

**Quantitative response to the operational risk
problems of external data scaling and dependence
structure optimization for capital diversification**

WN Reynecke

 **orcid.org 0000-0003-2883-9046**

Thesis submitted in fulfilment of the requirements for the degree
Doctor of Philosophy in Risk Analysis at the North-West
University

Promoter: Prof DCJ de Jongh

Co-promoter: Prof H Raubenheimer

Graduation May 2018

13032585

Acknowledgements

Utmost gratitude to my study supervisors, profs. Dawie de Jongh and Helgard Raubenheimer, for their assistance, tolerance and support.

A special word of thanks to JvZ, without whom this study would never have been possible.

Abstract

In this study we aim to provide a quantitative response to the two respective operational risk problems of i) external data scaling, and ii) dependence structure optimization for capital diversification. The study is hosted at a financial institute in South Africa which utilizes the Advanced Measurement Approach (AMA) to calculate capital requirements for operational risk.

For Problem I on external data scaling, our study tracks the usage of power law transformation in order to gauge the proportional effects of operational risk losses. We consider an extended technique incorporating a ratio-type scaling technique originating from the basic power law transformation study. Our proposed solution then diverts to quantile regression. We apply said theory to the regression problem of compiling a ratio-type scaling mechanism.

We conclude our study on Problem I by providing a scaling mechanism for scaling down internationally-sourced external loss data to South African-based internal loss data allowing direct combination in a pooled loss dataset. Using this loss dataset, we consider an impact study where we note an increase in undiversified capital estimates of approximately 9% when comparing to internal loss data only.

For Problem II on dependence structure optimization our study considers the utilization of copulas to express the dependence structure of operational risk losses over time. We specifically investigate the application of factor copulas as derived from (exploratory) factor analysis. Using factor-based copulas allows for significant reduction in the dimensions of dependence structures which is a major problem when considering the high dimensionality associated with operational risk categories (ORCs). Our proposed solution enables us to construct dependence structures for a large number of ORCs using only two factors. We build the study around elliptical copulas and investigate dependence structure and diversification benefits when adjusting for the presence and magnitude of tail dependence.

We conclude our study on Problem II by providing a two-step method for constructing a dependence structure via factor analysis and then using this low-dimensional dependence structure result to easily construct a high-dimensional copula from which we simