms of business line root cause analysis for amended, terminated and/or voided deals (a) 11,760 cases of pending, nearly missed and/or realised losses, and (b) 1,345 realised losses of the total 11,760 losses

in this manner maximising the certainty of meeting business objectives.

Proposition 6.7.1:

An AI aided EBOR framework with forward looking modeling aspects improves operational risk capital-at-risk (OpCaR) requirements over traditional models who lack AI "learning" capabilities hence suffer from this as a limitation.

Capital Calculation

This section will focus the latter part of the new EBOR framework for `OpVaR` calculation, based on the LDA modeling of a bank's operational losses over a one-year time horizon. The previously computed estimates of the frequencies for near misses, pending and subsequent severities associated with actual OpRisk losses can now be used in the calculation toward the total CaR computation as defined in 2.6.5

by definition 2.6.5.1 in equation 2.24. The prior results show that the accuracy of the frequency and severity estimates are reasonable and consistent with industry standards.

Computational analysis

After modeling OpRisk losses $(X_1, \ldots, X_N)$ at the individual level, aggregated operational losses can now be seen as a sum S of a random number N of the individual operational losses; when combined S can be expressed as $\vartheta(t) = \sum_{n=1}^{N(t)} X_n$. By common consent, Frachot et al. (2001); Aue & Kalkbrener (2006); Panjer (2006); & others define total *Capital-at-Risk* (CaR), $\mathbf{G}$: 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:

$$\begin{aligned} \text{CaR} &= \sum_{k=1}^{m} \text{VaR}_\alpha(X_k) \\ &= G_k^{-1}(x) \end{aligned} \tag{6.4}$$

$$\mathbf{G}_{\vartheta(t)}(x) = Pr[\vartheta(t) \leq x] = Pr\left(\sum_{n=1}^{N(t)} X_n \leq x\right) \tag{6.5}$$

As determined in chapter 2, subsection 2.6.4 equation 2.22.

Let $\mathbf{X}$ be the total loss of the bank. By equation 2.12 and considering that the losses X_k are independent; the distribution $\mathbf{G}$ is the convolution of the distributions $\mathbf{G}_k$:

$$\mathbf{G}(x) = \underbrace{G_1 \circledast G_2 \circledast \cdots \circledast G_m}_{m} = \overset{m}{\underset{k=1}{\circledast}} \mathbf{G}_k(x) \tag{6.6}$$

As previously defined, the CaR of the bank is:

$$\text{CaR}(\alpha) = \mathbf{G}^{-1}(\alpha) \tag{6.7}$$

OpVaR basics and sampling criteria

The study undertakes an m.l.e fitting method of distribution fitting, as opposed to the range of other common practices in defining the distribution e.g., extreme value theory (EVT) fitting and quantile matching. This is due to granular level detail inherent in the LDCE, whereby OpRisk losses are detailed on a BL/ET basis, in addition to other fields useful in root cause analysis. This makes it possible for a scenario where only individual events and the losses accompanying them are considered, therefore circumventing considerations for the need to use these other replacement estimation methods rather than the better suited m.l.e.

As opposed most older studies who's only focus is on severity modeling e.g., Larsen (2019) and Ganegoda & Evans (2013), the study in question focuses on both frequency and severity modeling, disputing the not very well founded assumptions which dictate that OpVaR is more sensitive to severity modeling than frequency modeling. By virtue of our chosen approach the experimental outcomes weigh less on the restrictive constructs of common wisdom based on intuition, who significantly contribute to the varied concerns typical in the OpVaR modeling. In the domain of the severity distribution, loss amounts are modeled using a spliced (piecewise-defined) severity distribution, with losses below a chosen threshold modeled by one parametric distribution and by another fitted for losses above this threshold (Aue & Kalkbrener, 2006).

Computation of the aggregate OpRisk loss using R OpVaR package

The key element to computing OpRisk capital estimates using the R OpVaR package (consisting of multi-platform, interactive state-of-the-art statistical procedures, random number generators and graphics), is in the structure of a hypothetical dataset generated from the augmented dataset as sketched out earlier in section 6.6.2.

Experimental setup for using R OpVaR package

Our hypothetical dataset's structure is inspired by the 56 BL/ET combination matrix risk type cells as depicted in figure 6.4, arranged in a list, each representing a cube element as depicted. Within each cube comprises a loss frequency model, a loss severity model and if specified a dependency model between the frequencies and severities. Subsection 8.13.4 provides demonstration of the hypothetical dataset, and the list of cells storing the requisite data structure comprising a dataframe of 16 cells/list elements and an integer (weekly) time period assignment for the model 8.13.4 is as initialized in Appendix 8.12.

The BL/ET combination matrix sample

The 3-dimensional diagram depicted in figure 6.4 shows the formation of a $7 \times 8 = 56$ BL/ET combination of matrix risk-type cells: A duration parameter denoted by time $[T, T + \tau]$ representing the next period's loss (usually the next year's annual loss) over which CaR is defined is shown along the depth k ordinate. Each individual OpRisk BL/ET is linked by a unique identifying trait viz., the trade number, which restricts a trade repetitively appearing across several BL/ET occasions thereby avoiding double counting. This trade by trade nature of loss reporting is essential

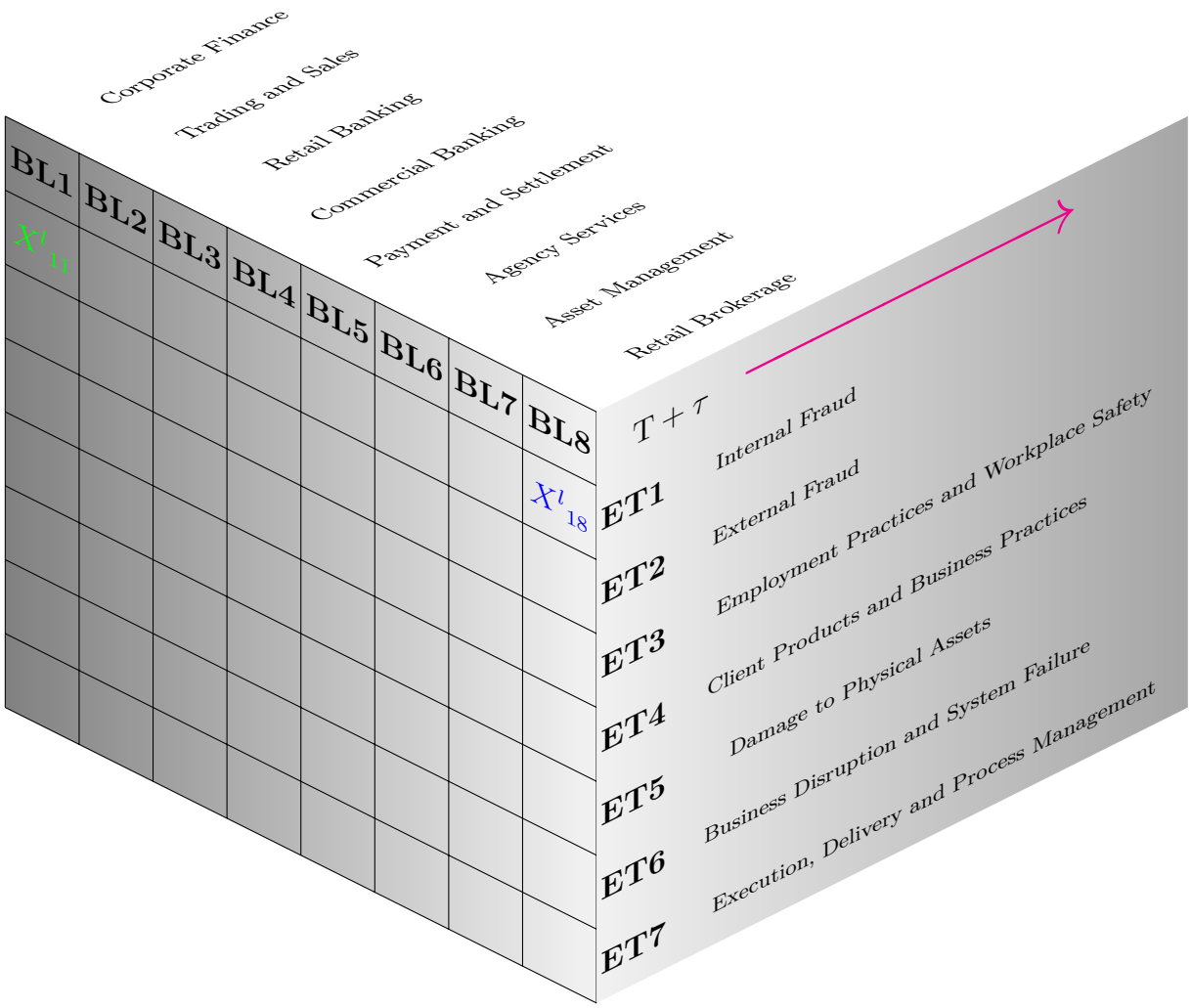


Figure 6.4
The 3-Dimensional grid of the BL/ET matrix for 7 event types and 8 business lines

for frequency calibration, whose loss dataset directly coincides with severity data as losses suffered are specific to trade level events.

Experimental design and analysis

In the construction of the opriskmodel, loss frequency is typically selected from the better fitting choice between two discrete distribution options viz., Poisson or Negative-Binomial distributions (Group et al., 2013), fitted taking into account the

weekly periods of discrete counts classified in the structure of the afore-mentioned dataset. In the case for loss severities, three forms of model types as numbered below can be selected:

1. "plain" - a single distribution (e.g., gamma, lognormal, Weibull, Tukey's g-h distribution)
2. "spliced" - a flexible combination of two distributions and a threshold (m.l.e)
3. "mixing" - a dynamically weighted mixture model.

For consistency, the "spliced" ex post facto design in scenario (2) above takes precedence over scenario's (1) & (3) for parameter estimation in order to obtain OpVaR statistics on m.l.e error. The choice of the splice threshold was chosen at T = ZAR2000. For the collective risk model parametrization of tail distribution for the hypothetical fixed loss-dataset of sample size $t = 52$ weeks is performed by R OpVar objects of type freqdist for frequency distributions and sevdist for severity distributions. This further degree of complexity to the computationally expensive a priori process of parametrization from the true severity distribution as prescribed in chapter 5.

According to Goodfellow et al. (2016) in their established theories underlying dataset augmentation, and the theory of asymptotic normality of m.l.e Frachot et al. (2004), it is easier to improve the generalisation of a target variable (LossIndicator of Loss) by increasing the size of the training set by adding extra copies of the training examples, therefore achieving statistically significant approximations accomplished by eliminating the "variance-bias" dilemma in the construction of our hypothetical dataset. In effect, this overcomes the allegation in Larsen (2019) formed of the restrictive phenomena that generality comes at a high price with respect to stability, more precisely since the GB2() distribution nests the Weibull, Lognormal, Log-logistic, Burr distributions.

Goodness-of-fit tests and diagnostic fitted distribution plots

If an R OpVaR object is of type sevdist (for the severity distribution), then the Kolmogorov-Smirnov test is performed on the loss data of a single BL/ET cell and the respected fitted distribution given by the sevdist object. Otherwise, if the R OpVaR object is of type freqdist, then a χ^2 goodness-of-fit test will be performed.

Table 6.1 generated by goodness-of-fit tests for severity distributions and frequency distributions.

Table 6.1
Testing the goodness-of-fit of the distributions given in a sevdist or a freqdist object to the loss data

Test	**χ^2 test for given probabilities**			**Two-sample Kolmogorov-Smirnov test**	
Data -	Frequencies			Loss severity and week in which loss occurred	
Statistics -	**χ^2 test**	**df**	**p-value**	**D**	**p-value**
BL/ET_Pairs					
BL1_ET1	5,1548	2	0,07597		0,6
BL1_ET6	14,263	3	0,002568	0,58333	0,0337
BL2_ET1	10,161	10	0,4265	0,25616	3,28E-06
BL2_ET4	21,778	4	0,000222	0,45455	0,002187
BL2_ET6	16,933	23	0,8125	0,33096	2,20E-16
BL2_ET7	20,416	4	0,000413	0,39286	0,0003528
BL3_ET1	14,803	4	0,005127	0,36842	0,01151
BL3_ET6	25,383	7	0,000649	0,34906	4,92E-06
BL4_ET1	26,197	4	2,89E-05	0,41176	0,006273
BL4_ET6	12,242	5	0,03162	0,25714	0,01954
BL5_ET1	15,783	3	0,001256	0,46667	0,002908
BL5_ET6	18,801	4	0,00086	0,25	0,06039
BL6_ET1	26,449	3	7,68E-06	0,32432	0,04081
BL6_ET6	10	6	0,1288	0,27368	0,001624
BL7_ET6	5	2	0,07597	0,66667	0,6
BL9_ET7	5	2	0,07597	0,33333	1

Experimental results

Bivariate dependence effects were detected via the Pearson test for correlation using SPSS, by evaluating significance at the $0,05$ level (2 tailed) as indicated (by the *. shaded entries) in Table 6.2 generated. As such, copulas are conveniently incorpo-

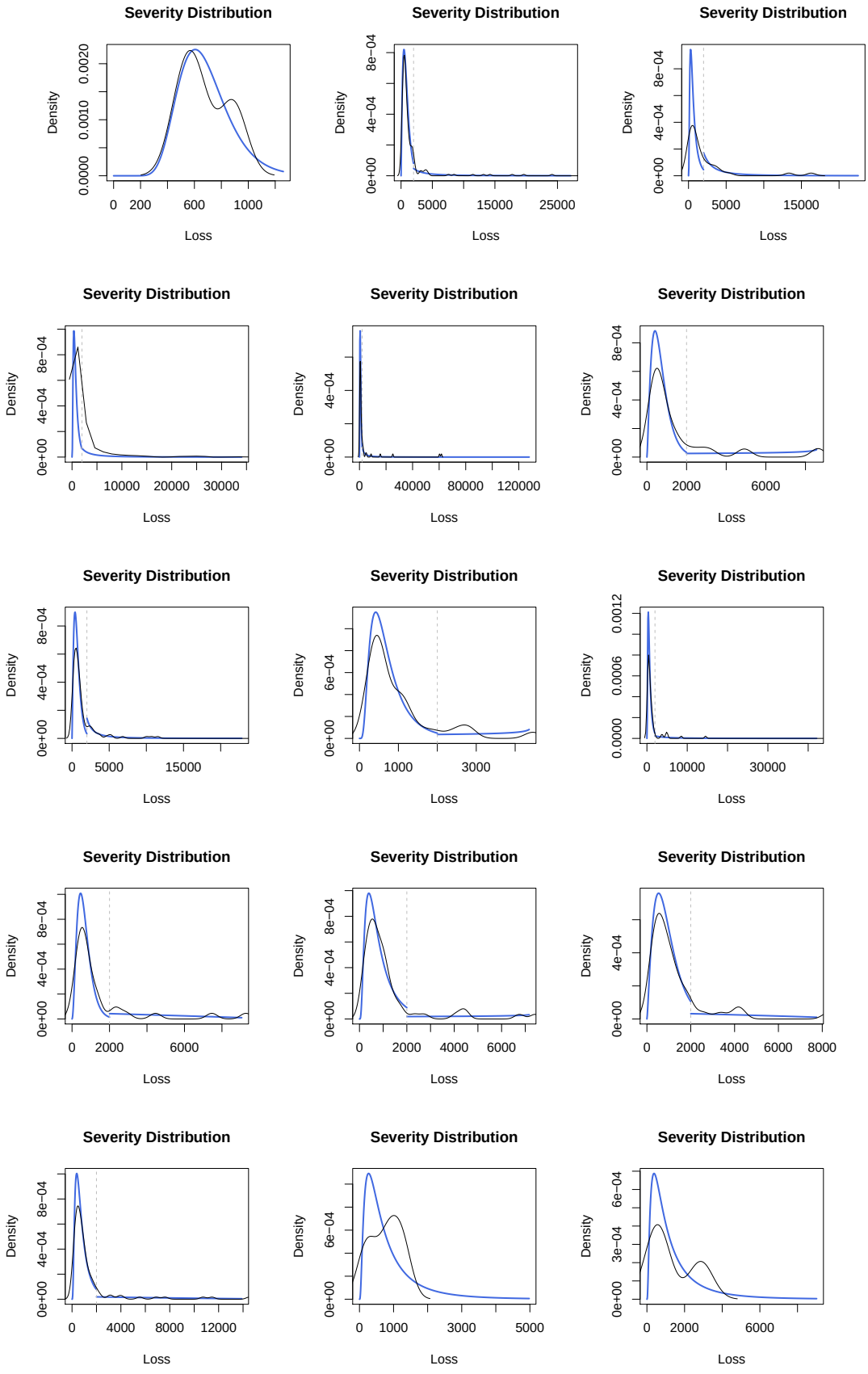


Figure 6.5
Diagnostic plots for the fitted severity distributions parameters of the fitted terms to BL/ET cells in the model.

rated into the model using the R OpVaR fitDependency function, to account for the risk-reducing impact of correlation effects between the BL/ET pair combinations.

Table 6.2
Matrix of correlations of BL/ET cells

	BL1_ET1	BL1_ET6	BL2_ET1	BL2_ET4	BL2_ET6	BL2_ET7	BL3_ET1	BL3_ET6	BL4_ET1	BL4_ET6	BL5_ET1	BL5_ET6	BL6_ET1	BL6_ET6	BL7_ET6	BL9_ET7
BL1_ET1	1															
BL1_ET6	-0,1	1														
BL2_ET1	0,1	0	1													
BL2_ET4	0,1	-0,1	0,1	1												
BL2_ET6	-0,1	0,4	0,2	-0	1											
BL2_ET7	0,3	0,3	0,2	-0,1	0,4	1										
BL3_ET1	0,2	0,2	-0	0,1	0,3	0,2	1									
BL3_ET6	0	-0,2	0,4	0,1	0,2	-0	-0,1	1								
BL4_ET1	-0,2	-0,1	0	-0,1	0,1	0	0	-0	1							
BL4_ET6	-0	0,2	0,3	-0,3	0,3	0,4	0,1	0,1	0,2	1						
BL5_ET1	0,3	0,1	-0,1	-0,2	0,3	0,2	-0,1	0,1	-0,2	0,1	1					
BL5_ET6	0,3	-0	0,3	0,2	0,4	0,4	0,3	-0	-0,1	0,1	0	1				
BL6_ET1	-0	0	-0,2	0,2	0,2	0,2	0,1	0	0	-0,1	-0,1	0,3	1			
BL6_ET6	-0	0	-0,3	0	0,1	-0,1	0	0,1	0,1	-0,1	0,2	-0,1	0	1		
BL7_ET6	0,3	-0,1	0	-0,1	0,1	0,1	-0,1	0,2	-0,1	-0,1	0,5	-0	-0,2	0	1	
BL9_ET7	-0,1	0,1	-0,2	0	0,2	-0,1	0,1	0,1	0	-0	0,3	-0	0,1	0,3	-0,1	1

*. Correlation is significant at the 0.05 level (2-tailed).

Total loss estimation by Monte Carlo simiulation

Table generated by aggregating a multi Monte Carlo loss model for Operational Risk

Conclusion

Best practice debates underway in this study propose leveraging ML models; through their innovative "learning"capabilities, towards a disruptive OpRisk framework design in the banking sector better equipped (by virtue of training on more data) to generalise on OpRisk measurement/assessment. Data augmentation gets around the data hungry requirements for more efficient ML training algorithms; by generat-

Table 6.3
The specified dependence copula 3 = Clayton copula is specified

Frequency Copula	CopulaBiClaytonFit	
BL/ET_Pair	**Copula Values**	**Direction**
BL1_ET6/BL2_ET6	0,380610413	4
BL1_ET6/BL2_ET7	0,316157205	4
BL1_ET1/BL7_ET6	0,067030398	4
BL2_ET1/BL3_ET6	0,594911937	1
BL2_ET1/BL4_ET6	0,607669617	1
BL2_ET1/BL5_ET6	0,525714286	4
BL2_ET6/BL2_ET7	0,511363636	1
BL2_ET6/BL4_ET6	0,492481203	4
BL2_ET6/BL5_ET6	0,567279768	4
BL2_ET7/BL4_ET6	0,644067797	4
BL2_ET7/BL5_ET6	0,537799043	4
BL5_ET1/BL7_ET6	0,221105528	4
BL5_ET6/BL6_ET1	0,487804878	4
BL6_ET6/BL9_ET7	0,1216	4

ing numerous fake data points, and dovetails as a remedy into the HFLS & LFHS dilemma of large and rare loss events by resolving the variance-bias problem associated with small sample sizes, aggregate units of measurement, and a wide spectrum of the loss BL/ET mix to the model.

More precisely, data augmentation redundifies aforementioned metrics former implications on OpRisk accuracy and their compounding effects on OpVaR stability, therefore aiding in estimation of OpRisk loss frequency and severity true distributions, and hence on capital calculation. Nevertheless, there must be a trade-off here, more 'data' so that one can do 'learning', BUT the quality of the 'data' is suspect as they are constructs not observations.

Subsequent to a prolonged duration extending additional data points from the 3 month data series to an annual dataspan via data augmentation, wherein forward looking targets are estimated assisted via the application of unsupervised ML model predictions, the current research embarked on a capital calculation exercise via R OpVaR package criteria, revealing a CaR value estimated at ZAR2.9bn ($\sim$ 3 billion South African Rand). Hoohlo (2014)'s comparabe CaR estimation, obtained on the

Table 6.4
Ordered losses for 1E6 MC simulations to estimate total CaR. The mean of the cumulative distribution is R2 711 000 and standard deviation R306 100 000

Var form fit

99,9%	**2 906 445 287**	**Capital-at-Risk (CAR)**
99,0%	**1 102 506 359**	
95,0%	**447 013 222**	

Distribution Fit

	FITTED DISTRIBUTION PAIRS	
BL/ET_Pairs	**Frequency Distr**	**Severity Distr**
BL1_ET1	Poisson(0.0577)	Lognormal(673.033;156.453)
BL1_ET6	Poisson(0.231)	Lognormal(4790.2581;1756.407)
BL2_ET1	ZTPoisson(3.818)	GPD(2023;3054.433;0.793)
BL2_ET4	Poisson(0.635)	GPD(2112;1349.637;0.806)
BL2_ET6	ZTNegBin(7;0.393)	GPD(2013;3331.3516;0.7316)
BL2_ET7	Poisson(1.077)	GPD(2029;2884.8086;1.4026)
BL3_ET1	Poisson(0.731)	GPD(2154.7696;8977.705;-1.3678)
BL3_ET6	Poisson(2.038)	GPD(2010;1852.475;0.7632)
BL4_ET1	Poisson(0.654)	GPD(2412;497.473;0.359)
BL4_ET6	Poisson(1.346)	GPD(3646;1984.99;1.447)
BL5_ET1	Poisson(0.577)	GPD(2243;3648.829;-0.303)
BL5_ET6	Poisson(1.077)	GPD(2303;4038.083;-0.748)
BL6_ET1	Poisson(0.712)	GPD(2578;3391.314;-0.43)
BL6_ET6	Poisson(1.827)	GPD(2103;5759.58;-0.2)
BL7_ET6	Poisson(0.058)	Weibull(1.849;856.025)
BL9_ET7	Poisson(0.058)	Weibull(1.287;1448.13)

Distr Tested

Frequency	**Severity**
Poisson	Gamma
Negartive Binomial	Lognormal
ZTPoisson	Weibull
	GPD
	Erlang

real 3 month dataset through the application traditional statistical techniques to calculate an OpVaR proxy of $\sim$ 3 billion South African Rand.

Evidently, the results indicate that at worst the new framework's performance matches Hoohlo (2014)'s OpRisk VaR measure and is appropriately calculated, opening up the horizon for the exploration modeling beyond historical accounts of significance, but incorporate forward-looking aspects of the BEICF henceforth. Furthermore, future projections which look reasonable and believable in practical terms, demonstrates that data augmentation, which is also cheap and efficient, set forth an improved ORMF when coupled with ML for the purpose of measurement and comparability of data achieved through heuristics, lowering costs of monitoring OpRisk in banking as recommended by BIS.

CHAPTER 7
THEORETICAL INVESIGATIONS INTO THE QUANTIFICATION OF MODERN ORMF

Introduction

Behavioral management theory is very much concerned with social factors which are often not easy to incorporate into finance models or to understand in the traditional framework. A critical component of behavioral finance is building models which reflect actual behavior such as motivation, support and employee relations, but many studies have revealed that these social factors are not easy to incorporate in models, until recently. In an adoptive fashion, it's become customary to shown that people regularly deviate from the traditional finance paradigm e.g., as put forward by Barberis & Thaler (2003) through evidence garnered in extensive experimental results compiled by cognitive psycologists on how people make decisions given their beliefs.

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. Many argue that individuals are fundamentally irrational because they do not behave rationally, more precisely, that neoclassical theory makes use of the assumption that all parties will behave rationally overlooks the fact that human nature is vulnerable to other forces; e.g., the essential feature of prospect theory (PT) are assumptions which are compatible with basic principles of perception and judgment for decisions taken under uncertainty which causes people to make irrational choices, giving rise to a literature and debates as to which heuristics and sociological and institutional priors are rational (Altman, 2008).

In this chapter, with the aid of Machine Learning (ML), more evidence on how human choice behavior can substantially deviate from neoclassical norms is provided. An unsupervised learning technique viz., Clustering analysis is used as a substitute to the traditional model based Autoregressive Moving Average (ARMA) method commonly used in analysing and representing stochastic processes. Clustering analysis 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.

Analysis and interpretation issues with behavioral finance theory

Investors systematically deviate from rationality when making financial decisions, yet as acknowledged by Kuhnen & 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 & Knutson, 2005).

As an example, risk-aversion mistakes occur when 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. Kuhnen & 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 phenomena; as evidenced in recent studies found from human brain imaging linking neural states to risk-related behaviours (Paulus, Rogalsky, Simmons, Feinstein, & Stein, 2003), which when activated can lead to a shift in risk preferences. Kuhnen & 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.

According to Barberis & Thaler (2003), within the variety of relations among risk preferences or how investors evaluate risky gambles, as a prerequisite assumption about investor preferences and an essential ingredient of any model trying to understand trading behavior, people have difficulty in grasping the concept of risk-neutrality. This concept can be characterised by a linear value function sometimes referred to as the utility function evidenced on a value/utility against outcomes (loss or gains) plot by symmetry around a neutral point where there is perceived indifference in losses and gains. A substantial body of evidence suggests that loss aversion, the tendency to be more sensitive to losses than to gains, or asymmetry around the nucleus whereby the perceived tendency is 2.5 times more powerful for losses as compared to gains is the more prominent of these opposing views, who play an important role in determining how people evaluate risky gambles.

Expected utility theory

The traditional finance paradigm seeks to understand financial markets using models in which agents are "rational". According to Barberis & 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 & Tversky, 2013).

Expected utility theory[1] (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:

$$U(x_1, p_1; \ldots; x_n, p_n) = p_1 u(x_1) + \ldots + p_n u(x_n) \tag{7.1}$$

corroborated by Morgenstern & Von Neumann (1953); Friedman & Savage (1948); Kahneman & 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, unless you have another, say, constant absolute risk aversion (CARA) utility. In this case one can determine how preferences change as risk aversion increases, however, it's still not enough to indicate whether the implications of their preferences due to levels of risk are "rational" (Babcock, Choi, & Feinerman, 1993). 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.

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

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

with a psycologically more realistic preference specification (List, 2004).

A new class of ORMF models approach

In recent work (Barberis & 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 have been several theories put forward advocating for the basis of a slightly different interpretation 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[2].

Kahneman & Tversky (2013) maintains that preferences between prospects which violate rational behaviour demonstrate 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, & 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

[2]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

economic theory[3] (Kuhnen & Knutson, 2005).

Prospect theory

Kahneman & 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 & 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[4] more times) than gains, and are especially sensitive to small losses unless accompanied by small gains[5]. Loss aversion is a strong differentiator when it comes to explaining exceptions to the general risk patterns that characterize prospect theory.

According to Kahneman & 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. Therefore, by relaxation of the expectation principle in equation 7.2.1, 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:

$$\bigvee = \pi(p)\nu(x) + \pi(q)\nu(y) \qquad \text{iff} \qquad p + q \leq 1 \tag{7.2}$$

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

[4]Deviation magnitude of the kink in the curve from the straight line in the value function in Prospect theory graph

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

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.

Sub-problem 2.

We challenge the weaknesses in current OpRisk theory which assumes banks are risk neutral, asserting they are more risk averse. Consequently, "predictive" future losses can be determined who's estimated capital requirements reflects the degree of risk tolerance and adapts to changes in the risk profile of the bank providing sufficient incentives for OpRisk management to mitigate OpRisk.

Sub aim (iv): 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. The wish is 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 making 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.

Theoretical foundations of risk-neutrality in risk quantification

In a market where securities are traded, risk-neutral probabilities are the cornerstone of trade largely 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 the idea compounding state prices defined in their risk-neutral probabilities. This gives rise to the concept of time value of money, wherein an artificial compounding rate can be reversed by a subsequent equivalent discounting, to find the way back to a determined arbitrage-free price (state price).

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.

Risk emerges on an imposed deviation from a risk-free state wherein interest rate different from zero implies the market allocates different utility to possessing goods for consumption now rather than only later. This particular risk-*free* deviation to risk-*neutral* appetite is founded in trust and stems from the nature the relationship/s amongst market participants.

For example, among games of (equal) chance i.e., individual players in a casino operate on the assumption that one can outwit the establishment, whereas the casino work on the assumption; of 50% chance outcome per turn, made through the law of large numbers: - On the one hand, chance outcomes exhibit mean regression which informs the establishments strategy of involving a many participants as required, since on the other hand, larger sample sizes in relation to variation leads to more consistent outcomes. This idea seems quite intuitive from an risk perspective since the very existence of it's management is the basis for the core purpose in banking.

Banks valuation of time value of money or "interest"is founded in groundbreaking concepts like the law of large numbers, and principles of inferential statistics, whereas banks themselves are found to stem from much simpler, rudimentary founding principles, like trust. In the interest of business sustainability, banks play a numbers game of attracting as many deposits as allowable to operate similarly to the afore-mentioned scenario involving the casino. Much like the casino example who attract funds through games of chance, the bank derives interest based on a degree of risk tolerance λ which determines risk-neutrality, risk-seeking or risk-averse attitudes. For example, a bank of risk tolerance λ with parametized utility curve $U(x) = \lambda[1 - e^{(-\frac{x}{\lambda})}]$, for large λ the utility function becomes linear as if there were symmetry about risk tolerance, whereas for small $\lambda \geq 0$, the bank would be risk averse. "Interest" for obvious reason of "fairness" is based on the underlying assumption of risk neutrality.

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

$$\Theta = \sum_{i=1}^{n} \Pr[\nu(x_i)] * X_i \tag{7.3}$$

where

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

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). 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 potential flaws in the LDA model of ORM by employing CA, by assuming managerial risk-taking attitudes are more risk averse.

More precisely, the goal is to use CA - an unsupervised learning algorithm to train a model on the unlabeled data and then use the results to learn deep hierarchies in the data and extract interesting features (Coates & Ng, 2012) found during banking

operations. Clustering of determining features are expected to distill whether due to risk aversion e.g., with the introduction of new products or in changes to the business mix of the portfolio (e.g. mergers and aquisitions, trade terminations, allocations or disinvestments) OpRisk controls would be in place to over-compensate for persistent loss event types and this would show in reducing losses particularly in those specific loss types or business lines over time.

K-means clustering of OpRisk loss data

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.

Proposition 7.6.1:

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.

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., This is a testable proposition which fits normal behavioral patterns and is consistent with risk averse behaviour.

Exploratory data analysis

The data is limited to the training dataset over the interval 01 January - 31 March 2013, in Figure 7.1, portrays detail of the trend of OpRisk losses against exposures for each of the 1631 observations and 16 variables. In the first plot, transactions 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

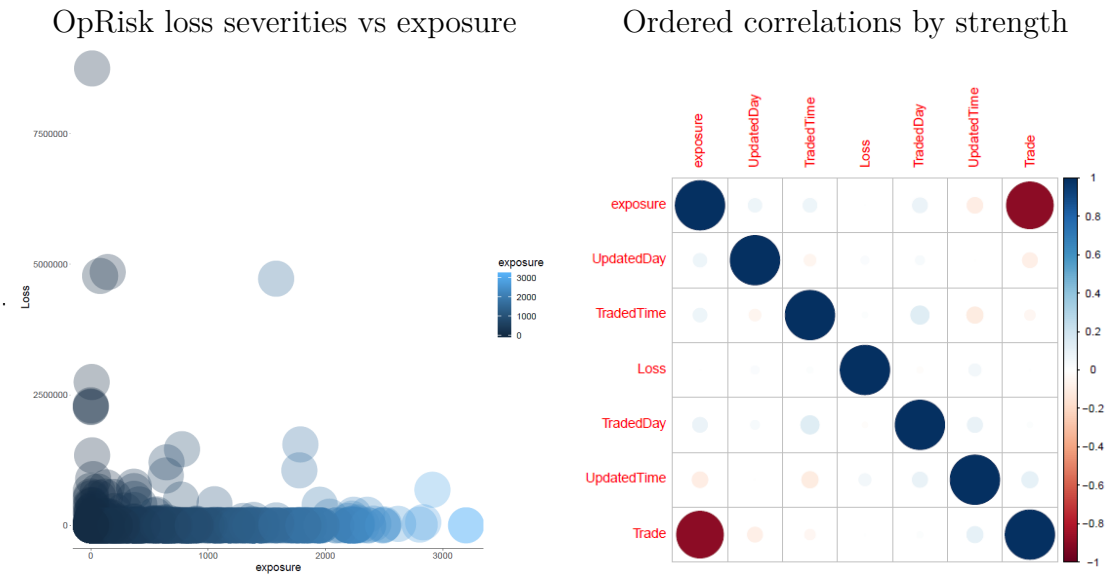


Figure 7.1
Graphically displayed correlations by strength and a plot of OpRisk loss severities vs exposure

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.

Choice of appropriate dataset attributes

The characteristics of the traded transactions or of the associated risk correction event are given by the following variables: Trade, UpdateTime, UpdatedDay, Trad-

edTime, 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 EventType-CategoryLevel 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.

Cluster analysis learns the deep heirachies of input features used in calculating OpRisk event frequencies and severities of losses and use the results to extract interesting features from the data that constitute risk tolerance perceived 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.

The estimation of *k*-means clustering solutions

A K-means clustering analysis is a methodology for discovering cohesive K clusters or subgroupings of observations in the Oprisk loss data . The resulting k clusters are represented by the mean or average values of each of the variables. The user whose task it is to specify k, may be hard done to guess it, or in practice they may obtain a priori the knowledge of how to select the appropriate k in advance.

Determining the number k *of clusters present*

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 7.2 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) - Silhouette Statistic(SS):Sum is a measure to approximate the optimal number of k groups by the Silhouette Statistic SS (blue line), as a performance metric indicates that in the near end approaching $k = 3$ and beyond $k = 4$ clusters the model overfits: Its computes the absolute error which is initially large, then monotonicaly decreases to zero at a point within the $k = 3$ and $k = 4$ range, it then begins to increase subsequent to the point where the Diffprevious Sum(WithinSS) (red line) intersects viz., and continues as such beyond $k = 4$ clusters. The task certain between $k = 3$ 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 beyond 4 clusters.

Determining a final clustering solution

Since the number of clusters tie at $k = 3, 4$, the next step for obtaining a (final) clustering solution is obtained on both choices of k accessed via the `rattle` R package. The `kmeans()` function calculates distances, dissimilarities and/or proximity between each entity to be clustered. For example, Table 7.1 is displays of the results of the "within cluster sum of squares" measures is given by

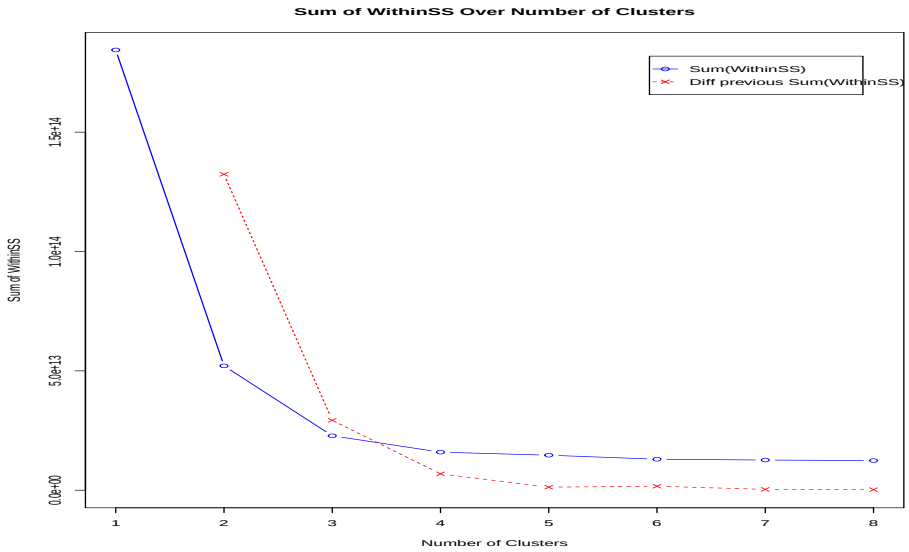


Figure 7.2
Finding the optimal number of k groups by Plotting the Within-groups sums of squares (Sum(WithinSS) - Silhouette Statistic(SS):Sum) vs. the number of clusters extracted. the sharp decreases suggest that a three-cluster solution may be a good fit to the data.

$$\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}(x_{ij}-\overline{x}_{kj})^2 \tag{7.4}$$

where S_k are the Oprisk observations in the k^{th} cluster and $\overline{x}_{kj}$ is the j^{th} variable of the cluster center for the k^{th} cluster.

Table 7.1 displays how well the k-means clustering uncovers the structure of the OpRisk data is contained in the average.between value measuring is the total dispersion between the clusters, and is found to maximise at 70%. By assigning the samples to 4 or 3 clusters rather than the maximum number achieved a reduction in sums of squares of 70. Whilst this represents a considerable improvement, the $K = 3$ means average.within measure minimises within sum of sqaures by 44% while the K=4 means measure decreased down to 35%. As such, the K=4 means performed better as the $k = 3$ means in this instance.

Table 7.1
A table showing Within cluster sum of squares sizes and average.within & average.between statistics

K=3	Within cluster sum of squares: 45.42597	45.75955	43.22250	45.79918
average	between	70%		
	within	44%		
K=4	Within cluster sum of squares: 85.02703	54.09784	139.64631	
average	between	70%		
	within	35%		

Descriptive analytics - intepreting the clusters

Clustering were formed applying 4 means and 3 means clustering to a dataset containing 5 covariates on 2330 OpRisk loss samples viz., UpdateTime - the transaction time over a 3 month period given by the realised OpRisk hazard; TradeTime - the time originating the trading transaction; Activity - an ordered discrete number representing the degree of OpRisk hazard based on the loss event type per business line; & lastly the exposure variable as defined in definition 2.10.1.1.

K-*means clustering models for OpRisk loss data*

The model from a K-means consists of Euclidean distances representing each cluster using a centroid (in vector of varaible means). The K-means clustering point of view consisting of four (4) vector model is visualized in figure 7.4 and the main content of the cluster means text view listed in Table 7.2. similarly K-means clustering point of view consisting of four (3) vector model is visualized in figure 7.3 and the main content of the cluster means text view listed in Table 7.2.

Table 7.2 generated below provides a summary of the mean values and a model for each of the k-means points consisting of k vectors of the mean values for each variable.

Table 7.2
K-means cluster sizes (means and centers) of the OpRisk loss banking data

K=3	Data means:				
	TradeTime	UpdateTime	Activity	Loss	exposure
	0,72175379	0,450255857	0,45872343	0,003032397	0,12190767
	Cluster centers:				
1	0,790161	0,7215703	0,4381482	0,004187844	0,06683204
2	0,3764403	0,5236885	0,5983117	0,002579653	0,47528165
3	0,7753703	0,2441683	0,4322296	0,00237686	0,05687265
K=4	Data means:				
	TradeTime	UpdateTime	Activity	Loss	exposure
	0,72175379	0,450255857	0,45872343	0,003032397	0,12190767
	Cluster centers:				
1	0,7877611	0,7496131	0,3374133	0,004425481	0,07242014
2	0,7662503	0,3731121	0,8452993	0,000607677	0,0617052
3	0,775061	0,2491971	0,271503	0,003567299	0,06173064
4	0,3607023	0,5438919	0,5923437	0,00273211	0,49524782

Interpretations from a K = 4 K-means model clusters

In comparing the four clusters cohesion measures for similarities between data points within the earlier recorded OpRisk loss (UpdateTime) periods: In cluster 3 we see OpRisk loss activity tendencies associated with the most severe OpRisk hazard loss types, which as expected reflects in the magnitudes of losses. According to ground truth defined by Proposition 7.6.1, the degree of Oprisk loss severities per cluster should go down in similar OpRisk hazard loss types (of comparable exposure) recorded late in the (UpdateTime) section.

Cluster 1 which is closest in OpRisk hazard loss type similarities to cluster 3,

appropriately recorded late in the period, the prediction result suggests Oprisk loss severities are slightly higher than expected later in the UpdateTime period. The sets back our version of ground truth, moreso since Activity at cluster 1 is at a lower degree OpRisk hazard loss type. A deeper parsing of OpRisk hazard loss type could serve as possible explanation of this inconsistency; given the greater exposures in cluster 1 on a lighter note, the degree of Oprisk loss severities do not increase drastically to dispute Proposition 7.6.1 altogether. Especially given the general ground rule of increasing degrees of Oprisk loss severities in the very late period due to increasing exposures.

The degree of Oprisk loss severities in recorded OpRisk loss (UpdateTime) periods drop drastically in cluster 2; notwithstanding we see losses associated with OpRisk loss activity tendencies associated with the least severe OpRisk loss hazard types residing in agency and other banking business lines. Cluster 4 measures are associated with the longest dated OpRisk transactions and are dated midway through the recorded OpRisk loss (UpdateTime) periods, as such extreme exposures have significant impact on the degree of Oprisk loss severities even with moderate to low associated OpRisk hazard loss types. The ground truth consistency with Proposition 7.6.1 is remarkably demonstrable here, since these trading transaction are longdated and OpRisk hazard loss types have possibly gone through several adjustment loops; it makes sense in our framing of the ground truth that the losses should drop significantly over time compared to cluster 3, since due to recurring OpRisk hazard loss types, they degree of Oprisk loss severities have adjusted for their natural risk aversion.

Interpretations from a K = 3 K-means model clusters

Clusters 3 and 1 have resembling features ideal in establishing core findings related measuring the accuracy in Proposition 7.6.1. Disappointingly, the degree

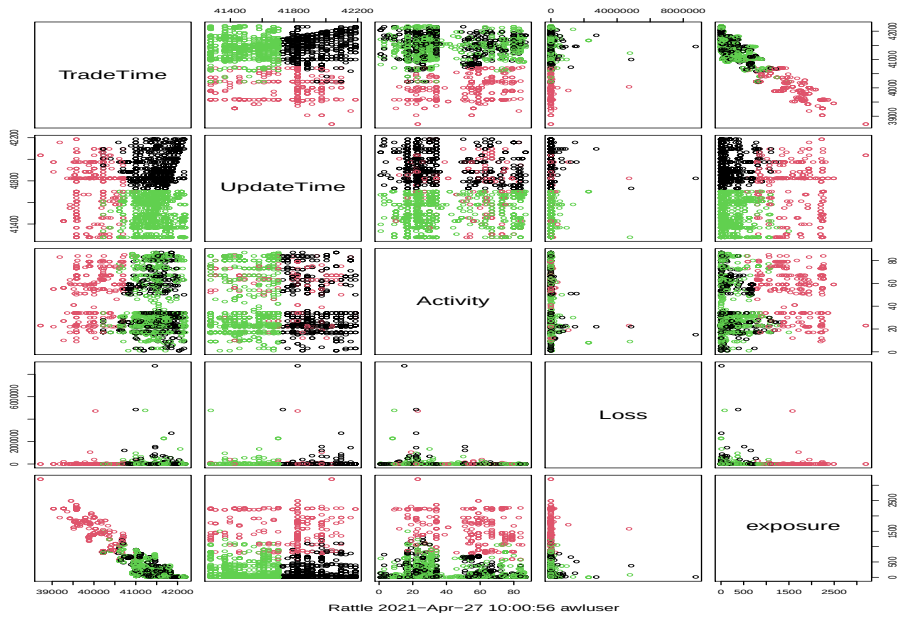


Figure 7.3
]
A scatterplot matrix for the k-means clustering of size 3 consisting of 369 loss event frequencies amounting to R 61 534 745 P&L loss impact.

of Oprisk loss severities in successive recorded OpRisk loss (UpdateTime) periods between clusters 3 and 1 approximately double in magnitudes for closely matching OpRisk hazard loss types and exposures. This may not come as a surprise though as the fewer data clusters as demonstrated earlier in Table 7.1 reduce compactness, as reflected in an their inability to distinguish between features in finer detail.

Model results using discriminant coordinates

Since OpRisk losses datasets are often characterised by rare and large losses compounded by small size datasets, the k-means clustering performance was significantly influenced by very large outlier losses e.g., ZAR$8,752,614$ realised very late in period (19 March 2013). The k-means clustering approach is particularly sensitive to outliers because it is based on means. A more robust representation is represented by a discriminant coordinates approach. The bivariate cluster plot presented by figure

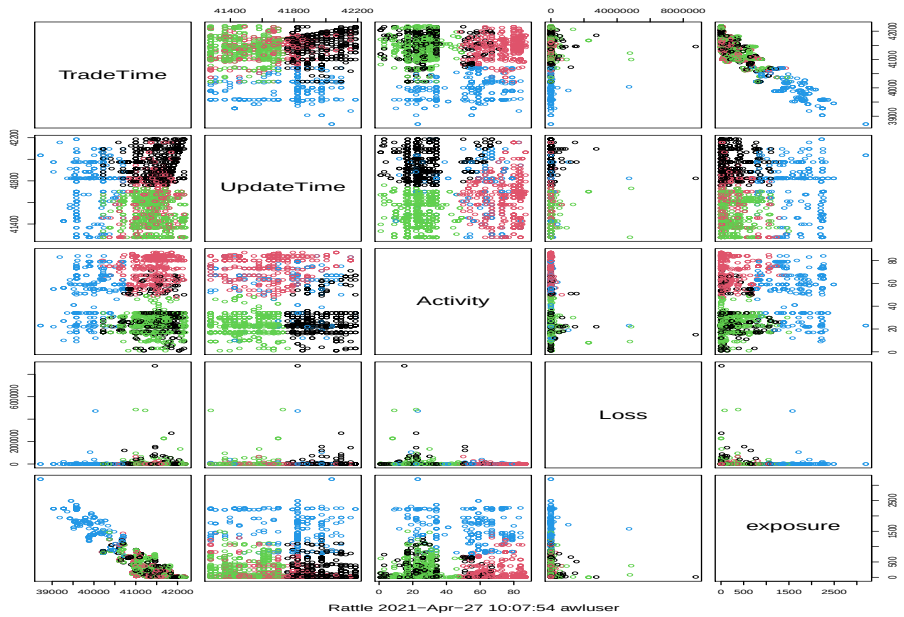


Figure 7.4
A scatterplot matrix for the k-means clustering of size 4 consisting of 369 loss event frequencies amounting to R 61 534 745 P&L loss impact.

7.5 is applied to the OpRisk loss data.

The bivariate plot is created by plotting the coordinates of each observation on the the first two principal components of the 5 Oprisk loss variables. Each cluster is represented by an ellipse with the smallest area containing all its points. Also, the two components only explain 60% of the dispersion around points, decreasing 10%in performance versus (the 70% point variability for)k-means.

Conclusion

Existing models in banking do not learn from past losses and are indifferent to risk tolerance levels. The interesting question is to test whether there is evidence from these data, through an unsupervised learning algorithm providing a k-dimensional sparse representation through a falsifyable Proposition, that the implied spreading of un-checked OpRisk negligence persists or not: The findings present hard positive

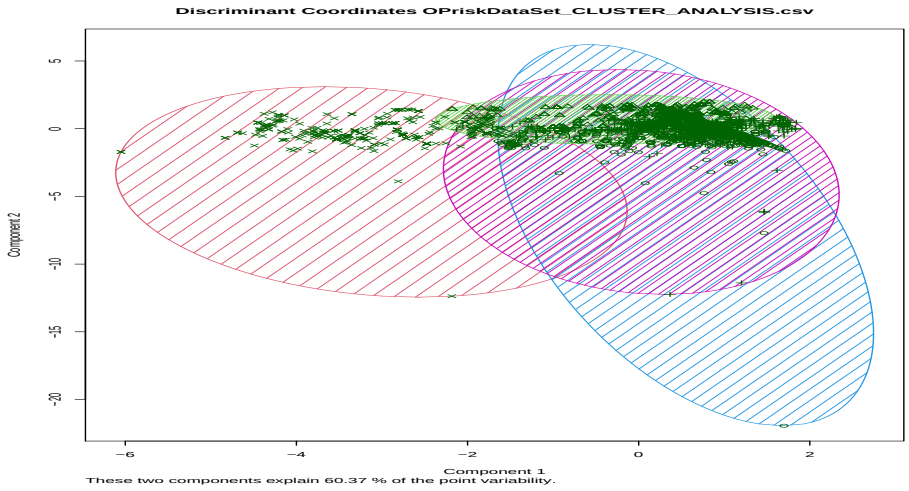


Figure 7.5
Discriminant coordinate plot for the four (4) group clusters of the OpRisk loss banking data

evidence obtained through an unsupervised learning k-means clustering modeling framework which learns from past losses that degrees of OpRisk loss severities are slowing based on successive recorded OpRisk loss (UpdateTime) periods, establishing Proposition 7.6.1 in truth. Furthermore, the new clustering model's predictive nature of learning techniques determining that OpRisk's are slowing move banking risk tolerances away from risk neutrality through a modeled future that over-provides for OpRisk hazard loss types that have already occurred, suggesting that banks are more risk averse.

CHAPTER 8
DISCUSSIONS, CONCLUSION AND FUTURE RECOMMENDATIONS

Introduction

This study calls for machine learning applications in OpRisk measurement through an EBOR framework by way of "learning" on three major variables; OpRisk exposure, loss frequency distributions, and loss severity distributions, via relationships that exist between LDA modeling and exposure-based modeling, which could potentially establish an improved practice to determine illicit conduct in investment banking operations. It then provides empirical evidence for the purpose of advancing a discussion on the compatibility of OpRisk attitudes in the context of risk aversion with the thinking of scholars in the field of behavioral economics.

True to the presentation (in the final sentence of Chapter 1), this chapter proceeds from Chapter 3 with a discussion of the findings on the hypothesis, propositions, and research questions of interest, arriving here in the current discussion and proceeding henceforth. Moreover, in this concluding chapter theoretical contributions, empirical contributions, and the methodological contribution are provided along with recommendations and suggestions for future research.

Holistic profile overview

The study's purpose is revealed through an intense series of ML analytics and new criteria toward an OpRisk framework that enlarges quarterly OpRisk data into annual data by adding "fake data" points through an artificially intelligent (AI) data augmentation technique, prolonging the duration of the dataset as a potentially rem-

edy for scarce and low quality datasets inherent in OpRisk management. Motivated by this abundance of larger and better quality datasets, for the first time an integrated framework for OpRisk loss assessment using data-intensive techniques is realised which offers an important tool for OpRisk management, simultaneously working toward Embrechts et al. (2018)'s call to support industry with a new class of OpRisk model that captures forward-looking aspects in real time. Three different machine learning techniques viz., GLM, GAMLSS, and CA, have been employed using the R software package.

Hoohlo (2014) applied traditional statisical parametrization techniques to calculate OpVaR based on Chavez-Demoulin et al. (2016) extreme value theory (EVT) method arriving at a proxy of ZAR3.02Bn. The new AI exposure-based OpVaR estimation technique on the augmented yearly data also yields an OpVaR value of precisely ZAR 2.9Bn. Thus, the research results reveal based on carefully designed accuracy metrics, that an AI guided data augmentation method should be permissible in the event of scarce and low quality OpRisk loss datasets.

It further evaluated the magnitudes of the effects of internal control factors inherent in each predictive, moderating and/or target variable on the aforesaid successes based on its performanc versus dynamic EVT fitting, which is very much based on the traditional EVT-POT model developed by Chavez-Demoulin et al. (2016), and various mix of supervised and unsupervised ML models viz., GLM, GAMLSS, and CA. For comparability, both Chavez-Demoulin et al. (2016) and the current research consider dynamic models based on covariates and in particular concentrate on the influence of internal root causes that prove to be useful for their proposed methodology.

Qualitative profile of the loss distribution data collection exercise (LDCE)

Twenty-one (21) key risk indicators (kri's) with eight feature groups including person identification, trade origination, root causes and market value sensitivities make up a comprehensive list of covariates that contribute to the accumulation of OpRisk loss events across eight (8) business units. For each risk event there is information about: trading risk exposure, trading characteristics, causal factor characteristics and the losses created by these factors viz., TradeID, UpdateTime, UpdatedDay, TradeTime, TradedDay, DeskID, CapturedBy, TradeStatus, TraderID, Instrument, Reason, Loss, etc. Using several classes of data-intensive techniques approximating the weights of the risk factors; i.e., the input features of the model on the LDCE data, we uncover never seen before findings and a tractable method of achieving greater model accuracy. In-built "learning" AI capabilities of ML models are better equipped to extract features in the data at this level of granularity and assess the impacts of exposure-based treatments, and other forward-looking aspects for improved forecasts of OpRisk losses.

Qualitative profile on forward-looking aspects compared to normal models

Using these computationally intensive OpRisk measurement techniques and mixing with a theory is not a new approach for modeling, particularly in calculating OpRisk RC; as evidenced through Agostini et al. (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 (ACT). However, applying AI is an original way of looking at the approximation issue as opposed to ACT. The idea at the basis of the use of AI, is to advance the very recent literature that a better estimation of

the OpRisk RC measurement can be obtained by testing the accuracy of these techniques against traditional "*ARMA*" statistical techniques, combining the historical simulations with forward-looking aspects through models that are weighted averages of the covariates, advocating for the combined use of both experiences. Consequently, 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 OpRisk. (Einemann et al., 2018).

Qualitative profile on risk averse attitudes in banking instead of risk-neutrality

Using these computationally intensive OpRisk measurement techniques the study seeks through empirical evidence to test the proofs of Beinhocker (1997)'s mathematics applied to economic phenomena of theories relating supply and demand to individual's rationality and their ability to maximize utility, insisting that when choosing between risky prospects systematic violations of EUT as evidenced in substantial bodies of evidence in recent work, implies a slightly different scientific intepretation which describes how individuals actually make decisions, namely Prospect Theory (PT). The essential feature of PT are assumptions which are more compatible with basic principles of perception and judgement for decisions taken under uncertainty, whereas the unsupervised ML technique (cluster analysis) is well suited to reveal additional chance probabilities determined through the natural clusters of unknown data feature findings from which new discoveries are made, consistent with those assumptions of PT, building towards a robust, precise, accurate, and scalable OpVaR measurement.

Discussion

From the presentation outlined in Section 1.8 Chapter 1, the following items are presented according to the structure of the (5) successive chapters - Chapters 3 - 7: Two (2) of the afore-mentioned chapters consist of OpRisk loss representation categories and three (3) OpRisk loss factor model categories, each with 15 risk variables. The findings for the representation categories, Chapters 3 and Chapter 7 are discussed according to the same structure as per propositions, research questions and the hypothesis detailed therein; and those findings related to the cases in the OpRisk loss factor model categories with the structure of the propositions detailed therein.

OpRisk loss representation: Exposure

This representation category focuses on the exploration of various descriptive statistics of the covariates. More specifically on how characteristics of the exposure variable moderate the relationship between predictive risk factors (covariates) and the target variable (LossIndicator).

Exposure response on risk influence

Conclusion 3.7.1: Equivalence between the two specifications 3.11 and 3.10 upholds Proposition 3.7.1 through the resultant relationship: $\lambda_i = \frac{d}{\mu_i}$: Demonstrating that time to detection (exposure d_i) is a quantity that is roughly proportional to the risk.

The exposure variable behaves as the proportionality constant i.e., d_i between the rate λ_i, and the magnitudes of the $p + 1$ parameters. Therefore, the exposure doesn't only represent a mechanism affecting the algorithm directly before and after regression estimation, but when it is doubled, it will indeed consequently double

OpRisk. According to literature (Parodi, 2014), there should be an expectation when time to detection (exposure measure for each independent r.v Y_i) doubles whilst everything else (e.g. interest on an interest rate swap) remains the same, for the risk also doubling due to direct proportionality. That is to say, the exposure measure is a known initial condition (constant) which is readily incorporated into the mean estimation procedure and is a quantity that is roughly proportional to the risk (Parodi, 2014).

Visual exposure response analysis

Conclusion 3.9.1: The empirical finding suggests a moderate exposure curve is a reasonable fit i.e., the exposure curve is neither strongly concave nor too close to linearity. It follows that a deductable of $d\%$ will achieve a reduction in the risk ceded (LGE) by the bank of exactly $d\%$, as conceptualised in Hypothesis 3.9.1, exposure is directly proportional to a reduction in the risk ceded by the bank.

The empirical finding determined through visual analysis illustrated in Figure 3.1, suggests a moderate exposure curve is a reasonable fit i.e., the exposure curve is neither strongly concave nor too close to linearity. A strongly concave exposure curve tells us that for even a small deductible (10% of the maximum probable loss), say due to an minor adjustment in the exposure (time to detection), whilst everything else (e.g. interest on an interest rate swap) remains the same for small effects in predictor variable/s, will achieve a reduction in the risk ceded (LGE) by the bank, and the probability of a total loss is minimal (Parodi, 2014).

On the other hand, if the exposure curve has no concavity at all, then the other extreme of a curve much nearer to the diagonal of a linearly shaped relationship, where a deductible of d% will achieve a reduction in the loss of exactly d%, and all losses are total losses. In this scenario deductibles are far less effective in reducing the

risk ceded by the bank. This translates into a growth phenomena which can either be flat, or increasing or decreasing with λ_i, but it cannot be increasing for some value of λ_i decreasing for others.

OpRisk loss frequency: Exposure-based OpRisk GLM model

The findings, conclusions and prectical implications discussed below relate to the development of an EBOR GLM modeling framework which not only captures past losses but also "learns", and therefore introduces forward-looking aspects to the OpRisk prediction problem.

Conclusion 4.6.1: EBOR OpRisk treatments through GLM modeling frameworks learn from data to provide a forward-looking LGE estimation model of the risk matrix solution to LDA loss frequency modeling problem, allowing for better estimation of effect size and measures of precision for inferences made in Proposition 4.6.1 to be valid.

The results from the confusion matrix 8.6 for the better fitting Binomial model vs the Poisson model case, of a greatly improved Sensitivity of 90.94% and Specificity of 36.51%, and Confusion Matrix 8.6 showing a slightly more concave ROC curve infer exceptional goodness-of-fit whereby model accuracy has been substantially improved by about 5%. The minimal K-L information loss using information-theoretic criteria i.e., AIC, AIC_c,QAIC is easily computed using model averaging based on Akaike weights.

Practical implications for OpRisk research

The selected model "AModel", (given a priori issues modeling the scientific proposition 4.6.1) supports the notion that "learning" from looking at evidence in

the deep hierarchies of features in the transformation matrix, $e^{\beta_i \mathbf{X}^T}$, acting on the exposure, achieves a "forward-looking" solution to the approximation of the LGE and therefore the random mechanism which generates the loss frequency distribution. GLM model frameworks with EBOR treatments aided by multimodel performance evaluation are found to effectively replace historical loss severity curves obtained from historical loss counts by models with learning capabilities. A better estimation model than the traditional "best model" approach focuses several competing models in the estimation process of relationships of variables, and on the estimation of effect size and measures of precision.

OpRisk loss severity: Exposure-based OpRisk GAMLSS models

The findings, conclusions and practical implications discussed below discuss efforts toward the development of an EBOR GAMLSS models accounting for variations around the mean of severities in order to account for heteroskedacity; notwithstanding an inappropriate assumption of a constant scale parameter for different ET/BL categories of OpRisk that do not capture tail behaviour properly.

Conclusion 5.4.1: AIC model selection criteria coupled with model diagnostics tools in the residuals of the fitted models for EBOR OpRisk treatments through GAMLSS modeling frameworks i.e., normalised quantile residuals and the worm plot, verify the $GB2()$ as an adequate GAMLSS model fit which outperforms its rivals given the reported information criteria, standard normal true residuals and the regions identified for an appropriate fit by the worm plot, and therefore are a complementary flexible model designed to "learn" from data providing forward-looking solution to LDA loss severity modeling, thus upholding Proposition 5.4.1.

GAMLSS modeling frameworks with EBOR treatments effectively replace his-

torical loss severity curves obtained from historical loss counts by models with "learning" capabilities which are forward-looking and account for variations around the mean of severities, notwithstanding capturing heavy-tailed behaviour properly compared to "normal" (ARMA) techniques.

Practical implications for OpRisk research and for practitioners

GAMLSS modeling frameworks with EBOR treatment for the first time eliminate the variance-bias dilemma wherein generality always came at a high cost with respect to stability by eliminating the theoretical concern involving highly unstable parameter fitting and findings which revise universally held views of infinite variance (Voudouris et al., 2012).

Thus, the revelation to risk practitioners to react appropriately in anticipation of "odds against" loss expectations/probabilities in one BL/ET mix combination (as the appropriate UOM) over another, based on whether their loss distributions were more or less variable than the other, now becomes possible by looking for evidence in the model design to affirm that truth i.e., OpRisk practitioners can now manage in favour of more flexible heavy tailed distributions with appropriate goodness of fit, thereby affirming "odds against" loss variations of one BL/ET combination against another.

OpRisk losses modeling and capital calculation depending on augmented covariates

The findings, conclusions and practical implications to be discussed below show how due to data scarcity an intense data augmentation analysis was embarked on that prolonged the 3-month data duration to cover and annual period as a remedy, and to reproduce plausible metrics for capital calculation based on proxy of $ZAR3,02$bn as obtained in Hoohlo (2014).

Conclusion 6.5.1: An improvised (AI aided) augmentation dataset as required to the extent data is deemed essential for "learning", in light the existing trade-off of more data versus quality (because they are constructs not observations), gets around the data hungry requirements of ML and eliminates the HFLS & LFHS dilemma of large and rare loss events, resolving the variance-bias problem associated with small sample sizes, aggregate units of measurement, and a wide spectrum of the loss BL/ET mix.

The findings suggest that dataset augmentation practically improves the ORMF's methodology, to the extent and accuracy inherent in modeled parameters for forward looking datasets obtained from the learned GLM and GAMLSS models. The results indicate that at worst the new framework's performance matches Hoohlo (2014)'s OpRisk VaR measure obtained on the real 3 month dataset through the application traditional statistical techniques to calculate an OpVaR proxy of $\sim$ 3 billion South African Rand opening up the horizon for the exploration modeling beyond historical accounts of significance, but also incorporating forward-looking aspects.

Practical implications for OpRisk research and for policy

As such, the HFLS & LFHS dilemma of large and rare loss events consisting of small sample sizes, aggregate units of measurement, and a wide spectrum of the loss BL/ET mix to model are eliminated, and in place of their negatively compounding effects are numerous positive implications borne out of OpVaR model stability; such as in replacing the persistent LDA capital estimates "upward bias" shortcoming of overcapitalizing for realised losses only after they have occurred, thereby innovating toward the estimation of OpRisk loss frequency and severity distributions and potentially gaining favour over traditional approaches and disrupting the modus operandi in the banking sectors operations frameworks.

Furthermore, future projections which look reasonable and believable in practical terms, demonstrates that data augmentation, which is also cheap and efficient, set forth an improved ORMF when coupled with ML for the purpose of measurement and comparability of data achieved through heuristics, lowering costs of monitoring OpRisk in banking as recommended by BIS. Thus, adding to "best practice" debates underway toward benefiting OpVaR model stability and in aid in the estimation of frequency and severity distributions of OpRisk losses.

An AI aided framework for OpRisk

Conclusion 6.7.1: An AI aided EBOR framework with forward looking modeling aspects improves operational risk capital-at-risk (OpCaR) requirements over traditional models who lack AI "learning" capabilities hence suffer from this as a limitation.

Evidently, the results indicate the AI aided framework improves or at worst matches the OpRisk capital adequacy requirements banks, implied by a $\geq 10\%$ capital requirement savings i.e CAR requirement $ZAR2,9$bn versus Hoohlo (2014)'s OpRisk VaR $ZAR3,02$bn measure obtained on the real 3 month dataset through the application traditional statistical techniques, and is also appropriately calculated, opening up the horizon for the exploration modeling beyond historical accounts of significance to include forward-looking aspects.

OpRisk loss representation on risk tolerance: A k*-means clustering algorithm*

The core of this study is to employ a representation learning k-means clustering model as the best model to fit the data for the task of providing a k dimensional vector that relates to features relevant to two major contributions: [1] AI models are forward looking compared to normal statistical models due "learning". [2] Instead of

existing OpRisk models which assume banks are risk neutral, the proposed AI aided OpRisk model captures possibilities that they may be more risk averse.

Conclusion 7.6.1: 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.

The findings present hard positive evidence obtained through an unsupervised learning k-means clustering modeling framework which learns from past losses that degrees of OpRisk loss severities are slowing based on successive recorded OpRisk loss (UpdateTime) periods, establishing Proposition 7.6.1 in truth. Furthermore, the new clustering model's predictive nature of learning techniques determining that OpRisk's are slowing move banking risk tolerances away from risk neutrality through a modeled future that over-provides for OpRisk hazard loss types that have already occurred, suggesting that banks are more risk averse.

Practical implications for OpRisk practitioners and for policy

The optimal K=4 k-means clustering model was best suited to uphold Proposition 7.6.1, finding it appropriate to declare through firm evidence in the data that things are improving and OpRisk losses slowing, dispelling the notion of the unchecked-spread of negligent behaviour in OpRisk. As a result of the learning the model predicts a better representation of the future providing risk practitioners the means to identify losses that would otherwise be missed, undercapitalised under capital adequacy requirements, or avoiding capital buffer redundancies in overcapitalised realised losses after they had occurred, thereby optimising OpRisk control functions' performance; infusing innovation and in change management by releasing human capital to focus on dilemma that require human judgement.

Limitations of the study

Though this study makes significant empirical, theoretical & methodological contributions, the dataset itself is not devoid of drawbacks:

- The sample was dominated by internal events occurring in Q1 of an emerging economy investment bank, namely from January to March 2013 in a South African investment bank which might pose a challenge when generalising to the remaining quarters in the year, and to more advanced economies like those in Europe, the United States of America, etc.
- Due to the timespan, the study omits many risk variables; especially those rare and far between events (black swans), that may result in catastrophic losses which would potentially contribute significantly to the predictive capacity of the OpRisk model. This puts to question the ”comprehensive”, all encompassing predictive capabilities of the OpRisk assessement model.
- The sample frame was dominated by risk practitioners subject to a sole institutional educational risk discipline which could be susceptible to bias, and thus caution must be exercised when generalising the findings.
- Although data was collected/extracted from two separate and independent research instruments, in few cases the same instrument and same respondents appeared for both covariate(s) and responses. Though precautions were carried out to ensure no issue of common method and response bias, careful intepretation of findings should still prevail.

Contributions and Recomendations

Many scholars and practitioners in OpRisk literature and industry support the call for a new class of OpRisk model that goes beyond simplistic methods, wanting to integrate new data-intensive AI/ML analytic techniques capable of extracting forward-looking actionable insights unique to OpRisk management systems. In response, this study aimed at developing a new AI/ML exposure-based OpVaR estimation technique: Consequently, a series of theoretical, empirical and methodological contributions are successively made to the OpRisk literature viz.:

[1] An empirical contribution marking the influence of a moderating variable viz., the *exposure* variable - a monotonic transformation of the mean λ_i, otherwise interpreted through a proportionality constant translated into a growth phenomena which can either be flat, increasing or decreasing with λ_i. [2] An empirical contribution to exposures-based OpRisk loss frequency modeling of forward-looking loss-given-event (LGE) estimates as a result "of the learning" in the model. [3] A methodological contribution by introducing a data augmentation algorithm that conveys statistical advantages (it naturally conveys forward looking output vectors representing learned frequencies and severities), circumventing the HFLS & LFHS OpRisk data dilemma and generalising on all distributional type attributes based on precise goodness-of-fit measures at no cost to OpVaR stability. [4] An theoretical contribution based on a generalised GAMLSS. [5] A theoretical contribution by addressing a comparative gap in literature which assumes banks' OpRisk tolerance is risk neutral by confirming banks OpRisk tolerance is risk averse through a falsifyable test.

Conclusion

In conclusion, investors are likely to be less influenced by accuracy measures reflecting risk tolerances which are inconsistent with their framing of the ground truth (risk neutrality as opposed their natural risk aversion). As such, new metrics may persist in emerging digital currency investing which more accurately reflect company fundamentals. This implies that firms should be wary of contagion effects amidst their customers due to risk averse as opposed to risk neutral tolerances i.e., contagion to systems which obey better suited governance mechanisms which tend to control for customer's risk tolerance while others do not.

REFERENCES

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

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

Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *36*(3), 332–339.

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

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

Babcock, B. A., Choi, E. K., & Feinerman, E. (1993). Risk and probability premiums for CARA utility functions. *Journal of Agricultural and Resource Economics*, 17–24.

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

Banking Supervision (BCBS), B. C. on. (2006). *International convergence of capital measurement and capital standards: A revised framework (comprehensive version).(june). Available at*.

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

Beinhocker, E. D. (1997). Strategy at the edge of c ao. *The McKinsey Quarterly*, (1), 25.

Boettrich, S., & Starykh, S. (2017). Recent trends in securities class action litigation: 2016 full-year review. *NERA, New York*.

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

Cameron, A. C., & 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., & 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., & Nešlehová, J. (2006). Quantitative models for operational risk: Extremes, dependence and aggregation. *Journal of Banking & Finance*, *30*(10), 2635–2658.

Coates, A., & Ng, A. Y. (2012). Learning feature representations with k-means. In *Neural networks: Tricks of the trade* (pp. 561–580). Springer.

Committee, B. et al. (2001). *QIS 2 - Operational Risk Loss Data – 4 May 2001*. 17–18. Retrieved from `https://www.bis.org/bcbs/qisoprisknote.pdf`

Committee, B. et al. (2003). *The 2002 Loss Data Collection Exercise for Operational Risk: Summary of the Data Collected*. 1–41. Retrieved from `https://www.bis.org/bcbs/qis/ldce2002.pdf`

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

Committee, B. et al. (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., et al. (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,.

Cruz, M. G., Peters, G. W., & Shevchenko, P. V. (2015). *Fundamental aspects of operational risk and insurance analytics: A handbook of operational risk*. John Wiley & Sons.

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

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

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models*. Chapman; Hall/CRC.

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`

Eilers, P. H., Marx, B. D., et al. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, *11*(2), 89–121.

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

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

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the

variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987–1007.

Evans, M., Hastings, N., & Peacock, B. (2001). *Statistical distributions*. IOP Publishing.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, methods and applications*. Springer Science & Business Media.

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., & Roncalli, T. (2001). *Loss distribution approach for operational risk*.

Frachot, A., Moudoulaud, O., & Roncalli, T. (2004). *Loss distribution approach in practice*.

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

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

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

Ganegoda, A., & Evans, J. (2013). A scaling model for severity of operational losses using generalized additive models for location scale and shape (GAMLSS). *Annals of Actuarial Science*, *7*(1), 61–100.

Gigerenzer, G., & 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*.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). MIT press Cambridge.

Gous, N. (2019). *Standard chartered pleads guilty to fixing the rand* (pp. 1–2). Retrieved from `https://www.timeslive.co.za/sunday-times/business/2019-02-05/standard-chartered-pleads-guilty-to-fixing-the-rand/`

Group, A. et al. (2013). AMA quantification challenges: AMAG range of practice and observations on "the thorny LDA topics." *Munich: Risk Management Association*.

Harvey, A. C. et al. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, *44*(3), 461–465.

Hemrit, W., & 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. (2014). *A new internal data measure for operational risk: A case study of a south african bank* (Master's thesis).

Jongh, R. de, De Wet, T., Raubenheimer, H., & 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. (1973). *Attention and effort* (Vol. 1063). Citeseer.

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

Kahneman, D., & 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.

Kennett, R., & Carrivick, L. (2018). *Annual banking loss report: operational risk loss data for banks submitted between 2012 and 2017* (pp. 1–16). Retrieved from `https://managingrisktogether.orx.org/orx-loss-data/annual-banking-loss-report`

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

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling*, *13*(4), 275–303.

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

Larsen, P. (2019). Maximum likelihood estimation error and operational value-at-risk stability. *Journal of Operational Risk*, *14*(1).

Laubsch, A. J., & Ulmer, A. (1999). *Risk management: A practical guide.* RiskMetrics Group.

Lewis, N. D. C. (2004). *Operational risk with excel and VBA: Applied statistical methods for risk management,+ website* (Vol. 244). John Wiley & Sons.

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

Mandelbrot, B. B. (1997). The variation of certain speculative prices. In *Fractals and scaling in finance* (pp. 371–418). Springer.

Martin, P. H. (2009). As risk management evolves, is operational risk management important? *The Journal of Operational Risk*, *4*(4), 75.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, *29*(2), 449–470.

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

Migueis, M. (2017). Forward-looking and incentive-compatible operational risk capital

framework. *Available at SSRN 2964945*.

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

Mori, T., Harada, E., et al. (2001). *Internal measurement approach to operational risk capital charge*. Bank of Japan.

Nelder, John A., & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, *74*(2), 221–232.

Nelder, John Ashworth, & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Ohlsson, E., & Johansson, B. (2010). *Non-life insurance pricing with generalized linear models* (Vol. 2). Springer.

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.

Paulus, M. P., Rogalsky, C., Simmons, A., Feinstein, J. S., & Stein, M. B. (2003). Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. *Neuroimage*, *19*(4), 1439–1448.

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

Rigby, R. A., & Stasinopoulos, D. (2010). A flexible regression approach using GAMLSS in r. *Lancaster: University of Lancaster*.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554.

Rigby, R. A., Stasinopoulos, D., Heller, G. Z., & De Bastiani, F. (2017). Distributions for modelling location, scale, and shape: Using GAMLSS in r. *URL Www. Gamlss. Org.(last Accessed 5 March 2018)*.

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*.

Rosa, P. (2012). Litigation-credit risk structural models in op risk-an exposure-based method using a credit risk structural model may help operational risk managers predict underwriter litigations. *Operational Risk and Regulation*, *13*(10), 34.

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.

Securitisation, A. (2001). Supporting document to the new basel capital accord. *Basel Committee on Banking Supervision. Consultative Document. Bank for International Settlements.–January*.

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

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society: Series B (Methodological)*, *51*(1), 47–60.

StaffWriter. (2017). *Citibank to pay massive fine in SA currency trading collusion scandal* (pp. 1–2). Retrieved from `https://businesstech.co.za/news/banking/158931/citibank-to-pay-massive-fine-in-sa-currency-trading-collusion-scandal/`

Stasinopoulos, M. D., Rigby, R. A., & Bastiani, F. D. (2018). GAMLSS: A distributional regression approach. *Statistical Modelling*, *18*(3-4), 248–273.

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

Stasinopoulos, M., Rigby, B., & Akantziliotou, C. (2008). *Instructions on how to use the gamlss package in r second edition*.

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

Verbyla, A. P. (1993). Modelling variance heterogeneity: Residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(2), 493–508.

Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J., & Stasinopoulos, D. (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, *39*(6), 1279–1293.

Williams, G. (2011). *Data mining with rattle and r: The art of excavating data for knowledge discovery*. Springer Science & Business Media.

Wiseman, R. M., & 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.

Yan, J., Guszcza, J., Flynn, M., & Wu, C.-S. P. (2009). Applications of the offset in property-casualty predictive modeling. *Casualty actuarial society e-forum, winter 2009*, 366.

APPENDICES

EDM for the Poisson variate

$$f_Y(y_i;\lambda_i) = \exp\{y_i \log\lambda_i - \lambda_i - \ln y!\} \tag{8.1}$$

The Log-likelihood is found by taking logs and summing over i:

$$\begin{aligned} \mathrm{D}(\bar{y},\bar{\mu}) &= 2\sum_i [y_i \log\lambda_i - \lambda_i] - 2\sum_i [y_i \log\hat{\mu}_i - \hat{\mu}_i] \\ &= 2\sum_i \left[y_i \log\Big(\frac{\lambda_i}{\hat{\mu}_i}\Big) + (\hat{\mu}_i - \lambda_i) \right] \end{aligned} \tag{8.2}$$

or GLM deviance

$$\mathrm{D}_{GLM} = -2\log\frac{\hat{L}_c}{\hat{L}_s} \tag{8.3}$$

*note that deviances may be interpreted as weighted sums of distances of estimated means from observations y_i.

%

%

%

%

Annex 1: Decision trees to determine LDCE event categorisation

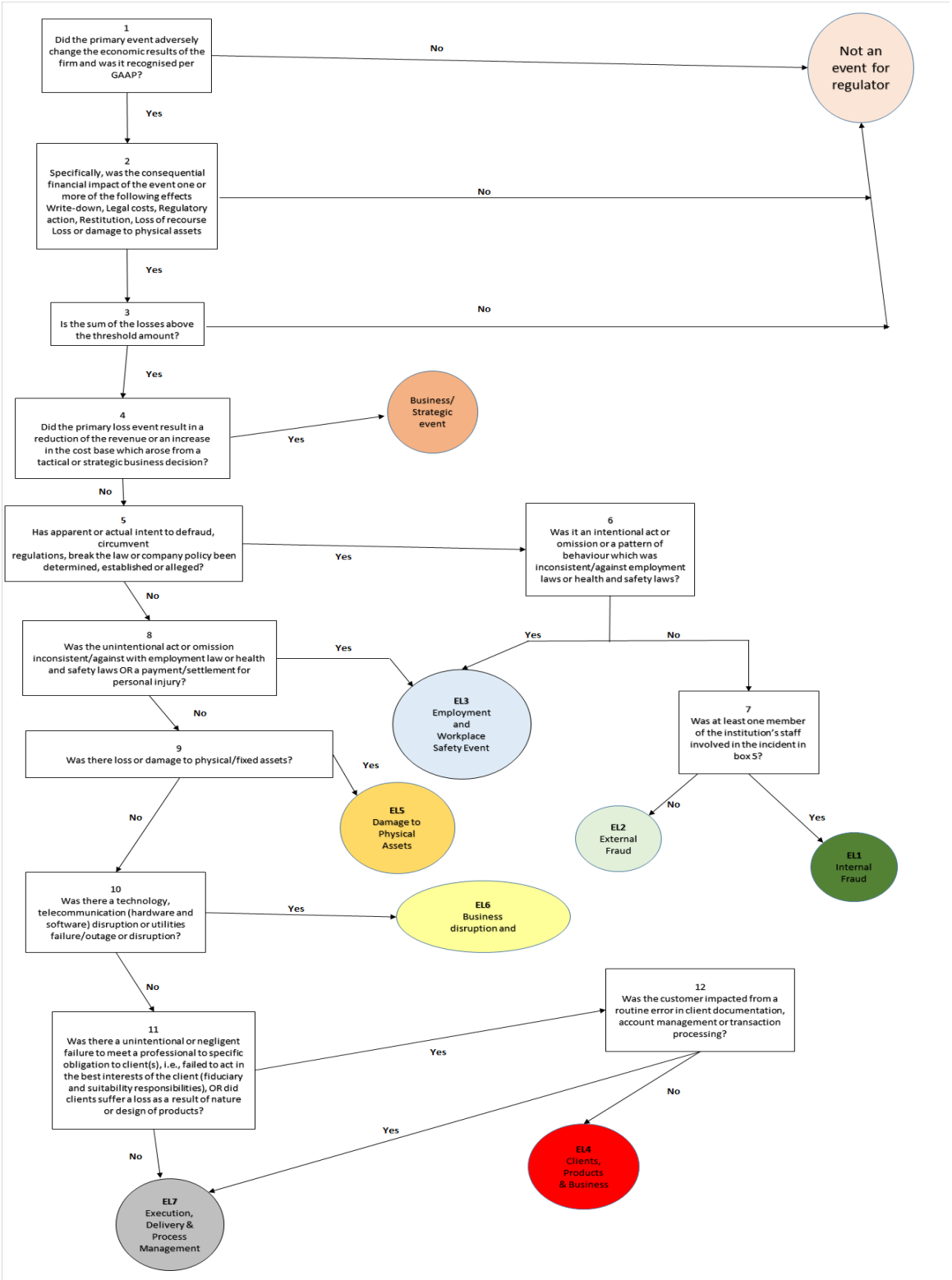


Figure 8.1
Categorisation flowchart for Loss data collection exercise (LDCE) (Committee et al., 2001)

Table 8.1
Add caption

Event type category - Level 1	Code	Event type category - Level 2	Code	Activity Examples (Level 3)	Code
Internal Fraud	EL1	Unauthorized Activity	EL11	Transactions not reported (intentional)	EL1101
				Trans type unauthorised (w/monetary loss)	EL1102
				Erroneously conducted operational activities (intentional)	EL1103
				Mismarking of position (intentional)	EL1104
				Intentional manipulation of documents	EL1105
				Misuse of vested responsibility	EL1106
		Theft and Fraud	EL12	Fraud / credit fraud / worthless deposits	EL1201
				Theft / extortion / embezzlement / robbery	EL1202
				Misappropriation of assets	EL1203
				Malicious destruction of assets	EL1204
				Forgery	EL1205
				Check kiting	EL1206
				Smuggling	EL1207
				Account take-over / impersonation / etc.	EL1208
				Tax non-compliance / evasion (wilful)	EL1209
				Bribes / kickbacks	EL1210
				Insider trading (not on firm's account)	EL1211
		Internal Security Systemuse of IT system by bank employees	EL13	Misuse of IT System	EL1301
				Manipulation of files or programmes	EL1302
				Inappropriate deployment of confidential data	EL1303
				Other types of computer crimes	EL1304
External Fraud	EL2	Theft and Fraud	EL21	Theft/Robbery	EL2101
				Forgery	EL2102
				Check kiting	EL2103
		Systems Security	EL22	Hacking damage	EL2201
				Theft of information (w/monetary loss)	EL2202
Employment Practices and Workplace Safety	EL3	Employee Relations	EL31	Compensation, benefit, termination issues	EL3101
				Organised labour activity	EL3102
		Safe Environment	EL32	General liability (slip and fall, etc.)	EL3201
				Employee health & safety rules events	EL3202
				Workers compensation	EL3203
		Diversity and Discrimination	EL33	All discrimination types	EL3301
Clients, Products and Business Practices	EL4	Suitability, Disclosure, and Fiduciary	EL41	Fiduciary breaches / guideline violations (breach of data confidentiality)	EL4101
				Suitability / disclosure issues (KYC, breach of code of conduct, etc.)	EL4102
				Retail consumer disclosure violations (Non-transparency toward client)	EL4103
				Breach of privacy	EL4104
				Unauthorised trading practices (aggressive sales)	EL4105
				Account churning (manipulations to create fictitious operations)	EL4106
				Misuse of confidential information	EL4107
				Lender Liability	EL4108
		Improper Business or Market Practices	EL42	Antitrust (violation of anti-monopoly regulations)	EL4201
				Improper trade / market practices	EL4202
				Market manipulation/unauthorised trading or market practices	EL4203
				Insider trading (on firm's account)	EL4204
				Engaging in unlicensed activity	EL4205
				Money laundering	EL4206

Table 8.2
Add caption

		Product Flaws	EL43	Product defects (unauthorised, ambiguous or punitive contract clauses, etc.)	EL4301
				Model errors	EL4302
		Selection, Sponsorship, and Exposure	EL44	Failure to investigate client per guidelines	EL4401
				Exceeding client exposure limits	EL4402
		Advisory Activity	EL45	Disputes over performance of advisory activities	EL4501
Damage to Physical Assets	EL5	Disasters and Other Events	EL51	Natural disaster losses	EL4502
				Human losses from external sources (terrorism, vandalism, industrial action, etc.)	EL4503
		Political and Legal Risk	EL52	Changes in law etc,..	EL5201
				Political changes	EL5202
Business Disruption and System Failures	EL6	Inadequacy, inefficiency, poor functioning or fall of the IT System	EL61	Inaccesibility of applications	EL6101
				Inability to send and receive data	EL6102
				Faulty automatic data processing (inaccurate and incomplete data)	EL6103
				Inaccesible or untimely systemic data offer	EL6104
				Fall of the system (application, framework..)	EL6105
		Unavailability of providers (Public services/Information)	EL62	Fall of the system of public services (Telecommunication, power supply, etc..)	EL6201
Execution, Delivery, and Process Management	EL7	Transaction Capture, Execution, and Maintenance	EL71	Miscommunication	EL7101
				Data entry, maintenance or loading error	EL7102
				Missed deadline or responsibility	EL7103
				Model / system misoperation	EL7104
				Accounting error / entity attribution error	EL7105
				Other task misperformance	EL7106
				Delivery failure	EL7107
				Collateral management failure	EL7108
				Reference Data Maintenance	EL7109
		Monitoring and Reporting	EL72	Failed mandatory reporting obligation	EL7201
				Inaccurate external report (loss incurred)	EL7202
		Customer Intake and Documentation	EL73	Client permissions / disclaimers missing	EL7301
				Legal documents missing / incomplete	EL7302
		Customer/Client Account Management	EL74	Unapproved access given to accounts	EL7401
				Incorrect client records (loss incurred)	EL7402
				Negligent loss or damage of client assets	EL7403
		Trade Counterparties	EL75	Non-client counterparty misperformance (poor work of business partners)	EL7501
				Misc. non-client counterparty disputes (disputes with business partners)	EL7502
		Vendors and Suppliers	EL76	Outsourcing	EL7601
				Vendor disputes	EL7602
Normal Activity	EL8	Not an event for regulator	ET	Not Applicable	ET ET

Table 8.3
Add caption

Business Lines - Level 1	Code1	Business Lines - Level 2	Code2	Internal Business Line	Code3
Corporate Finance	BL1	Corporate Finance	BL11	Corporate Bonds	BL1101
		Municipal/Government Finance	BL12		
		Merchant Banking	BL13		
		Advisory Services	BL14		
Trading and Sales	BL2	Sales	BL21	Prime Broking	BL2101
				Prime Services	BL2102
		Market Making	BL22	Absa Capital Securities	BL2201
				Equity	BL2202
				Flow Credit - MTM	BL2203
				Index	BL2204
				MM	BL2205
				MM Own Paper Trading	BL2206
				Non-ZAR Fair Value	BL2207
				Precious metals and Index ETN	BL2208
		Proprietary Positions	BL23	ACS Prime Risk	BL2301
				Africa BTB Trading	BL2302
				Agri_Softs Trading	BL2303
				Bond Trading	BL2304
				Cash Equities	BL2305
				CD - Offshore vs Domestic_20130312_127226.xls	BL2306
				Commodities	BL2307
				Complex Equity Risk FV	BL2308
				Counterparty Credit	BL2309
				Credit Derivatives	BL2310
				Delta One Trading	BL2311
				DH FX-Risk	BL2312
				Energy_Base Metals	BL2313
				Equity	BL2314
				Equity	BL2315
				ETF - Trading Accrued	BL2316
				FI - Options_CRE	BL2317
				Forex	BL2318
				FX Forwards Yieldx	BL2319
				Index	BL2320
				Inflation	BL2321
				Interest Rate - MTM	BL2322
				Linear Derivatives	BL2323
				Non-Linear Derivatives	BL2324
				Non-Linear Stock Derivatives	BL2325
				Options	BL2326
				Quanto Risk	BL2327
				Rates	BL2328
				SBL - Delta One Products	BL2329
				SND	BL2330
				SSA Structured - MTM	BL2331
				Structured Credit Trading - MTM	BL2332
				Structured Trading	BL2333
				Swap Flow	BL2334
		Treasury	BL24	Bonds/Repos	BL2401
				Group Treasury	BL2402
				Repos MTM	BL2403

Table 8.4
Add caption

Retail Banking	BL3	Retail Banking	BL31	Structured Accounting Mismatches	BL3101
				Structured Accounting Mismatches Accruals	BL3102
				Structured Accounting Mismatches Fair Value	BL3103
		Private Banking	BL32	ABCAP FUND	BL3201
				ABCAP Funding Mismatch	BL3202
				Africa	BL3203
				Corp Desk Deposits	BL3204
		Card Services	BL33		BL3301
Commercial Banking	BL4	Commercial Banking	BL41	ACBB	BL4101
				Corporate Short Term Lending	BL4102
				Funding	BL4103
				Funding - Fair Value	BL4104
				Funding - Funding Pool	BL4105
				GROUP FUND 2476	BL4106
				ILLIQUID BOOK	BL4107
				LIQUID BOOK	BL4108
				Strategic loan Portfolio(SA) - Accrual	BL4109
				Structured Credit - Secured Finance - MTM	BL4110
				Structured Credit - Unsecured Finance - MTM	BL4111
Payment and Settlement	BL5	External Clients	BL51	EXTERNALCLIENT - EC-IR	BL5101
				LEGACY - ECL-IR	BL5102
				Liabilities 2474	BL5103
Agency Services	BL6	Custody	BL61	Prime Services	BL6101
				Securities Lending	BL6102
		Corporate Agency	BL62	Securities - Lending (Agency)	BL6201
		Corporate Trust	BL63	Management	BL6301
				Regional Management	BL6302
Asset Management	BL7	Discretionary Fund Management	BL71	Bond Investment (incl Legacy) MTM	BL7101
		Non-Discretionary Fund Management	BL72	Investment Banking	BL7201
				Portfolio Management	BL7202
Retail Brokerage	BL8	Retail Brokerage	BL81		BL8101
Corporate Level – Non-Business Line Specific	BL9	Losses originating from a corporate/firm-wide function that cannot be linked to a specific business line.	BL91	Other	BL9100

Appendix A: R Code and Data Preparation for Chapter 3

Required: R Packages from CRAN

Data preparation for understanding the raw frequency and raw severity data mentioned in section 3.9 on page 83 of collected internally over the period between 1 January 2013 and 31 March 2013 i.e., Q12013 at an investment bank in SA

```
file_loc <- "C:/Users/Mphekeleli/Documents/R PROJECT/OpRiskPHDGitHub
/OpRisk_PHD_Thesis/Data"
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")
```

Data preparation providing the numbers of OpRisk events collected through pre-processing by following the LCDE as described in section 3.11 on 91 over Q12013, limited to the covariates selected to fit into the models

```
# Set parameter values
crv$seed <- 42 # set random seed
crv$training.proportion <- 1.0

# Load data
fname <- "file:///G:/PHD/OPRISK_PHD_DISS/Data/OPriskDataSet_exposure.csv"
crs$dataset <- read.csv(fname,
              sep=";",
              dec=",",
              na.strings=c(".", "NA", "", "?"),
              strip.white=TRUE, encoding="UTF-8")

# Build the train/validate/test datasets.

# nobs=2330 train=2330 validate=0 test=0
set.seed(crv$seed)
crs$nobs   <- nrow(crs$dataset)
crs$sample <- sample(crs$nobs, crv$training.proportion * crs$nobs)
```

```
crs$train  <- crs$sample <- sample(crs$nobs, crv$training.proportion * crs$nobs)
# crs$train  <- sample(crs$nobs, 1*crs$nobs)
crs$validate <- NULL
crs$test <- NULL

# The following variable selections have been noted.
crs$input     <- c("Trade", "UpdatedDay", "UpdatedTime",
                   "TradedDay", "TradedTime", "Desk", "CapturedBy",
                   "TradeStatus", "TraderId", "Instrument", "Reason",
                   "Loss", "EventTypeCategoryLevel1",
                   "BusinessLineLevel1", "exposure")

crs$numeric   <- c("Trade", "UpdatedDay", "UpdatedTime",
                   "TradedDay", "TradedTime", "Loss", "exposure")

crs$categoric <- c("Desk", "CapturedBy", "TradeStatus",
                   "TraderId", "Instrument", "Reason",
                   "EventTypeCategoryLevel1", "BusinessLineLevel1")

crs$target    <- "LossIndicator"
crs$risk      <- NULL
crs$ident     <- NULL
crs$ignore    <- c("UpdateTime", "TradeTime", "Nominal", "FloatRef",
                   "LastResetDate", "LastResetRate", "Theta", "Unexplained")
crs$weights   <- NULL
```

The algorithm only accepts numerical data and so categorical data is transformed into numeric. This is done using an approach where each value of a categoric variable is turned into a variable itself. Multi-level categoric variables are recoded by building dummy variables corresponding to each level by the following commands:

Numerical and categoric variables (to be transformed by below conversion code) depicted in table 3.3 on page 86:

```
# 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"]])
```

Table generated by conversion code wherein 'Dummy' multi-level categoric variables are created by recoded by remapping variables and transforming into a numeric

corresponding to each level Rattle

...Continued from previous

Scatterplots and Histograms plots of loss severities and frequency counts against selected explanatory variables showing basic summary statistics of intra-day trading activity

Figures 3.2a and 3.2b on page 92:

```
# Scatterplots for loss severities
plot(projdata$UpdatedTime, log(projdata$Loss+0.000000001), ylim = c(6, 18),
     col = "navy", xlab = "Updated Time", ylab = "Log. Loss")
xyplot(Loss ~ as.factor(TraderId) , data = projdata)

# Histograms for loss severities
hist(projdata$UpdatedDay, col = "#9999CC", main = "All losses", xlab = "Updated Day"
     , ylab = "Frequency")
hist(projdata$TradedDay)
```

Characteristics of exposure: Exposure data is used for several of the steps of the process in frequency and severity modelling

Figure 3.1 on page 90:

```
#######Display histogram plots for the selected variables##########
#============================================================
# Display histogram plots for the selected variables. Use ggplot2 to generate
# histogram plot for exposure. Generate the plot.

Exp <- crs %>%
  with(dataset[train,]) %>%
  dplyr::select(exposure) %>%
  ggplot2::ggplot(ggplot2::aes(x=exposure)) +
  ggplot2::geom_density(lty=1, lwd=1) +
  ggplot2::geom_density(ggplot2::aes(fill="", colour=""), alpha=0.55) +
  ggplot2::xlab("Exposure") +
  ggplot2::ggtitle("Pdf(exposure)") +
  ggplot2::labs(y="Density") +
  theme_bw(base_size = 15) +
  theme(legend.position = "none")

# Display the plots.
gridExtra::grid.arrange(Exp)

#============================================================
```

Table 8.5
Building 'Dummy' variables corresponding to each level

LossIndicator

n	missing	distinct	Info	Sum	Mean	Gmd
1631	0	2	0,4	262	0,16	0,3

exposure

n	missing	distinct	Info	Mean	Gmd	0,1	0,1	0,3	0,5	0,8	0,9	0,95
1631	0	1592	1	389	535	1,1	2	8	109	622	1247	1785

Desk

n	missing	distinct	Info	Mean	Gmd	0,1	0,1	0,3	0,5	0,8	0,9	0,95
1631	0	10	1	6,2	3,14	2	3	4	7	9	10	10
lowest : 1 2 3 4 5	highest:	6	7	8	9	10						

Value	1	2	3	4	5	6	7	8	9	10
Frequency	38	110	234	157	232	26	157	196	274	207
Proportion	0,023	0,067	0,1	0,1	0,14	0	0,1	0,1	0,2	0,1

CapturedBy

n	missing	distinct	Info	Mean	Gmd
1631	0	5	0,8	3,45	1
lowest : 1 2 3 4 5	highest:	1	2	3	4	5

Value	1	2	3	4	5
Frequency	168	64	365	943	91
Proportion	0,103	0,039	0,2	0,58	0,06

TradeStatus

n	missing	distinct	Info	Mean	Gmd
1631	0	4	0,8	1,62	0,6

Value	1	2	3	4
Frequency	708	832	87	4
Proportion	0,434	0,51	0,1	0

TraderId

n	missing	distinct	Info	Mean	Gmd
1631	0	7	1	3,91	2,05
lowest : 1 2 3 4 5	highest:	3	4	5	6	7

Value	1	2	3	4	5	6	7
Frequency	107	369	367	12	457	136	183
Proportion	0,066	0,226	0,2	0,01	0,28	0,1	0,1

Table 8.6
Building 'Dummy' variables corresponding to each level

Instrument

n	missing	distinct	Info	Mean	Gmd	0,1	0	0,3	0,5	1	1	1
1631	0	23	1	15,6	6,96	4	5	12	17	22	22	23
lowest : 1 2 3 4 5	highest:	19	20	21	22	23						

Reason

n	missing	distinct	Info	Mean	Gmd	0,1	0	0,3	0,5	1	1	1
1631	0	19	1	10,1	5,92	3	3	4	10	16	16	18
lowest : 1 2 3 4 5	highest:	15	16	17	18	19						

Value	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	3	5	219	247	14	38	35	125	111	21	10	274
Proportion	0,002	0,003	0,1	0,15	0,01	0	0	0,1	0,1	0	0	0,2

Value	13	14	15	16	17	18	19
Frequency	15	8	47	360	8	60	31
Proportion	0,009	0,005	0	0,22	0,01	0	0

EventTypeCategoryLevel1

n	missing	distinct	Info	Mean	Gmd							
1631	0	5	0,8	3,56	1,1							
lowest : 1 2 3 4 5	highest:	1	2	3	4	5						

Value	1	2	3	4	5
Frequency	109	252	102	955	213
Proportion	0,067	0,155	0,1	0,59	0,13

BusinessLineLevel1

n	missing	distinct	Info	Mean	Gmd							
1631	0	8	0,8	2,91	1,37							
lowest : 1 2 3 4 5	highest:	4	5	6	7	8						

Value	1	2	3	4	5	6	7	8
Frequency	17	1011	191	116	121	166	6	3
Proportion	0,01	0,62	0,1	0,07	0,07	0,1	0	0

```
# Generate just the data for an Ecdf plot of the variable 'exposure'.
ds <- rbind(data.frame(dat=crs$dataset[crs$train,][,"exposure"], grp="All"))

# The 'Hmisc' package provides the 'Ecdf' function.
library(Hmisc, quietly=TRUE)

# Plot the data.
Ecdf(ds[ds$grp=="All",1], col="red", xlab="exposure", lwd=2,
     ylab=expression(Proportion <= x), subtitles=FALSE)

# Add a title to the plot.
title(main="Cdf(exposure)")

#=============================================================
# Benford's Law

# The 'ggplot2' package provides the 'ggplot' function.
library(ggplot2, quietly=TRUE)

# The 'reshape' package provides the 'melt' function.
library(reshape, quietly=TRUE)

# Initialies the parameters.
var    <- "exposure"
digit  <- 1
len    <- 1

# Build the dataset
ds <- merge(benfordDistr(digit, len),
            digitDistr(crs$dataset[crs$train,][var], digit, len, "All"))

# Plot the digital distribution
p <- plotDigitFreq(ds)
p <- p + ggtitle("Digital Analysis") +
  theme(
    legend.position = c(.95, .95),
    legend.justification = c("right", "top"),
    legend.box.just = "right",
    legend.margin = margin(6, 6, 6, 6)
    ) +
  # custom box around legend
 theme(
    legend.box.background = element_rect(color="red", size=2),
    legend.box.margin = margin(6, 6, 6, 6)
) +
# custom the key
 theme(legend.key = element_rect(fill = "white", colour = "black")) +
```

```
# custom the text
 theme(legend.text = element_text(size = 8, colour = "red")) +
# custom the title
 theme(legend.title = element_text(face = "bold"))
print(p)
# #============================================================
```

Histograms for frequency characteristics of daily and monthly operations for daily losses and/or pending/near misses

Figures 3.3a and 3.3b on page 95:

```
################   EXPLORATORY DATA ANALYSIS  ###################
#_______________________________________________________________________________________
# Update Time
### summary statistics
 summary(projdata$UpdatedTime)

### Histograms - ALL Losses and Near Misses/Pending Losses and Realised losses
par(mfrow=c(1,3))
hist(projdata$UpdatedTime, col = "blue", main = "All losses",
     xlab = "Update Time", ylab = "Frequency")
hist(projdata$UpdatedTime[projdata$LossIndicator == 0], col = "red",
     main = "Near Misses", xlab = "Update Time", ylab = "Frequency")
hist(projdata$UpdatedTime[projdata$LossIndicator == 1], col = "green",
     main = "Realised losses", xlab = "Update Time", ylab = "Frequency")
par(mfrow=c(1,1))
#_______________________________________________________________________________________
# # Update Day
 summary(projdata$UpdatedDay)

### Histograms - ALL Losses and Near Misses/Pending Losses and Realised losses
par(mfrow=c(1,3))
hist(projdata$UpdatedDay, col = "#9999CC", main = "All losses",
     xlab = "Updated Day", ylab = "Frequency")
hist(projdata$UpdatedDay[projdata$LossIndicator == 0], col = "#CC6666",
     main = "Near Misses", xlab = "Updated Day", ylab = "Frequency")
hist(projdata$UpdatedDay[projdata$LossIndicator == 1], col = "#66CC99",
     main = "Realised losses", xlab = "Updated Day", ylab = "Frequency")
par(mfrow=c(1,1))
```

Density plots of overlaid trade proportions of realised losses vs pending losses/near misses

Figures 3.4 on page 96:

```
# Density Plot for Updated Day
p01 <- crs %>%
  with(dataset[sample,]) %>%
  dplyr::mutate(LossIndicator=as.factor(LossIndicator)) %>%
  dplyr::select(UpdatedDay, LossIndicator) %>%
  ggplot2::ggplot(ggplot2::aes(x=UpdatedDay)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::geom_density(ggplot2::aes(fill=LossIndicator, colour=LossIndicator)
                        , alpha=0.55) +
  ggplot2::ggtitle("Distr. of Updated DaY") +
  ggplot2::labs(fill="LossIndicator", y="Density") +
  ggplot2::xlab("Day of Month that Trade Was Updated") +
  ggplot2::theme(legend.position=c(.7,.2))

# Density Plot for Traded day
p02 <- crs %>%
  with(dataset[sample,]) %>%
  dplyr::mutate(LossIndicator=as.factor(LossIndicator)) %>%
  dplyr::select(TradedDay, LossIndicator) %>%
  ggplot2::ggplot(ggplot2::aes(x=TradedDay)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::geom_density(ggplot2::aes(fill=LossIndicator, colour=LossIndicator)
                        , alpha=0.55) +
  ggplot2::ggtitle("Distr. of TradedDay") +
  ggplot2::labs(fill="LossIndicator", y="Density") +
  ggplot2::xlab("Day of Month for Trade") +
  ggplot2::theme(legend.position=c(.7,.2))

# Display the plots.
gridExtra::grid.arrange(p01, p02, nrow = 1)
```

Table 3.5 on 94

```
# addmargins(table(projdata$Desk, projdata$LossIndicator), 2)
```

Histograms of overlaid trade proportions of realised losses vs pending loss/near misses

Figures 3.5 on page 97:

```
# Plot Desk category distribution
p03 <- crs %>%
  with(dataset[sample,]) %>%
  dplyr::mutate(LossIndicator=as.factor(LossIndicator)) %>%
  dplyr::select(Desk, LossIndicator) %>%
  dplyr::group_by(Desk, LossIndicator) %>%
  dplyr::summarise(n = n()) %>%
  ggplot2::ggplot(ggplot2::aes(x=Desk, y=n, fill=LossIndicator)) +
  ggplot2::geom_bar(stat="identity") +
```

```
  ggplot2::ggtitle("Desk category distribution") +
  ggplot2::theme_minimal() +
  ggplot2::theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  ggplot2::ylab("Frequency")

# Create new variable to proportion no. of realised losses
T01 <- crs %>%
  with(dataset[sample,]) %>%
  dplyr::mutate(LossIndicator=as.factor(LossIndicator)) %>%
  dplyr::select(Desk, LossIndicator) %>%
  dplyr::group_by(Desk, LossIndicator) %>%
  dplyr::summarise(n = n())

T02 <- T01 %>%
  group_by(Desk) %>%
  summarise(N=sum(n))

T03 <- inner_join(T01, T02)

# Plot Desk category by proportion
T04 <- T03 %>%
  mutate(Prob=n/N) %>%
  filter(LossIndicator==1) %>%
  select(Desk, Prob) %>%
  arrange(desc(Prob)) %>%
  ggplot2::ggplot(ggplot2::aes(x=Desk, y=Prob, fill=Desk), alpha=0.55) +
  ggplot2::geom_bar(stat="identity", fill="grey", colour="black", show.legend = FALSE)+
  ggplot2::ggtitle("Proportion of losses per Desk") +
  ggplot2::theme_minimal()+
  ggplot2::theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  ggplot2::ylab("Loss Ratio (n/N)")+
  ggplot2::xlab("Desk")

#Display both plots in one row
gridExtra::grid.arrange(p03, T04, nrow = 1)
####============================================================
```

Mosaic grid plots for the bidimensional distributions by traded ***Instrument****,* ***TraderId*** *and number*

Figures 3.6 on page 98:

```
par(mfrow=c(1,2))
plot(table(projdata$LossIndicator, projdata$Instrument), main="By Instrument",
     col=rainbow(20), las=1)
plot(table(projdata$LossIndicator, projdata$TraderId), main="By Trader",
     col=rainbow(20), las=1)
```

```
plot(table(projdata$LossIndicator, projdata$CapturedBy), main="By Tech Support",
     col=rainbow(20), las=1)
par(mfrow = c(1, 1))
####============================================================
```

Mosaic grid plots for the structure of OpRisk portfolio by traded ***TradeStatus****,* ***TraderId*** *by the number of realised losses vs pending loss/near misses*

Figures 3.6 on page 98:

```
# Contingency Table
library(vcd)
STD <- structable(~TradeStatus + TraderId + LossIndicator, data = projdata)
par(cex.axis = 1.5, cex.lab = 0.1)
# Mosaic plot
MS01 <- mosaic(STD, condvars = 'TradeStatus', col=rainbow(20),
      split_horizontal = c(TRUE, FALSE, TRUE), shade = TRUE, legend = TRUE)
MS01
```

Appendix B: R Code for Chapter 4

Required: R Packages from CRAN (in addition to packages already found in Chapter 3)

Data partitioning of the pre-processed OpRisk dataset into Training/Validation/Testing proportions, in preparation for machine learning model building treatments. The original dataset is partitioned into three random subsets initiated by a random number sequence with a randomly selected seed.

The function `getmode` specifies the modal class as the reference level in the GLM from which the corresponding observations are estmated and weighted against, see chapter 3 section 3.9 on page 83

```
building <- TRUE
scoring <- ! building

# Load data
fname <- "file:///C:/Users/Mphekeleli/Documents/R PROJECT/OpRiskPHDGitHub
/OpRisk_PHD_Thesis/Data/OPriskDataSet_exposure.csv"
crs$dataset <- read.csv(fname,
              sep=";",
              dec=",",
              na.strings=c(".", "NA", "", "?"),
              strip.white=TRUE, encoding="UTF-8")
exposure <- crs$dataset[,ncol(crs$dataset)]

# Select variables for loss incident model
crs$dataset <- as.data.frame(crs$dataset)

# The following varaible selections have been noted

crs$input <- crs$dataset %>%
  group_by(UpdatedDay,
           UpdatedTime,
           TradedDay,
           TradedTime,
           Desk,
           CapturedBy,
           TradeStatus,
           TraderId,
           Instrument,
           Reason,
           EventTypeCategoryLevel1,
           BusinessLineLevel1) %>%
```

```
  transmute(LossesIndicator = LossIndicator,
            Losses = Loss,
            Exposure = exposure)

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

for (i in 5:(ncol(crs$input) - 3)){
     crs$input[[i]] <- relevel(crs$input[[i]], getmode(crs$input[[i]]))
}

#============================================================================================
# A predefined value is used to reset the random seed so that results are repeatable
crv$seed <- 42

# Build the training/validation/testing datasets. Set parameter values


set.seed(crv$seed)    # set random seed to make your partition reproducible

crs$nobs <- nrow(crs$input)                 # nobs=2331

crs$train <- sample(crs$nobs, 0.7*crs$nobs) # proportion of training data = 1632

crs$nobs %>%
  seq_len() %>%
  setdiff(crs$train) %>%
  sample(0.15*crs$nobs) ->                  # proportion of validation data = 350
  crs$validate

crs$nobs %>%
  seq_len() %>%
  setdiff(crs$train) %>%
  setdiff(crs$validate) ->                  # proportion of testing data = 349
  crs$test


crs$training <- as.data.frame(crs$input[crs$train,])
crs$validation <- as.data.frame(crs$input[crs$validate,])
crs$testing <- as.data.frame(crs$input[crs$test,])
```

Models

Estimation of some poisson regression models for OpRisk loss frequency distribution: To build the model we pass on to the model building function `glm` *i.e., the formula that describes the model to build.*

use "LossesIndicator" as the dependent variable, `TradedDay` and `Desk` variables are predictor variables.

Basic model build summary

```
Call:
glm(formula = LossesIndicator ~ TradedDay + Desk, family = poisson(link = "log"),
    data = crs$training, offset = log(Exposure))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8706  -0.5300  -0.2286  -0.0545   4.3750

Coefficients:
                      Estimate Std. Error z value             Pr(>|z|)
(Intercept)          -8.053221   0.193632 -41.590 < 0.0000000000000002 ***
TradedDay            -0.014087   0.006381  -2.208              0.027266 *
DeskAfrica            1.457695   0.413087   3.529              0.000417 ***
DeskBonds/Repos       1.764230   0.254571   6.930      0.0000000000042 ***
DeskCommodities       0.924033   0.235749   3.920      0.0000887114575 ***
DeskDerivatives      -0.577626   0.344672  -1.676              0.093763 .
DeskEquity            1.365152   0.225948   6.042      0.0000000015232 ***
DeskManagement/Other -1.410706   1.014052  -1.391              0.164177
DeskMM                0.339561   0.241501   1.406              0.159711
DeskPrime Services    2.129594   0.223030   9.548 < 0.0000000000000002 ***
DeskSND              -0.716361   0.283796  -2.524              0.011596 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1994.9  on 1630  degrees of freedom
Residual deviance: 1764.7  on 1620  degrees of freedom
AIC: 2310.7

Number of Fisher Scoring iterations: 8
```

Let us now fit a broader GLM to be our global model, still ***LossesIndicator*** *as the dependent variable, while the rest of the predictive variables will be predictor variables.*

```
freqfit <- glm(formula = LossesIndicator ~ UpdatedDay + UpdatedTime + TradedDay +
    TradedTime + Desk + CapturedBy + TradeStatus + TraderId +
    Instrument + Reason + EventTypeCategoryLevel1 + BusinessLineLevel1,
    family = poisson(link = "log"), data = crs$training,
    offset = log(Exposure)))
```

Global model build summary

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7605  -0.3575  -0.1105  -0.0315   3.9754

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7605  -0.3575  -0.1105  -0.0315   3.9754

Coefficients:
                       Estimate     Std. Error z value             Pr(>|z|)
(Intercept)          -7.8113982    0.6701090 -11.657 < 0.0000000000000002 ***
UpdatedDay            0.0043298    0.0097374   0.445             0.656565
UpdatedTime           0.5850748    0.7146037   0.819             0.412935
TradedDay            -0.0001216    0.0082899  -0.015             0.988297
TradedTime           -0.8201083    0.7856849  -1.044             0.296572
DeskAfrica            1.5831846    0.5281939   2.997             0.002723 **
DeskBonds/Repos       2.3432611    0.3958404   5.920     0.00000000322505 ***
DeskCommodities       0.6910019    0.4627163   1.493             0.135343
DeskDerivatives       0.3832286    0.5020354   0.763             0.445255
DeskEquity            1.1778919    0.3969116   2.968             0.003001 **
DeskManagement/Other -14.7535497  367.5957685  -0.040             0.967985
DeskMM                1.2027479    0.5903580   2.037             0.041618 *
DeskPrime Services   -0.1311216    1.3579826  -0.097             0.923079
DeskSND               0.0518179    0.7105464   0.073             0.941864
CapturedByMIDOFFICE   0.2618294    0.3319433   0.789             0.430242

Some omissions
```

```
BusinessLineLevel
1BL5                -1.1905846    0.5243289  -2.271             0.023166 *
BusinessLineLevel
1BL6                 1.7544025    1.4007670   1.252             0.210403
BusinessLineLevel
1BL7                 2.2553305    2.2524125   1.001             0.316684
BusinessLineLevel
1BL9                 0.2496745 1613.4575790   0.000             0.999877
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1898.7  on 1630  degrees of freedom
Residual deviance: 1228.1  on 1553  degrees of freedom
AIC: 1888.1

Number of Fisher Scoring iterations: 15
```

Estimation of some binomial regression modelS: The formula that describes the model to build.

The target variable is “remapped” or transformed into a factor variable from a numerical variable by use of the code below:

```
crs$dataset[["TFC_LossIndicator"]] <-
  as.factor(crs$dataset[["LossIndicator"]])

ol <- levels(crs$dataset[["TFC_LossIndicator"]])
lol <- length(ol)
nl <- c(sprintf("[%s,%s]", ol[1], ol[1]),
        sprintf("(%s,%s]", ol[-lol], ol[-1]))
levels(crs$dataset[["TFC_LossIndicator"]]) <- nl
```

We will use ”LossesIndicator” as the dependent variable.

```
 freqfit <- glm(LossesIndicator ~ UpdatedDay + UpdatedTime + TradedDay
        + TradedTime + Desk + CapturedBy + TradeStatus + TraderId +
   Instrument + Reason + EventTypeCategoryLevel1 + BusinessLineLevel1,
   data=crs$training, family=binomial(link="logit"), offset=log(Exposure))

 summary(freqfit)
```

Basic model build summary

```
Call:
```

```
glm(formula = LossesIndicator ~ UpdatedDay + UpdatedTime + TradedDay +
    TradedTime + Desk + CapturedBy + TradeStatus + TraderId +
    Instrument + Reason + EventTypeCategoryLevel1 + BusinessLineLevel1,
    family = binomial(link = "logit"), data = crs$training, offset = log(Exposure)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2957  -0.4346  -0.0987  -0.0001   4.0153

Coefficients:
                       Estimate      Std. Error z value           Pr(>|z|)
(Intercept)          -7.9749408     1.0332350  -7.718 0.0000000000000118 ***
UpdatedDay            0.0158775     0.0144454   1.099           0.271707
UpdatedTime          -0.5014732     0.9397316  -0.534           0.593594
TradedDay            -0.0007336     0.0120301  -0.061           0.951373
TradedTime            0.9590326     1.0673571   0.899           0.368913
DeskAfrica            4.3345648     0.9075486   4.776 0.0000017870595853 ***
DeskBonds/Repos       5.0241306     0.6499654   7.730 0.0000000000000108 ***
DeskCommodities       2.0972543     0.6780211   3.093           0.001980 **
DeskDerivatives      -0.1864882     0.7413218  -0.252           0.801380
DeskEquity            1.8915605     0.5960148   3.174           0.001505 **
DeskManagement/Other -18.2048936  1520.1898190  -0.012           0.990445
DeskMM                1.9981588     0.7940263   2.516           0.011853 *
DeskPrime Services    0.5307160     1.3075756   0.406           0.684832
DeskSND               2.3274858     1.0644292   2.187           0.028771 *
CapturedByMIDOFFICE   0.8804690     0.4547261   1.936           0.052836 .

Some omissions

BL6                     0.8510447     1.4106248   0.603           0.546302
BusinessLineLevel1
BL7                   -16.5197686  2691.6096464  -0.006           0.995103
BusinessLineLevel1
BL9                    24.2348398  1520.1905636   0.016           0.987281
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2380.8  on 1630  degrees of freedom
Residual deviance: 1270.6  on 1553  degrees of freedom
AIC: 1426.6

Number of Fisher Scoring iterations: 18
```

Model selection and multimodel inference: MuMIn

"All possible models" are considered by subjectively ad iteratively searching the data for patterns and "significance".

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.

```
cl <- makeCluster(2) # Assign R cores to the job
options(na.action=na.fail)
freqfits <- dredge(freqfit)
stopCluster(cl)
freqfits
```

"MuMLn::dredge" returns a list of 4097 models, below is the build summary

We 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.*

```
cl <- makeCluster(2) # Assign R cores to the job
Amodel <- model.avg(get.models(freqfits, subset = TRUE))
summary(Amodel)
stopCluster(cl)
```

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

Evaluate model performance on the test dataset

Obtain the response from the Linear model.

```
Av.PredTT <- predict(Amodel, crs$testing, type = "response")

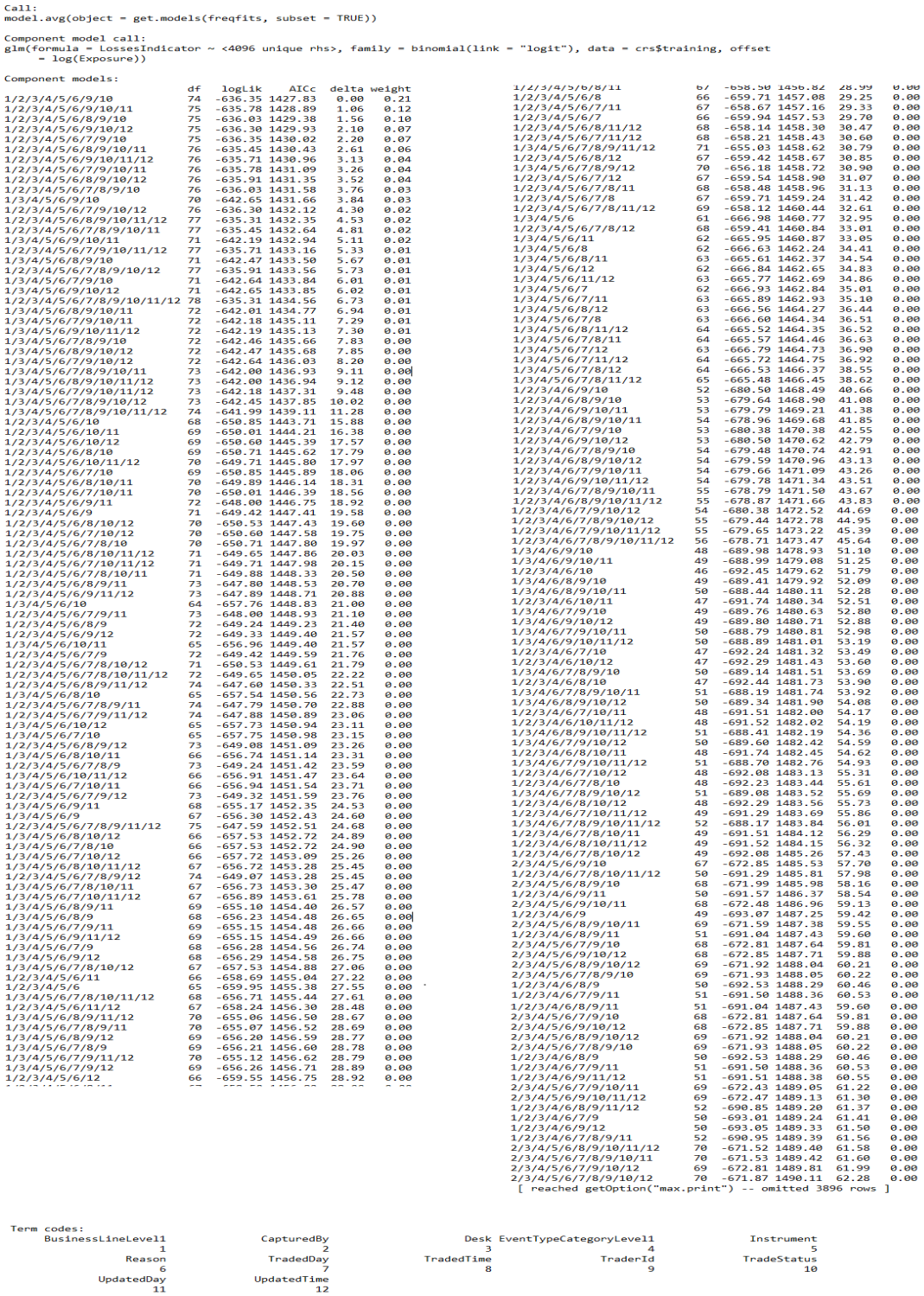
# Export into excel

HTMLStart(); HTML(data.frame(Av.PredTT)); w <- HTMLStop()
browseURL(w)

shell(paste("start excel", w))
Est <- "file:///C:/Users/Mphekeleli/Documents/R PROJECT/OpRiskPHDGitHub/
OpRisk_PHD_Thesis/Data/OPriskDataSet_GoF_Amodel_InCode.csv"
pred <- read.csv(Est,
```

```
Global model call: glm(formula = LossesIndicator ~ UpdatedDay + UpdatedTime + TradedDay +
    TradedTime + Desk + CapturedBy + TradeStatus + TraderId +
    Instrument + Reason + EventTypeCategoryLevel1 + BusinessLineLevel1,
    family = binomial(link = "logit"), data = crs$training, offset = log(Exposure))
---
Model selection table
     (Intrc) BsLL1 CptrB Desk ETCL1 Instr Reasn        TrddD     TrddT TrdrI TrdSt       UpdtD     UpdtT df    logLik   AICc delta weight
832   -7.504     +     +    +     +     +     +                                +     +                    74 -636.347 1427.8  0.00  0.210
1856  -7.754     +     +    +     +     +     +                                +     +  0.01535000          75 -635.780 1428.9  1.06  0.123
960   -7.928     +     +    +     +     +     +              0.830900          +     +                    75 -636.026 1429.4  1.56  0.096
2880  -7.349     +     +    +     +     +     +                                +     +           -0.290900 75 -636.297 1429.9  2.10  0.073
896   -7.512     +     +    +     +     +     +  0.000552900                   +     +                    75 -636.346 1430.0  2.20  0.070
1984  -8.187     +     +    +     +     +     +              0.843000          +     +  0.01546000          76 -635.452 1430.4  2.61  0.057
3904  -7.579     +     +    +     +     +     +                                +     +  0.01559000 -0.335600 76 -635.714 1431.0  3.13  0.044
1920  -7.753     +     +    +     +     +     + -0.000085550                   +     +  0.01536000          76 -635.780 1431.1  3.26  0.041
3008  -7.740     +     +    +     +     +     +              0.932100          +     +           -0.451000 76 -635.910 1431.4  3.52  0.036
1024  -7.928     +     +    +     +     +     +  0.000052290 0.830700          +     +                    76 -636.026 1431.6  3.76  0.032
830   -7.394     +           +     +     +     +                                +     +                    70 -642.646 1431.7  3.84  0.031
2944  -7.357     +     +    +     +     +     +  0.000517300                   +     +           -0.290600 76 -636.296 1432.1  4.30  0.024
4032  -7.984     +     +    +     +     +     +              0.955900          +     +  0.01583000 -0.500200 77 -635.310 1432.4  4.53  0.022
2048  -8.180     +     +    +     +     +     + -0.000582600 0.845400          +     +  0.01550000          77 -635.450 1432.6  4.81  0.019
1854  -7.623     +           +     +     +     +                                +     +  0.01348000          71 -642.193 1432.9  5.11  0.016
3968  -7.577     +     +    +     +     +     + -0.000153300                   +     +  0.01560000 -0.335800 77 -635.714 1433.2  5.33  0.015
958   -7.708     +           +     +     +     +              0.609800          +     +                    71 -642.470 1433.5  5.67  0.012
3072  -7.739     +     +    +     +     +     + -0.000050860 0.932300          +     +           -0.451100 77 -635.910 1433.6  5.73  0.012
894   -7.374     +           +     +     +     + -0.001349000                   +     +                    71 -642.639 1433.8  6.01  0.010
2878  -7.403     +           +     +     +     +                                +     +            0.017400 71 -642.646 1433.8  6.02  0.010
4096  -7.975     +     +    +     +     +     + -0.000733600 0.959000          +     +  0.01588000 -0.501500 78 -635.308 1434.6  6.73  0.007
1982  -7.944     +           +     +     +     +              0.619100          +     +  0.01355000          72 -642.012 1434.8  6.94  0.007
1918  -7.601     +           +     +     +     + -0.001573000                   +     +  0.01352000          72 -642.184 1435.1  7.29  0.005
3902  -7.610     +           +     +     +     +                                +     +  0.01350000 -0.024400 72 -642.192 1435.1  7.30  0.005
1022  -7.687     +           +     +     +     + -0.001837000 0.620300          +     +                    72 -642.458 1435.7  7.83  0.004
3006  -7.673     +           +     +     +     +              0.627300          +     +           -0.082280 72 -642.466 1435.7  7.85  0.004
2942  -7.383     +           +     +     +     + -0.001347000                   +     +            0.016580 72 -642.639 1436.0  8.20  0.003
2046  -7.921     +           +     +     +     + -0.002056000 0.630600          +     +  0.01359000          73 -641.997 1436.9  9.11  0.002
4030  -7.891     +           +     +     +     +              0.646000          +     +  0.01365000 -0.126800 73 -642.003 1436.9  9.12  0.002
3966  -7.587     +           +     +     +     + -0.001577000                   +     +  0.01354000 -0.025820 73 -642.183 1437.3  9.48  0.002
3070  -7.650     +           +     +     +     + -0.001860000 0.638500          +     +           -0.085210 73 -642.454 1437.8 10.02  0.001
4094  -7.866     +           +     +     +     + -0.002097000 0.658500          +     +  0.01369000 -0.130600 74 -641.987 1439.1 11.28  0.001
576   -7.893     +     +    +     +     +     +                                      +                    68 -650.852 1443.7 15.88  0.000
1600  -8.138     +     +    +     +     +     +                                      +  0.01823000          69 -650.012 1444.2 16.38  0.000
2624  -7.582     +     +    +     +     +     +                                      +           -0.636900 69 -650.603 1445.4 17.57  0.000
704   -7.633     +     +    +     +     +     +             -0.510100                +                    69 -650.714 1445.6 17.79  0.000
3648  -7.802     +     +    +     +     +     +                                      +  0.01883000 -0.701600 70 -649.712 1445.8 17.97  0.000
640   -7.894     +     +    +     +     +     +  0.000037860                         +                    69 -650.852 1445.9 18.06  0.000
1728  -7.886     +     +    +     +     +     +             -0.490300                +  0.01811000          70 -649.885 1446.1 18.31  0.000
1664  -8.127     +     +    +     +     +     + -0.000773300                         +  0.01828000          70 -650.010 1446.4 18.56  0.000
1344  -8.589     +     +    +     +     +     +                                +     +     0.02395000          72 -648.003 1446.8 18.92  0.000
320   -8.273     +     +    +     +     +     +                                +                          71 -649.424 1447.4 19.58  0.000
2752  -7.423     +     +    +     +     +     +             -0.383400                +           -0.562200 70 -650.529 1447.4 19.60  0.000
2688  -7.583     +     +    +     +     +     +  0.000088980                         +           -0.636900 70 -650.603 1447.6 19.75  0.000
768   -7.639     +     +    +     +     +     +  0.000470200 -0.513000               +                    70 -650.713 1447.8 19.97  0.000
3776  -7.656     +     +    +     +     +     +             -0.347100                +  0.01869000 -0.633700 71 -649.651 1447.9 20.03  0.000
3712  -7.791     +     +    +     +     +     + -0.000750400                         +  0.01888000 -0.701400 71 -649.710 1448.0 20.15  0.000
1792  -7.882     +     +    +     +     +     + -0.000367600 -0.488100               +  0.01814000          71 -649.885 1448.3 20.50  0.000
1472  -8.930     +     +    +     +     +     +              0.665100          +     0.02418000          73 -647.796 1448.5 20.70  0.000
3392  -8.356     +     +    +     +     +     +                                +     0.02415000 -0.440600 73 -647.886 1448.7 20.88  0.000
574   -7.692     +           +     +     +     +                                   +                    64 -657.760 1448.8 21.00  0.000
1408  -8.566     +     +    +     +     +     + -0.001441000                   +     0.02403000          73 -647.996 1448.9 21.10  0.000
448   -8.586     +     +    +     +     +     +              0.619900          +                          72 -649.242 1449.2 21.40  0.000
2368  -8.059     +     +    +     +     +     +                                +                 -0.400800 72 -649.326 1449.4 21.57  0.000
1598  -7.937     +           +     +     +     +                                      +  0.01736000          65 -656.959 1449.4 21.57  0.000
 [ reached getOption("max.print") -- omitted 4041 rows ]
Models ranked by AICc(x)
```

Figure 8.2
Model selection data mining exercise



```
Call:
model.avg(object = get.models(freqfits, subset = TRUE))

Component model call:
glm(formula = LossesIndicator ~ <4096 unique rhs>, family = binomial(link = "logit"), data = crs$training, offset
    = log(Exposure))

Component models:
                                df   logLik    AICc  delta weight
1/2/3/4/5/6/9/10                74 -636.35 1427.83   0.00   0.21
1/2/3/4/5/6/9/10/11             75 -635.78 1428.89   1.06   0.12
1/2/3/4/5/6/8/9/10              75 -636.03 1429.38   1.56   0.10
1/2/3/4/5/6/9/10/12             75 -636.30 1429.93   2.10   0.07
1/2/3/4/5/6/7/9/10              75 -636.35 1430.02   2.20   0.07
1/2/3/4/5/6/8/9/10/11           76 -635.45 1430.43   2.61   0.06
1/2/3/4/5/6/9/10/11/12          76 -635.71 1430.96   3.13   0.04
1/2/3/4/5/6/7/9/10/11           76 -635.78 1431.09   3.26   0.04
1/2/3/4/5/6/8/9/10/12           76 -635.91 1431.35   3.52   0.04
1/2/3/4/5/6/7/8/9/10            76 -636.03 1431.58   3.76   0.03
1/3/4/5/6/9/10                  70 -642.65 1431.66   3.84   0.03
1/2/3/4/5/6/7/9/10/12           76 -636.30 1432.12   4.30   0.02
1/2/3/4/5/6/8/9/10/11/12        77 -635.31 1432.35   4.53   0.02
1/2/3/4/5/6/7/8/9/10/11         77 -635.45 1432.64   4.81   0.02
1/3/4/5/6/9/10/11               71 -642.19 1432.94   5.11   0.02
1/2/3/4/5/6/7/9/10/11/12        77 -635.71 1433.16   5.33   0.01
1/3/4/5/6/8/9/10                71 -642.47 1433.50   5.67   0.01
1/2/3/4/5/6/7/8/9/10/12         77 -635.91 1433.56   5.73   0.01
1/3/4/5/6/7/9/10                71 -642.64 1433.84   6.01   0.01
1/3/4/5/6/9/10/12               71 -642.65 1433.85   6.02   0.01
1/2/3/4/5/6/7/8/9/10/11/12      78 -635.31 1434.56   6.73   0.01
1/3/4/5/6/8/9/10/11             72 -642.01 1434.77   6.94   0.01
1/3/4/5/6/7/9/10/11             72 -642.18 1435.11   7.29   0.01
1/3/4/5/6/9/10/11/12            72 -642.19 1435.13   7.30   0.01
1/3/4/5/6/7/8/9/10              72 -642.46 1435.66   7.83   0.00
1/3/4/5/6/8/9/10/12             72 -642.47 1435.68   7.85   0.00
1/3/4/5/6/7/9/10/12             72 -642.64 1436.03   8.20   0.00
1/3/4/5/6/7/8/9/10/11           73 -642.00 1436.93   9.11   0.00
1/3/4/5/6/8/9/10/11/12          73 -642.00 1436.94   9.12   0.00
1/3/4/5/6/7/9/10/11/12          73 -642.18 1437.31   9.48   0.00
1/3/4/5/6/7/8/9/10/12           73 -642.45 1437.85  10.02   0.00
1/3/4/5/6/7/8/9/10/11/12        74 -641.99 1439.11  11.28   0.00
1/2/3/4/5/6/10                  68 -650.85 1443.71  15.88   0.00
1/2/3/4/5/6/10/11               69 -650.01 1444.21  16.38   0.00
1/2/3/4/5/6/10/12               69 -650.60 1445.39  17.57   0.00
1/2/3/4/5/6/8/10                69 -650.71 1445.62  17.79   0.00
1/2/3/4/5/6/10/11/12            70 -649.71 1445.80  17.97   0.00
1/2/3/4/5/6/7/10                69 -650.85 1445.89  18.06   0.00
1/2/3/4/5/6/8/10/11             70 -649.89 1446.14  18.31   0.00
1/2/3/4/5/6/7/10/11             70 -650.01 1446.39  18.56   0.00
1/2/3/4/5/6/9/11                72 -648.00 1446.75  18.92   0.00
1/2/3/4/5/6/9                   71 -649.42 1447.41  19.58   0.00
1/2/3/4/5/6/8/10/12             70 -650.53 1447.43  19.60   0.00
1/2/3/4/5/6/7/10/12             70 -650.60 1447.58  19.75   0.00
1/2/3/4/5/6/7/8/10              70 -650.71 1447.80  19.97   0.00
1/2/3/4/5/6/8/10/11/12          71 -649.65 1447.86  20.03   0.00
1/2/3/4/5/6/7/10/11/12          71 -649.71 1447.98  20.15   0.00
1/2/3/4/5/6/7/8/10/11           71 -649.88 1448.33  20.50   0.00
1/2/3/4/5/6/8/9/11              73 -647.80 1448.53  20.70   0.00
1/2/3/4/5/6/9/11/12             73 -647.89 1448.71  20.88   0.00
1/3/4/5/6/10                    64 -657.76 1448.83  21.00   0.00
1/2/3/4/5/6/7/9/11              73 -648.00 1448.93  21.10   0.00
1/2/3/4/5/6/8/9                 72 -649.24 1449.23  21.40   0.00
1/2/3/4/5/6/9/12                72 -649.33 1449.40  21.57   0.00
1/3/4/5/6/10/11                 65 -656.96 1449.40  21.57   0.00
1/2/3/4/5/6/7/9                 72 -649.42 1449.59  21.76   0.00
1/2/3/4/5/6/7/8/10/12           71 -650.53 1451.29  21.79   0.00
1/2/3/4/5/6/7/8/10/11/12        72 -649.65 1450.05  22.22   0.00
1/2/3/4/5/6/8/9/11/12           74 -647.60 1450.33  22.51   0.00
1/3/4/5/6/8/10                  65 -657.54 1450.56  22.73   0.00
1/2/3/4/5/6/7/8/9/11            74 -647.79 1450.70  22.88   0.00
1/2/3/4/5/6/7/9/11/12           74 -647.88 1450.89  23.06   0.00
1/3/4/5/6/10/12                 65 -657.73 1450.94  23.11   0.00
1/3/4/5/6/7/10                  65 -657.75 1450.98  23.15   0.00
1/2/3/4/5/6/8/9/12              73 -649.08 1451.09  23.26   0.00
1/3/4/5/6/8/10/11               66 -656.74 1451.14  23.31   0.00
1/2/3/4/5/6/7/8/9               73 -649.24 1451.42  23.59   0.00
1/3/4/5/6/10/11/12              66 -656.91 1451.47  23.64   0.00
1/3/4/5/6/7/10/11               66 -656.94 1451.54  23.71   0.00
1/2/3/4/5/6/7/9/12              73 -649.32 1451.59  23.76   0.00
1/3/4/5/6/9/11                  68 -655.17 1452.35  24.53   0.00
1/3/4/5/6/9                     67 -656.30 1452.43  24.60   0.00
1/2/3/4/5/6/7/8/9/11/12         75 -647.59 1452.51  24.68   0.00
1/3/4/5/6/8/10/12               66 -657.53 1452.72  24.89   0.00
1/3/4/5/6/7/8/10                66 -657.53 1452.72  24.90   0.00
1/3/4/5/6/7/10/12               66 -657.72 1453.09  25.26   0.00
1/3/4/5/6/8/10/11/12            67 -656.72 1453.28  25.45   0.00
1/2/3/4/5/6/7/8/9/12            74 -649.07 1453.28  25.45   0.00
1/3/4/5/6/7/8/10/11             67 -656.73 1453.30  25.47   0.00
1/3/4/5/6/7/10/11/12            67 -656.89 1453.61  25.78   0.00
1/3/4/5/6/8/9/11                69 -655.10 1454.40  26.57   0.00
1/3/4/5/6/8/9                   68 -656.23 1454.48  26.65   0.00
1/3/4/5/6/7/9/11                69 -655.15 1454.48  26.66   0.00
1/3/4/5/6/9/11/12               69 -655.15 1454.49  26.66   0.00
1/3/4/5/6/7/9                   68 -656.28 1454.56  26.74   0.00
1/3/4/5/6/9/12                  68 -656.29 1454.58  26.75   0.00
1/3/4/5/6/7/8/10/12             67 -657.53 1454.88  27.06   0.00
1/2/3/4/5/6/11                  66 -658.69 1455.04  27.22   0.00
1/2/3/4/5/6                     65 -659.95 1455.38  27.55   0.00
1/3/4/5/6/7/8/10/11/12          68 -656.71 1455.44  27.61   0.00
1/2/3/4/5/6/11/12               67 -658.24 1456.30  28.48   0.00
1/3/4/5/6/8/9/11/12             70 -655.06 1456.50  28.67   0.00
1/3/4/5/6/7/8/9/11              70 -655.07 1456.52  28.69   0.00
1/3/4/5/6/8/9/12                69 -656.20 1456.59  28.77   0.00
1/3/4/5/6/7/8/9                 69 -656.21 1456.60  28.78   0.00
1/3/4/5/6/7/9/11/12             70 -655.12 1456.62  28.79   0.00
1/3/4/5/6/7/9/12                69 -656.26 1456.71  28.89   0.00
1/2/3/4/5/6/12                  66 -659.55 1456.75  28.92   0.00
1/2/3/4/5/6/8/11                67 -658.50 1456.82  28.99   0.00
1/2/3/4/5/6/8                   66 -659.71 1457.08  29.25   0.00
1/2/3/4/5/6/7/11                67 -658.67 1457.16  29.33   0.00
1/2/3/4/5/6/7                   66 -659.94 1457.53  29.70   0.00
1/2/3/4/5/6/8/11/12             68 -658.14 1458.30  30.47   0.00
1/2/3/4/5/6/7/11/12             68 -658.21 1458.43  30.60   0.00
1/3/4/5/6/7/8/9/11/12           71 -655.03 1458.62  30.79   0.00
1/2/3/4/5/6/8/12                67 -659.42 1458.67  30.85   0.00
1/3/4/5/6/7/8/9/12              70 -656.18 1458.72  30.90   0.00
1/2/3/4/5/6/7/12                67 -659.54 1458.90  31.07   0.00
1/2/3/4/5/6/7/8/11              68 -658.48 1458.96  31.13   0.00
1/2/3/4/5/6/7/8                 67 -659.71 1459.24  31.42   0.00
1/2/3/4/5/6/7/8/11/12           69 -658.12 1460.44  32.61   0.00
1/3/4/5/6                       61 -666.98 1460.77  32.95   0.00
1/2/3/4/5/6/7/8/12              68 -659.41 1460.84  33.01   0.00
1/3/4/5/6/11                    62 -665.95 1460.87  33.05   0.00
1/3/4/5/6/8                     62 -666.63 1462.24  34.41   0.00
1/3/4/5/6/8/11                  63 -665.61 1462.37  34.54   0.00
1/3/4/5/6/12                    62 -666.84 1462.65  34.83   0.00
1/3/4/5/6/11/12                 63 -665.77 1462.69  34.86   0.00
1/3/4/5/6/7                     62 -666.93 1462.84  35.01   0.00
1/3/4/5/6/7/11                  63 -665.89 1462.93  35.10   0.00
1/3/4/5/6/8/12                  63 -666.56 1464.27  36.44   0.00
1/3/4/5/6/7/8                   63 -666.60 1464.34  36.51   0.00
1/3/4/5/6/8/11/12               64 -665.52 1464.35  36.52   0.00
1/3/4/5/6/7/8/11                64 -665.57 1464.46  36.63   0.00
1/3/4/5/6/7/12                  63 -666.79 1464.73  36.90   0.00
1/3/4/5/6/7/11/12               64 -665.72 1464.75  36.92   0.00
1/3/4/5/6/7/8/12                64 -666.53 1466.37  38.55   0.00
1/3/4/5/6/7/8/11/12             65 -665.48 1466.45  38.62   0.00
1/2/3/4/6/9/10                  52 -680.50 1468.49  40.66   0.00
1/2/3/4/6/8/9/10                53 -679.64 1468.90  41.08   0.00
1/2/3/4/6/9/10/11               53 -679.79 1469.21  41.38   0.00
1/2/3/4/6/8/9/10/11             54 -678.96 1469.68  41.85   0.00
1/2/3/4/6/7/9/10                53 -680.38 1470.38  42.55   0.00
1/2/3/4/6/9/10/12               53 -680.50 1470.62  42.79   0.00
1/2/3/4/6/7/8/9/10              54 -679.48 1470.74  42.91   0.00
1/2/3/4/6/8/9/10/12             54 -679.59 1470.96  43.13   0.00
1/2/3/4/6/7/9/10/11             54 -679.66 1471.09  43.26   0.00
1/2/3/4/6/9/10/11/12            54 -679.78 1471.34  43.51   0.00
1/2/3/4/6/7/8/9/10/11           55 -678.79 1471.50  43.67   0.00
1/2/3/4/6/8/9/10/11/12          55 -678.87 1471.66  43.83   0.00
1/2/3/4/6/7/9/10/12             54 -680.38 1472.52  44.69   0.00
1/2/3/4/6/7/8/9/10/12           55 -679.44 1472.78  44.95   0.00
1/2/3/4/6/7/9/10/11/12          55 -679.65 1473.22  45.39   0.00
1/2/3/4/6/7/8/9/10/11/12        56 -678.71 1473.47  45.64   0.00
1/3/4/6/9/10                    48 -689.98 1478.93  51.10   0.00
1/3/4/6/9/10/11                 49 -688.99 1479.08  51.25   0.00
1/2/3/4/6/10                    46 -692.45 1479.62  51.79   0.00
1/3/4/6/8/9/10                  49 -689.41 1479.92  52.09   0.00
1/3/4/6/8/9/10/11               50 -688.44 1480.11  52.28   0.00
1/2/3/4/6/10/11                 47 -691.74 1480.34  52.51   0.00
1/3/4/6/7/9/10                  49 -689.76 1480.63  52.80   0.00
1/3/4/6/9/10/12                 49 -689.80 1480.71  52.88   0.00
1/3/4/6/7/9/10/11               50 -688.79 1480.81  52.98   0.00
1/3/4/6/9/10/11/12              50 -688.89 1481.01  53.19   0.00
1/2/3/4/6/7/10                  47 -692.24 1481.32  53.49   0.00
1/2/3/4/6/10/12                 47 -692.29 1481.43  53.60   0.00
1/3/4/6/7/8/9/10                50 -689.14 1481.51  53.69   0.00
1/2/3/4/6/8/10                  47 -692.44 1481.73  53.90   0.00
1/3/4/6/7/8/9/10/11             51 -688.19 1481.74  53.92   0.00
1/3/4/6/8/9/10/12               50 -689.34 1481.90  54.08   0.00
1/2/3/4/6/7/10/11               48 -691.51 1482.00  54.17   0.00
1/2/3/4/6/10/11/12              48 -691.52 1482.02  54.19   0.00
1/3/4/6/8/9/10/11/12            51 -688.41 1482.19  54.36   0.00
1/3/4/6/7/9/10/12               50 -689.60 1482.42  54.59   0.00
1/2/3/4/6/8/10/11               48 -691.74 1482.45  54.62   0.00
1/3/4/6/7/9/10/11/12            51 -688.70 1482.76  54.93   0.00
1/2/3/4/6/7/10/12               48 -692.08 1483.13  55.31   0.00
1/2/3/4/6/7/8/10                48 -692.23 1483.44  55.61   0.00
1/3/4/6/7/8/9/10/12             51 -689.08 1483.52  55.69   0.00
1/2/3/4/6/8/10/12               48 -692.29 1483.56  55.73   0.00
1/2/3/4/6/7/10/11/12            49 -691.29 1483.69  55.86   0.00
1/3/4/6/7/8/9/10/11/12          52 -688.17 1483.84  56.01   0.00
1/2/3/4/6/7/8/10/11             49 -691.51 1484.12  56.29   0.00
1/2/3/4/6/8/10/11/12            49 -691.52 1484.15  56.32   0.00
1/2/3/4/6/7/8/10/12             49 -692.08 1485.26  57.43   0.00
2/3/4/5/6/9/10                  67 -672.85 1485.53  57.70   0.00
1/2/3/4/6/7/8/10/11/12          50 -691.29 1485.81  57.98   0.00
2/3/4/5/6/8/9/10                68 -671.99 1485.98  58.16   0.00
1/2/3/4/6/9/11                  50 -691.57 1486.37  58.54   0.00
2/3/4/5/6/9/10/11               68 -672.48 1486.96  59.13   0.00
1/2/3/4/6/9                     49 -693.07 1487.25  59.42   0.00
2/3/4/5/6/8/9/10/11             69 -671.59 1487.38  59.55   0.00
1/2/3/4/6/8/9/11                51 -691.04 1487.43  59.60   0.00
2/3/4/5/6/7/9/10                68 -672.81 1487.64  59.81   0.00
2/3/4/5/6/9/10/12               68 -672.85 1487.71  59.88   0.00
2/3/4/5/6/8/9/10/12             69 -671.92 1488.04  60.21   0.00
2/3/4/5/6/7/8/9/10              69 -671.93 1488.05  60.22   0.00
1/2/3/4/6/8/9                   50 -692.53 1488.29  60.46   0.00
1/2/3/4/6/7/9/11                51 -691.50 1488.36  60.53   0.00
1/2/3/4/6/8/9/11                51 -691.04 1487.43  59.60   0.00
2/3/4/5/6/7/9/10                68 -672.81 1487.64  59.81   0.00
2/3/4/5/6/9/10/12               68 -672.85 1487.71  59.88   0.00
2/3/4/5/6/8/9/10/12             69 -671.92 1488.04  60.21   0.00
2/3/4/5/6/7/8/9/10              69 -671.93 1488.05  60.22   0.00
1/2/3/4/6/8/9                   50 -692.53 1488.29  60.46   0.00
1/2/3/4/6/7/9/11                51 -691.50 1488.36  60.53   0.00
1/2/3/4/6/9/11/12               51 -691.51 1488.38  60.55   0.00
2/3/4/5/6/7/9/10/11             69 -672.43 1489.05  61.22   0.00
2/3/4/5/6/9/10/11/12            69 -672.47 1489.13  61.30   0.00
1/2/3/4/6/8/9/11/12             52 -690.85 1489.20  61.37   0.00
1/2/3/4/6/7/9                   50 -693.01 1489.24  61.41   0.00
1/2/3/4/6/9/12                  50 -693.05 1489.33  61.50   0.00
1/2/3/4/6/7/8/9/11              52 -690.95 1489.39  61.56   0.00
2/3/4/5/6/8/9/10/11/12          70 -671.52 1489.40  61.58   0.00
2/3/4/5/6/7/8/9/10/11           70 -671.53 1489.42  61.60   0.00
2/3/4/5/6/7/9/10/12             69 -672.81 1489.81  61.99   0.00
2/3/4/5/6/7/8/9/10/12           70 -671.87 1490.11  62.28   0.00
 [ reached getOption("max.print") -- omitted 3896 rows ]


Term codes:
      BusinessLineLevel1                  CapturedBy                        Desk EventTypeCategoryLevel1                  Instrument
                       1                           2                           3                           4                           5
                  Reason                  TradedDay                 TradedTime                      TraderId                 TradeStatus
                       6                           7                           8                           9                          10
             UpdatedDay                 UpdatedTime
                      11                          12
```

Figure 8.3
Estimation of some Poisson distribution for target variable `LossesIndicator` showing the 4096 component models and their respective AICc values for our models giving rise to the average (mean) model `Amodel`

Model-averaged coefficients:
(full average)

	Estimate	Std. Error	Adjusted SE	z value	Pr(>|z|)	
(Intercept)	-7.67274874	0.88134801	0.88198452	8.699	< 0.0000000000000002	***
BusinessLineLevel1BL1	2.75817912	1.01714718	1.01793710	2.710	0.006737	**
BusinessLineLevel1BL3	-1.68663571	0.88871394	0.88940429	1.896	0.057912	.
BusinessLineLevel1BL4	-4.15598982	0.71986432	0.72041781	5.769	< 0.0000000000000002	***
BusinessLineLevel1BL5	-2.32395162	0.69224043	0.69277670	3.355	0.000795	***
BusinessLineLevel1BL6	0.98066548	1.41041943	1.41149881	0.695	0.487200	
BusinessLineLevel1BL7	-16.40673180	2706.21778957	2708.32407004	0.006	0.995166	
BusinessLineLevel1BL9	23.98386853	1515.45494761	1516.63446709	0.016	0.987383	
CapturedByMIDOFFICE	0.78315372	0.50952279	0.50979350	1.536	0.124485	
CapturedByPROD ACCOUNTANT	0.40816445	0.66783818	0.66833211	0.611	0.541384	
CapturedByPROD CONTROLLER	-1.08579393	0.59483324	0.59508929	1.825	0.068063	.
CapturedByUNAUTHORISED	0.39005509	0.76909787	0.76967583	0.507	0.612310	
DeskAfrica	4.40227494	0.90481658	0.90551795	4.862	0.0000012	***
DeskBonds/Repos	5.06476675	0.65418976	0.65468554	7.736	< 0.0000000000000002	***
DeskCommodities	1.85999819	0.73511032	0.73556714	2.529	0.011450	*
DeskDerivatives	-0.11630021	0.74992884	0.75047722	0.155	0.876846	
DeskEquity	1.88124888	0.59068584	0.59114409	3.182	0.001461	**
DeskManagement/Other	-18.03094378	1515.45421098	1516.63372990	0.012	0.990514	
DeskMM	2.05252967	0.79042564	0.79102782	2.595	0.009466	**
DeskPrime Services	0.33643654	1.32369843	1.32469261	0.254	0.799516	
DeskSND	2.32314753	1.05903192	1.05984894	2.192	0.028382	*
EventTypeCategoryLevel1EL1	44.98315636	1550.59911616	1551.80599541	0.029	0.976874	
EventTypeCategoryLevel1EL4	1.29626733	0.60840285	0.60886853	2.129	0.033256	*
EventTypeCategoryLevel1EL6	82.40468924	1902.47603399	1903.95679116	0.043	0.965478	
EventTypeCategoryLevel1EL8	-5.45492979	1.35306713	1.35409678	4.028	0.0000561	***
InstrumentBill	-0.87804917	0.95326994	0.95399896	0.920	0.357370	
InstrumentBond	0.26836138	0.60440594	0.60487483	0.444	0.657285	
InstrumentBuySellback	1.31366297	1.03843899	1.03924653	1.264	0.206211	
InstrumentCall Deposit	-2.20958551	0.59653314	0.59699611	3.701	0.000215	***
InstrumentCombination	-17.10828000	742.87301135	743.45121878	0.023	0.981641	
InstrumentCreditDefaultSwap	-1.45956141	0.95915219	0.95989694	1.521	0.128375	
InstrumentCurr	-2.80763440	0.72155874	0.72211891	3.888	0.000101	***
InstrumentCurrSwap	-2.02191314	0.77001102	0.77060371	2.624	0.008695	**
InstrumentDeposit	0.27396391	0.50022088	0.50060623	0.547	0.584197	
InstrumentEquityIndex	-20.82006732	4310.04835103	4313.40254587	0.005	0.996149	
InstrumentETF	0.83286782	2.51672330	2.51868006	0.331	0.740889	
InstrumentFRA	-18.37756626	2217.66952531	2219.39557059	0.008	0.993393	
InstrumentFRN	1.12211650	0.69182246	0.69235869	1.621	0.105079	
InstrumentFuture/Forward	-1.27630580	0.50669400	0.50707570	2.517	0.011836	*
InstrumentIndexLinkedBond	-18.48306179	1994.96229046	1996.51503444	0.009	0.992614	
InstrumentIndexLinkedSwap	-1.29230205	0.82889991	0.82954251	1.558	0.119269	
InstrumentOption	-0.33433058	0.73675506	0.73732668	0.453	0.650235	
InstrumentOther	-0.70919696	0.42421547	0.42454499	1.670	0.094823	.
InstrumentRepo/Reverse	-1.63632522	0.87164256	0.87231642	1.876	0.060677	.
InstrumentSecurityLoan	-1.61173947	0.50587140	0.50626463	3.184	0.001455	**
InstrumentStock	-15.64175662	916.44030501	917.15359818	0.017	0.986393	
InstrumentTotalReturnSwap	-2.10575078	0.50792453	0.50831768	4.143	0.0000343	***
ReasonAcquirer	-16.18300432	10691.13203139	10699.45264887	0.002	0.998793	
ReasonBrokerage Related	4.12922953	2.27268995	2.27445258	1.815	0.069450	.
ReasonCalendar Related	2.83525534	0.52701967	0.52742873	5.376	0.0000001	***
ReasonCapture Errors - Direction / Amount / Rate / CP	1.47596117	0.52463852	0.52504428	2.811	0.004937	**
ReasonClient Request to Amend Economics of Deal	-0.37233958	1.20712640	1.20806252	0.308	0.757921	
ReasonCommodities Early Delivery	2.30721212	1.22548586	1.22643465	1.881	0.059940	.
ReasonCorrected NDEUSSA Reset Level	0.44188166	0.94714707	0.94788370	0.466	0.641089	
ReasonCPI Fixings	-16.49019139	1175.33889976	1176.25370325	0.014	0.988815	
ReasonEarly Termination / Delivery / Close out	-43.73201578	1550.59939524	1551.80627465	0.028	0.977518	
ReasonFees/Commissions Related	-0.89549786	1.09520075	1.09604086	0.817	0.413911	
ReasonOperations request to change Economics of Trade	-41.48456832	1550.59950623	1551.80638576	0.027	0.978673	
ReasonPayments Related	0.66959105	0.77279276	0.77338539	0.866	0.386604	
ReasonPortfolio Move / Restructure	-14.76941522	850.11970155	850.78137763	0.017	0.986150	
ReasonSales Credits	2.69384329	1.40178104	1.40279472	1.920	0.054815	.
ReasonSystem Update Call Accounts	-80.33483411	1902.47618417	1903.95694146	0.042	0.966344	
ReasonTrade Restructure	-56.15162037	2910.67448815	2912.93993018	0.019	0.984620	
ReasonTri-Optima	-2.03259346	1.46443081	1.46553876	1.387	0.165464	
ReasonValuation Group Request	2.29287069	0.96874037	0.96949319	2.365	0.018029	*
TraderIdAMBA	-0.06898980	0.78797950	0.78858829	0.087	0.930286	
TraderIdANALYST	-0.62114195	0.43438322	0.43471851	1.429	0.153051	
TraderIdASSOCIATE	-0.92937483	0.53206218	0.53247536	1.745	0.080918	.
TraderIdATS	2.49644634	0.86929120	0.86994828	2.870	0.004109	**
TraderIdMNGDIRECTOR	1.35793363	0.51989652	0.52029804	2.610	0.009057	**
TraderIdVICE PRINCIPAL	-0.74521632	0.50533173	0.50571837	1.474	0.140595	
TradeStatusBO-BO Confirmed	-1.56247589	0.32928745	0.32954116	4.741	0.0000021	***
TradeStatusTerminated	1.60721945	1.37926613	1.38030264	1.164	0.244263	
TradeStatusTerminated/Void	-15.68530720	4643.11569269	4646.72943324	0.003	0.997307	
UpdatedDay	0.00562264	0.01142398	0.01142917	0.492	0.622751	
TradedTime	0.26424360	0.70621368	0.70659358	0.374	0.708428	
UpdatedTime	-0.08515101	0.49928589	0.49963602	0.170	0.864675	
TradedDay	-0.00001636	0.00601969	0.00602436	0.003	0.997833	

Figure 8.4
Continued from 8.3

(conditional average)

	Estimate	Std. Error	Adjusted SE	z value	Pr(>|z|)	
(Intercept)	-7.67274874	0.88134801	0.88198452	8.699	< 0.0000000000000002	***
BusinessLineLevel1BL1	2.75817912	1.01714718	1.01793710	2.710	0.006737	**
BusinessLineLevel1BL3	-1.68663571	0.88871394	0.88940429	1.896	0.057912	.
BusinessLineLevel1BL4	-4.15598982	0.71986432	0.72041781	5.769	< 0.0000000000000002	***
BusinessLineLevel1BL5	-2.32395162	0.69224043	0.69277670	3.355	0.000795	***
BusinessLineLevel1BL6	0.98066548	1.41041943	1.41149881	0.695	0.487200	
BusinessLineLevel1BL7	-16.40673180	2706.21778957	2708.32407004	0.006	0.995166	
BusinessLineLevel1BL9	23.98386853	1515.45494761	1516.63446709	0.016	0.987383	
CapturedByMIDOFFICE	0.88771856	0.44883438	0.44918269	1.976	0.048121	*
CapturedByPROD ACCOUNTANT	0.46266160	0.69306854	0.69360803	0.667	0.504749	
CapturedByPROD CONTROLLER	-1.23076657	0.47184870	0.47221452	2.606	0.009151	**
CapturedByUNAUTHORISED	0.44213433	0.80465070	0.80527687	0.549	0.582974	
DeskAfrica	4.40227494	0.90481658	0.90551795	4.862	0.0000012	***
DeskBonds/Repos	5.06476675	0.65418976	0.65468554	7.736	< 0.0000000000000002	***
DeskCommodities	1.85999819	0.73511032	0.73556714	2.529	0.011450	*
DeskDerivatives	-0.11630021	0.74992884	0.75047722	0.155	0.876846	
DeskEquity	1.88124888	0.59068584	0.59114409	3.182	0.001461	**
DeskManagement/Other	-18.03094378	1515.45421098	1516.63372990	0.012	0.990514	
DeskMM	2.05252967	0.79042564	0.79102782	2.595	0.009466	**
DeskPrime Services	0.33643654	1.32369843	1.32469261	0.254	0.799516	
DeskSND	2.32314753	1.05903192	1.05984894	2.192	0.028382	*
EventTypeCategoryLevel1EL1	44.98315636	1550.59911616	1551.80599541	0.029	0.976874	
EventTypeCategoryLevel1EL4	1.29626733	0.60840285	0.60886853	2.129	0.033256	*
EventTypeCategoryLevel1EL6	82.40468924	1902.47603399	1903.95679116	0.043	0.965478	
EventTypeCategoryLevel1EL8	-5.45492979	1.35306713	1.35409678	4.028	0.0000561	***
InstrumentBill	-0.87804917	0.95326994	0.95399896	0.920	0.357370	
InstrumentBond	0.26836138	0.60440594	0.60487483	0.444	0.657285	
InstrumentBuySellback	1.31366297	1.03843899	1.03924653	1.264	0.206211	
InstrumentCall Deposit	-2.20958552	0.59653313	0.59699611	3.701	0.000215	***
InstrumentCombination	-17.10828003	742.87301202	743.45121945	0.023	0.981641	
InstrumentCreditDefaultSwap	-1.45956141	0.95915219	0.95989694	1.521	0.128375	
InstrumentCurr	-2.80763440	0.72155873	0.72211890	3.888	0.000101	***
InstrumentCurrSwap	-2.02191314	0.77001101	0.77060370	2.624	0.008695	**
InstrumentDeposit	0.27396391	0.50022088	0.50060623	0.547	0.584197	
InstrumentEquityIndex	-20.82006736	4310.04835493	4313.40254977	0.005	0.996149	
InstrumentETF	0.83286782	2.51672330	2.51868006	0.331	0.740889	
InstrumentFRA	-18.37756630	2217.66952732	2219.39557260	0.008	0.993393	
InstrumentFRN	1.12211650	0.69182246	0.69235868	1.621	0.105079	
InstrumentFuture/Forward	-1.27630580	0.50669399	0.50707569	2.517	0.011836	*
InstrumentIndexLinkedBond	-18.48306183	1994.96229227	1996.51503624	0.009	0.992614	
InstrumentIndexLinkedSwap	-1.29230205	0.82889991	0.82954251	1.558	0.119269	
InstrumentOption	-0.33433058	0.73675507	0.73732669	0.453	0.650235	
InstrumentOther	-0.70919696	0.42421547	0.42454499	1.670	0.094823	.
InstrumentRepo/Reverse	-1.63632522	0.87164256	0.87231642	1.876	0.060677	.
InstrumentSecurityLoan	-1.61173948	0.50587139	0.50626462	3.184	0.001455	**
InstrumentStock	-15.64175665	916.44030584	917.15359901	0.017	0.986393	
InstrumentTotalReturnSwap	-2.10575078	0.50792452	0.50831767	4.143	0.0000343	***
ReasonAcquirer	-16.18300432	10691.13203139	10699.45264887	0.002	0.998793	
ReasonBrokerage Related	4.12922953	2.27268995	2.27445258	1.815	0.069450	.
ReasonCalendar Related	2.83525534	0.52701967	0.52742873	5.376	0.0000001	***
ReasonCapture Errors - Direction / Amount / Rate / CP	1.47596117	0.52463852	0.52504428	2.811	0.004937	**
ReasonClient Request to Amend Economics of Deal	-0.37233958	1.20712640	1.20806252	0.308	0.757921	
ReasonCommodities Early Delivery	2.30721212	1.22548586	1.22643465	1.881	0.059940	.
ReasonCorrected NDEUSSA Reset Level	0.44188166	0.94714707	0.94788370	0.466	0.641089	
ReasonCPI Fixings	-16.49019139	1175.33889976	1176.25370325	0.014	0.988815	
ReasonEarly Termination / Delivery / Close out	-43.73201578	1550.59939524	1551.80627465	0.028	0.977518	
ReasonFees/Commissions Related	-0.89549786	1.09520075	1.09604086	0.817	0.413911	
ReasonOperations request to change Economics of Trade	-41.48456832	1550.59950623	1551.80638576	0.027	0.978673	
ReasonPayments Related	0.66959105	0.77279276	0.77338539	0.866	0.386604	
ReasonPortfolio Move / Restructure	-14.76941522	850.11970155	850.78137763	0.017	0.986150	
ReasonSales Credits	2.69384329	1.40178104	1.40279472	1.920	0.054815	.
ReasonSystem Update Call Accounts	-80.33483411	1902.47618417	1903.95694146	0.042	0.966344	
ReasonTrade Restructure	-56.15162037	2910.67448815	2912.93993018	0.019	0.984620	
ReasonTri-Optima	-2.03259346	1.46443081	1.46553876	1.387	0.165464	
ReasonValuation Group Request	2.29287069	0.96874037	0.96949319	2.365	0.018029	*
TraderIdAMBA	-0.06901594	0.78812762	0.78873654	0.088	0.930273	
TraderIdANALYST	-0.62137731	0.43429717	0.43463265	1.430	0.152814	
TraderIdASSOCIATE	-0.92972698	0.53185526	0.53226877	1.747	0.080685	.
TraderIdATS	2.49739227	0.86809629	0.86875453	2.875	0.004044	**
TraderIdMNGDIRECTOR	1.35844817	0.51932249	0.51972460	2.614	0.008955	**
TraderIdVICE PRINCIPAL	-0.74549868	0.50521917	0.50560604	1.474	0.140356	
TradeStatusBO-BO Confirmed	-1.56259946	0.32900716	0.32926111	4.746	0.0000021	***
TradeStatusTerminated	1.60734656	1.37924661	1.38028321	1.165	0.244220	
TradeStatusTerminated/Void	-15.68654770	4643.29929107	4646.91317452	0.003	0.997307	
UpdatedDay	0.01525549	0.01439242	0.01440361	1.059	0.289534	
TradedTime	0.83807203	1.04922901	1.05003989	0.798	0.424793	
UpdatedTime	-0.32359393	0.93283894	0.93355109	0.347	0.728872	
TradedDay	-0.00006542	0.01203573	0.01204507	0.005	0.995667	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure 8.5
Continued from 8.4

```
                  sep=",",
                  dec=".",
                  na.strings=c(".", "NA", "", "?"),
                  strip.white=TRUE, encoding="UTF-8")
```

Generate the confusion matrix showing counts and generate an ROC curve for the GLM model on the appropriate dataset partition

```
confusionMatrix(table(pred$response, crs[["testing"]][["LossesIndicator"]]))

crs$pr <- Av.PredTT

# Remove observations with missing target.

no.miss   <- na.omit(crs[["testing"]][["LossesIndicator"]])
miss.list <- attr(no.miss, "na.action")
attributes(no.miss) <- NULL

if (length(miss.list))
{
  predic <- prediction(crs$pr[-miss.list], no.miss)
} else
{
  predic <- prediction(crs$pr, no.miss)
}

pe <- performance(predic, "tpr", "fpr")
au <- performance(predic, "auc")@y.values[[1]]
pd <- data.frame(fpr=unlist(pe@x.values), tpr=unlist(pe@y.values))
p <- ggplot(pd, aes(x=fpr, y=tpr))
p <- p + geom_line(colour="red")
p <- p + xlab("False Positive Rate") + ylab("True Positive Rate")
p <- p + ggtitle("ROC Curve Linear OPriskDataSet_exposure.csv [validate] LossesInd
p <- p + theme(plot.title=element_text(size=10))
p <- p + geom_line(data=data.frame(), aes(x=c(0,1), y=c(0,1)), colour="grey")
p <- p + annotate("text", x=0.50, y=0.00, hjust=0, vjust=0, size=5,
                  label=paste("AUC =", round(au, 2)))
print(p)

# Calculate the area under the curve for the plot.


# Remove observations with missing target.

no.miss   <- na.omit(crs[["testing"]][["LossesIndicator"]])
miss.list <- attr(no.miss, "na.action")
```

```
attributes(no.miss) <- NULL

if (length(miss.list))
{
  predic <- prediction(crs$pr[-miss.list], no.miss)
} else
{
  predic <- prediction(crs$pr, no.miss)
}
performance(predic, "auc")
```

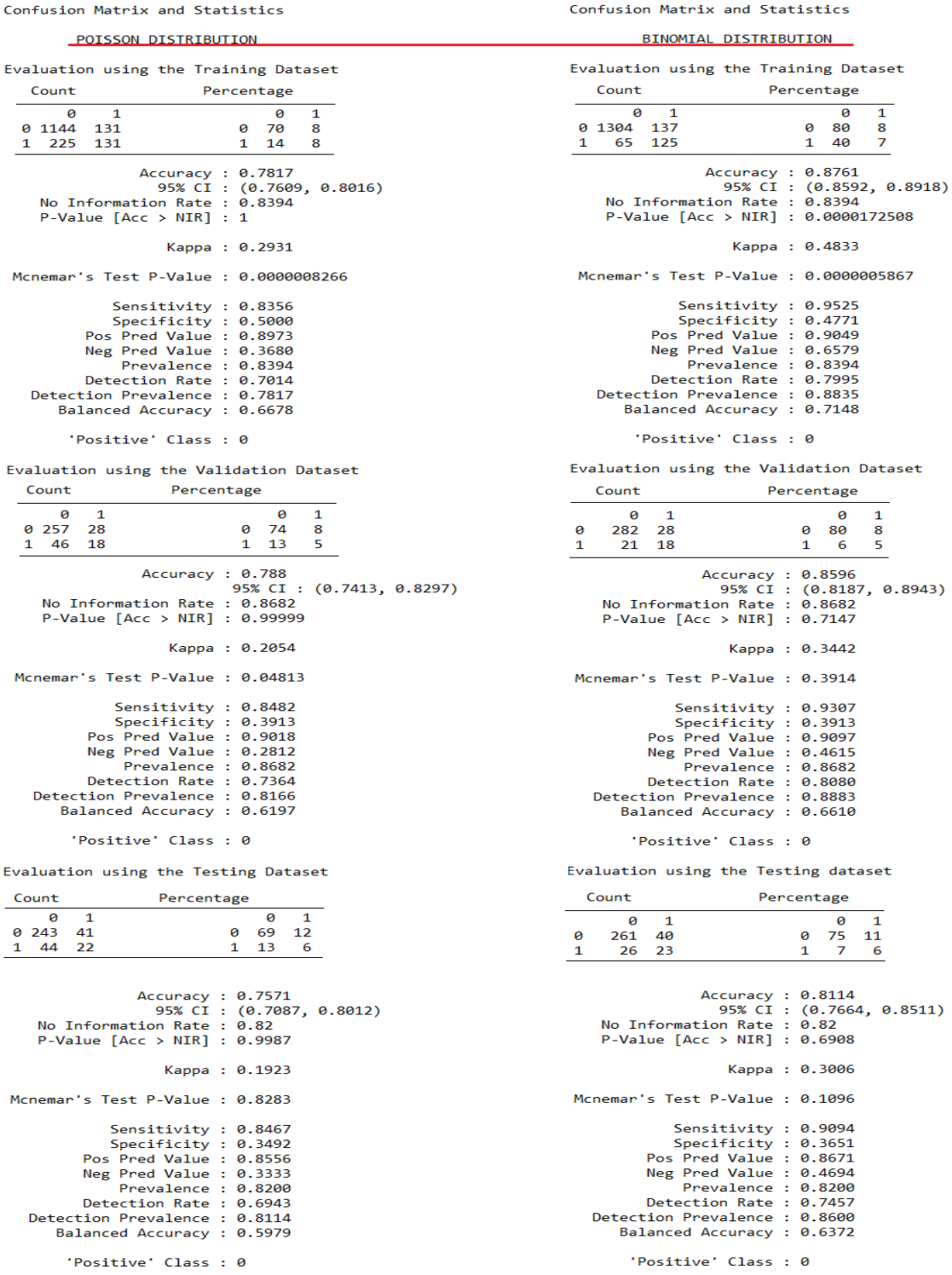
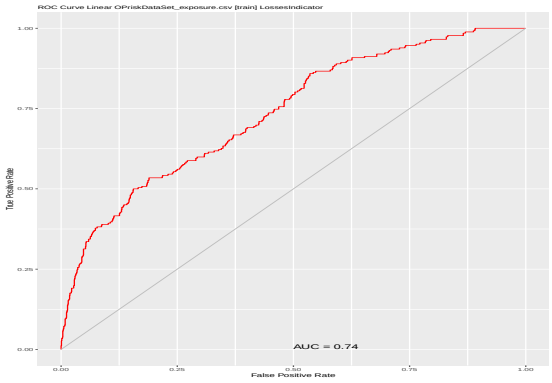
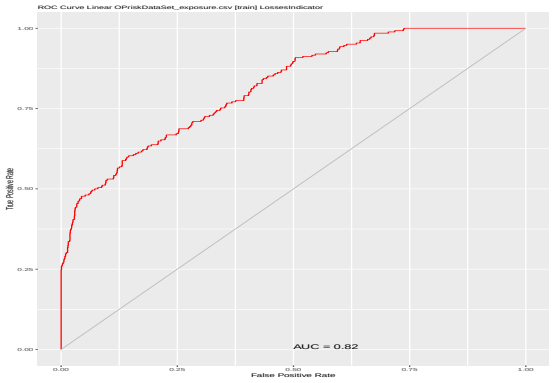


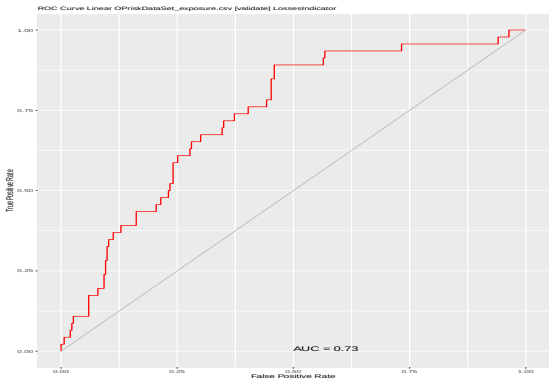
Figure 8.6
Comparison of Poisson, Binomial GLM's confusion matrices for Training, Validation & Testing datasets



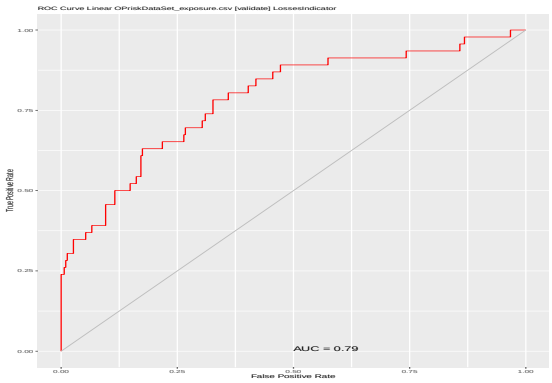
(a) AUC for poisson parameter depicting model's performance on the training set



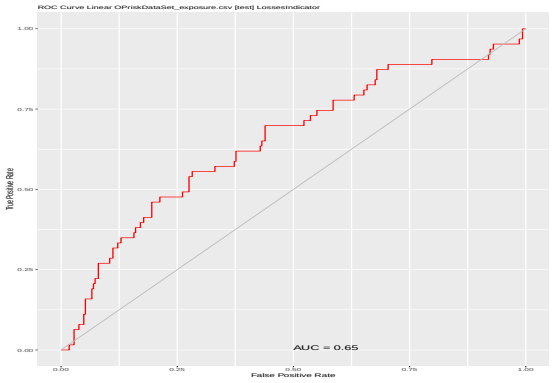
(b) AUC for binomial parameters depicting model's performance on the training set



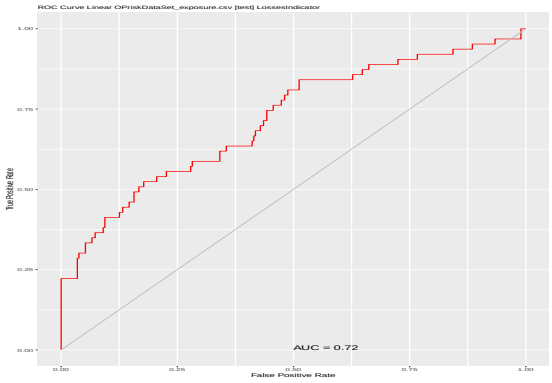
(c) As in (a) but depicting model's performance on the validation set



(d) As in (b) but depicting model's performance on the validation set



(e) As in (a) & (c) depicting model's performance on the testing set



(f) As in (b) & (d) depicting model's performance on the testing set

Figure 8.7
Comparison of area under the curve (AUC) for the Poisson, Binomial GLM ROC curve for training, validation and testing datasets

Appendix C: R Code for Chapter 5

Appendix C: R Code for Chapter 5 (GAMLSS)

Required: R Packages from CRAN (in addition to packages already found in Chapter 3)

```
if (!require(gamlss)){
  install.packages("gamlss")
  library(gamlss)
}
```

Data exploration of OpRisk loss severity dataset in preparation for GAMLSS machine learning treatment. The raw data and the pre-processed datasets are initiated and analysed.

Plots exploring loss severity characteristics, see Chapter 5 Section 5.2.1 on page 133

```
options(scipen = 999)
file_loc <- "C:/Users/Mphekeleli/Documents/R PROJECT/OpRiskPHDGitHub/
OpRisk_PHD_Thesis/Data"
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_GAMLSS.xlsx",
                                check.names = TRUE, sheet = "CleanedData")

names(projdata) <- sub("\\.", "", names(projdata))
dput(names(projdata))

par(mar=c(1,1,1,1))
PPP <- par(mfrow=c(2,2))
plot(Loss ~ UpdateTime, data = projdata, col=gray(0.7), pch=15, cex=0.5)
plot(Loss ~ UpdatedDay, data = projdata, col=gray(0.7), pch=15, cex=0.5)
par(PPP)
```

Data partitioning of the pre-processed OpRisk dataset into Training/Validation/Testing proportions, in preparation for machine learning model building treatments. The original dataset is partitioned into three random subsets initiated by a random number sequence with a randomly selected seed.

```
# Load packages
library(rattle, quietly = TRUE)
library(magrittr, quietly = TRUE) # Utilize the %>% and %>% pipeline operators
library(Hmisc, quietly = TRUE)
library(chron, quietly = TRUE)
library(dplyr, quietly = TRUE)
library(ggplot2)
library(caTools)
library(caret)
library(gamlss)

building <- TRUE
scoring <- ! building

# A predefined value is used to reset the random seed so that results are repeatable

crv$seed <- 42 # set random seed to make your partition reproducible
# Load the dataset OPriskDataSet_exposure
#============================================================================
fname <- "file:///C:/Users/Mphekeleli/Documents/R PROJECT/OpRiskPHDGitHub/
OpRisk_PHD_Thesis/Data/OPriskDataSet_exposure_severity.csv"
crs$dataset <- read.csv(fname,
                        sep=",",
                        dec=".",
                        na.strings=c(".", "NA", "", "?"),
                        strip.white=TRUE, encoding="UTF-8")

exposure <- crs$dataset[,ncol(crs$dataset)]
crs$dataset <- as.data.frame(crs$dataset)

# The following varaible selections have been noted
crs$input <- crs$dataset %>%
  group_by(UpdatedDay,
           UpdatedTime,
           TradedDay,
           TradedTime,
           Desk,
           CapturedBy,
           TradeStatus,
           TraderId,
           Instrument,
```

```
           Reason,
           EventTypeCategoryLevel1,
           BusinessLineLevel1) %>%
  transmute(LossesIndicator = LossIndicator,
            Losses = Loss,
            Exposure = exposure)

# 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(crs$input) - 3)){
  crs$input[[i]] <- relevel(crs$input[[i]], getmode(crs$input[[i]]))
}

# Build the training/validation/testing datasets
# nobs=2331 training=1632 validation=350 testing=349

set.seed(crv$seed)

crs$nobs <- nrow(crs$input)

crs$train <- sample(crs$nobs, 0.7*crs$nobs)

crs$nobs %>%
  seq_len() %>%
  setdiff(crs$train) %>%
  sample(0.15*crs$nobs) ->
  crs$validate

crs$nobs %>%
  seq_len() %>%
  setdiff(crs$train) %>%
  setdiff(crs$validate) ->
  crs$test


crs$training <- as.data.frame(crs$input[crs$train,])
crs$validation <- as.data.frame(crs$input[crs$validate,])
crs$testing <- as.data.frame(crs$input[crs$test,])
```

Selection of the distribution.

Begin to model the response Losses (y) using the three parameter Zero adjusted gamma (ZAGA) distribution, followed by the Zero adjusted inverse gamma (ZAIG) distribution, and lastly using the four parameter Generalized beta type 2 (GB2) distribution. Start by fitting the full linear model μ including all explanatory variables.

```
mod1 <- gamlss(Losses ~ UpdatedDay + Desk + CapturedBy + TradeStatus + TraderId
     + Instrument + Reason + EventTypeCategoryLevel1 + BusinessLineLevel1 + Exposure,
     mu.start = NULL,  sigma.start = NULL, nu.start = NULL, tau.start = NULL,
     sigma.fo = ~1, nu.fo = ~1, data=crs$training,  family = ZAGA, n.cyc=80)
mod1

mod2 <- gamlss(Losses ~ UpdatedDay + Desk + CapturedBy + TradeStatus + TraderId
     + Instrument + Reason + EventTypeCategoryLevel1+ BusinessLineLevel1 + Exposure,
     mu.start = NULL,  sigma.start = NULL, nu.start = NULL, tau.start = NULL,
     sigma.fo = ~1, nu.fo = ~1, data=crs$training,  family = ZAIG, n.cyc=80)
mod2

mod3 <- gamlss(Losses ~ UpdatedDay + Desk + CapturedBy + TradeStatus + TraderId
     + Instrument + Reason + EventTypeCategoryLevel1 + BusinessLineLevel1 + Exposure,
     mu.start = NULL,  sigma.start = NULL, nu.start = NULL, tau.start = NULL,
     sigma.fo = ~1, nu.fo = ~1, tau.fo = ~1, data=crs$training,  family = GB2, n.cyc=170
mod3

GAIC(mod1, mod2, mod3)
```

Selection of terms.

Now we use ***drop1()*** to check whether any linear terms can be dropped and then using ***add1()***, we consider adding a two-way interaction term into the linear model mod1

```
drop1(mod3)
add1(mod3, scope=~(UpdatedDay + Desk + CapturedBy + TradeStatus + TraderId +
Instrument + Reason + EventTypeCategoryLevel1 + BusinessLineLevel1 + Exposure)^2)
```

We use ***FORM*** as an upper bound for ***scope***, starting from all explanatory terms sos that all interactions are considered

```
FORM <- as.formula("~(UpdatedDay + Desk + CapturedBy + TradeStatus + TraderId
+ Instrument + Reason + EventTypeCategoryLevel1 + BusinessLineLevel1 + Exposure)^2
+pb(UpdatedDay) + pb(Desk) + pb(CapturedBy) + pb(TradeStatus)+ pb(TradeStatus)
+ pb(TraderId) + pb(Instrument) + pb(Reason)+ pb(EventTypeCategoryLevel1)
+ pb(BusinessLineLevel1) + pb(Exposure)")
```

The estimation of some GAMLSS for OpRisk loss severity distribution: To build the model we pass on to the model building function **`gamlss`** *i.e., the formula that describes the model to build.*

```
mod14 <- stepGAICAll.A(mod1, scope=list(lower=~1, upper=FORM), k=log(371))

GAIC(mod14,mod24,mod34)
mod14
plot(mod14)
mod14$anova
```

```
Generalized Akaike Information Criteria (GAIC)

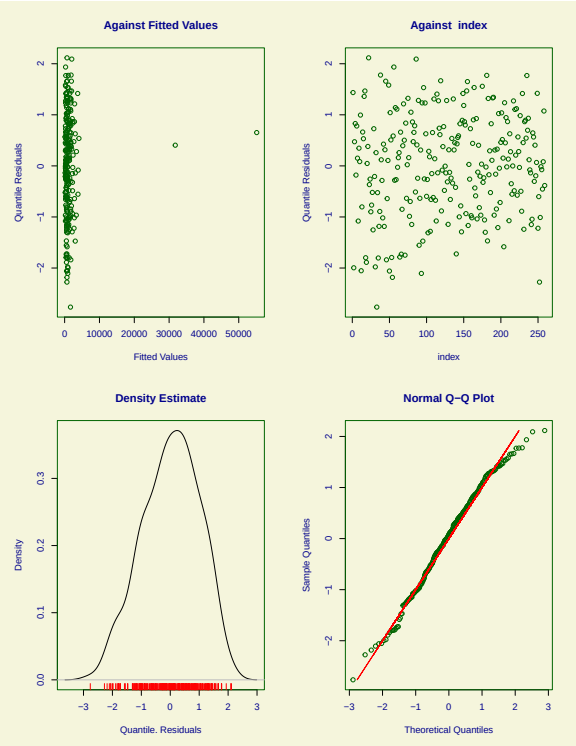
                df          AIC
GB2      22.27608      5955.438
ZAIG     13.00028      5964.959
ZAGA     25.13312      6066.610


GB2 summary statistics
*****************************************************************
             Summary of the Quantile Residuals
                         mean   =  0.02081528
                     variance   =  0.9648917
             coef. of skewness  =  -0.2544155
             coef. of kurtosis  =  2.485503
Filliben correlation coefficient  =  0.9945067
*****************************************************************

ZAIG summary statistics
*****************************************************************
         Summary of the Randomised Quantile Residuals
                         mean   =  0.03217758
                     variance   =  1.058743
             coef. of skewness  =  0.01371644
             coef. of kurtosis  =  2.528278
Filliben correlation coefficient  =  0.9975217
*****************************************************************

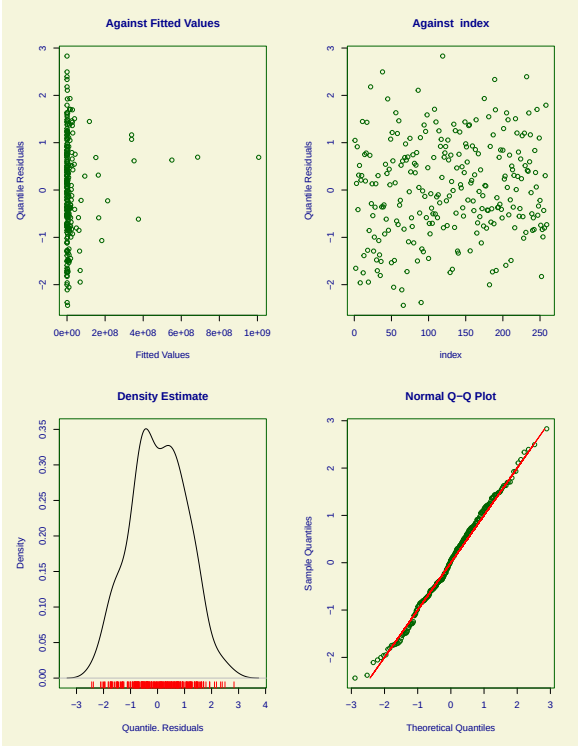
ZAGA summary statistics
*****************************************************************
         Summary of the Randomised Quantile Residuals
                         mean   =  -0.03874095
                     variance   =  0.9474831
             coef. of skewness  =  0.8438338
             coef. of kurtosis  =  3.773944
Filliben correlation coefficient  =  0.9780408
*****************************************************************
```

(a)

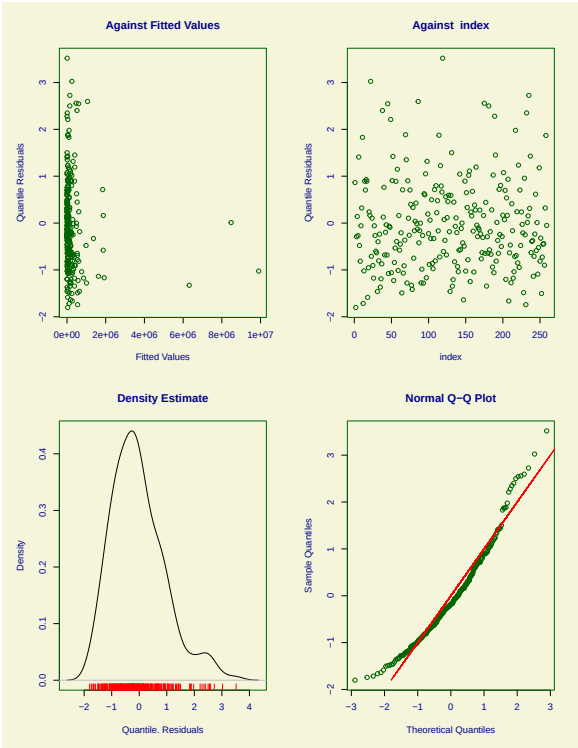


(b)

Figure 8.8
(a)Summary statistics of quantile residuals from models GB2, ZAIG & ZAGA (b)Displays (normalized quantile) residuals from model GB2(μ, σ, ν, τ).



(a)



(b)

Figure 8.9
(a) Display of (normalized quantile) residuals from model ZAIG(μ, σ, ν) (b) As for (a) but on the model ZAGA(μ, σ, ν)

Appendix D: R Code for Chapter 6

Required: R Packages from CRAN (in addition to packages already found in Chapter 3, 4, 5)

```
if (!require(gamlss)){
  install.packages("gamlss")
  library(gamlss)
}
```

Evaluate Poisson rate

Predicting validation set results and estimating the mean.

```
 Av.PredTT <- predict(Amodel, crs$testing, type = "response")

MASS::fitdistr(Av.Pred_Validation, "Poisson")
```

Data augmentation code: Extrapolation 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


% SOME OMITTED LINES




%% 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')
```

Deploying the GLM and GAMLSS Models

Import extrapolated data (1 year data), use the estimated model `Amodel` to predict forecasted estimates

```
 newfname <- "file:///C:/Users/Mphekeleli/Documents/R PROJECT/OpRiskPHDGitHub
/OpRisk_PHD_Thesis/Data/Extrap_Data_Model.csv"

 crs$newdataset <- read.csv(newfname,
                sep=",",
                dec=".",
                na.strings=c(".", "NA", "", "?"),
```

```
                strip.white=TRUE, encoding="UTF-8",header=T)

 head(crs$newdataset)

 Forecast <- predict(Amodel, crs$newdataset, type = "link")
 Forecast
```

Deploying the estimated GLM frequency model

Begin to deploy the estimated model for the response (the target variables y) LossesIndicator, generating new pairs $(\mathbf{x}, y)$ by transforming the $\mathbf{x}$ inputs using the augmented dataset *Augmentation_Dataset_Model.csv*. Fit the full GLM model μ including all explanatory variables.

```
# Remap variables.

# Transform into a factor.

crs$dataset[["TFC_LossIndicator"]] <- as.factor(crs$dataset[["LossIndicator"]])

ol <- levels(crs$dataset[["TFC_LossIndicator"]])
lol <- length(ol)
nl <- c(sprintf("[%s,%s]", ol[1], ol[1]), sprintf("(%s,%s]", ol[-lol], ol[-1]))
levels(crs$dataset[["TFC_LossIndicator"]]) <- nl

#class(exposure)
#length(exposure)

crs$dataset <- as.data.frame(crs$dataset)

# The following varaible selections have been noted

crs$input <- crs$dataset %>%
  group_by(UpdatedDay,
           UpdatedTime,
           TradedDay,
           TradedTime,
           Desk,
           CapturedBy,
           TradeStatus,
           TraderId,
           Instrument,
           Reason,
           EventTypeCategoryLevel1,
           BusinessLineLevel1) %>%
  transmute(LossesIndicator = TFC_LossIndicator,
```

```
            Losses = Loss,
            Exposure = exposure)

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

for (i in 5:(ncol(crs$input) - 3)){
     crs$input[[i]] <- relevel(crs$input[[i]], getmode(crs$input[[i]]))
}

crs$target  <- "LossesIndicator"
crs$risk    <- NULL
crs$ident   <- NULL
crs$ignore  <- c("UpdateTime", "TradeTime", "Nominal", "FloatRef", "LastResetDate", "Las
crs$weights <- NULL

# Build the training/validation/testing datasets
# nobs=2331 training=1632 validation=350 testing=349

set.seed(crv$seed)

crs$nobs <- nrow(crs$input)

crs$train <- sample(crs$nobs, 0.7*crs$nobs)

crs$nobs %>%
  seq_len() %>%
  setdiff(crs$train) %>%
  sample(0.15*crs$nobs) ->
  crs$validate

crs$nobs %>%
  seq_len() %>%
  setdiff(crs$train) %>%
  setdiff(crs$validate) ->
  crs$test


crs$training <- as.data.frame(crs$input[crs$train,])
crs$validation <- as.data.frame(crs$input[crs$validate,])
crs$testing <- as.data.frame(crs$input[crs$test,])

## Import extrapolated data nine months into the horizon, use the estimated model Amodel

newfname <- "file:///C:/Users/user/Documents/PHD/OPRISK_PHD_DISS/Data/Extrap_Data_Model.
```

```
crs$newdataset <- read.csv(newfname,
               sep=",",
               dec=".",
               na.strings=c(".", "NA", "", "?"),
               strip.white=TRUE, encoding="UTF-8",header=T)

head(crs$newdataset)

### Deploying the R model
## Predict forecasted estimates

Forecast <- predict(Amodel, crs$newdataset, type = "link")
Forecast

# Export into excel

library(R2HTML)
HTMLStart(); HTML(data.frame(Forecast)); w <- HTMLStop()
browseURL(w)

shell(paste("start excel", w))
```

Deploying the estimated GAMLSS severity model

Begin to deploy the estimated "best model "for the response (the target variables y) Losses, generating new pairs $(\mathbf{x}, y)$ by transforming the $\mathbf{x}$ inputs using the modified augmented dataset *Augmentation_Dataset_Model.csv*. Fit the full GAMLSS model μ, σ, ν, τ including all explanatory variables.

```
# ## Imported extrapolated data nine months into the horizon, after use the estimated mo

PREDfname <- "file:///C:/Users/user/Documents/PHD/OPRISK_PHD_DISS/Data/Results/Actual_Mo

crs$NEWDATAset <- read.csv(PREDfname,
                           sep=";",
                           dec=",",
                           na.strings=c(".", "NA", "", "?"),
                           strip.white=TRUE, encoding="UTF-8",header=T)

head(crs$NEWDATAset)

# ## Deploying the R model
# ## Use the chosen model mod34 for Loss to predict all the parameters mu, sigma, tau, a
# ## variables in newdata
```

```
PREDICTED <- predictAll(mod34, newdata = crs$NEWDATAset, type = "response")
PREDICTED

# # Export into excel

library(R2HTML)
HTMLStart(); HTML(data.frame(PREDICTED)); w <- HTMLStop()
browseURL(w)

shell(paste("start excel", w))
```

Experimental setup for using R OpVaR package

Begin to deploy the estimated "best model "for the response (the target variables y) Losses, generating new pairs $(\mathbf{x}, y)$ by transforming the $\mathbf{x}$ inputs using the modified augmented dataset *Augmentation_Dataset_Model.csv*.

- Fit the full GAMLSS model μ, σ, ν, τ including all explanatory variables. Key element of OpVaR is the opriskmodel structure.
- Inspired by the regulatory 8x7 business line/event type matrix, the opriskmodel structure is a list object where the single elements correspond e.g., to the cells of a business line/event type matrix. Each element comprises a loss frequency model, a loss severity model and if specified a dependency model between the frequencies and severities.
- Our lossdat dataset provides a minimal (2x2) example. As for the opriskmodel structure, the loss data needs to be stored in a list of cells with each cell comprising a data frame of loss severities and an integer time period assignment (e.g., years or quarters).

```
library(dplyr)
library(lubridate)

LossDataSet <- c(list(data.frame(Loss=c(887,624,509),
               Period = c(12,19,73),
               UpdateTime = as.Date(c('2013-01-17','2013-01-28','2013-04-17')))),

               list(data.frame(Loss=c(4311,549,2978,775,378,7086,182,775,166,966,919,233
               Period = c(9,18,18,26,27,33,35,40,42,43,48,51),
               UpdateTime = as.Date(c('2013-02-28','2013-05-02','2013-05-02','2013-06-26
                                      '2013-07-04','2013-08-12','2013-08-28','2013-10-03
                                      '2013-10-17','2013-10-23',
                                      '2013-11-26','2013-12-17')))),

# Some omitted lines
```

```
list(data.frame(Loss=c(220,821,1243),
Period = c(5,28,44),
UpdateTime = as.Date(c('2013-02-01','2013-07-10','2013-10-31')))),

list(data.frame(Loss=c(467,2864,657),
Period = c(27,31,46),
UpdateTime = as.Date(c('2013-07-01','2013-07-31','2013-11-11')))))
```

Appendix E: R Code for Chapter 7

Required: R Packages from CRAN (in addition to packages already found in Chapter 3, 4, 5, 6)

```
if (!require(gamlss)){
  install.packages("gamlss")
  library(gamlss)
}
```

Results: Rattle program code

Predicting validation set results and estimating the mean.

```
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
```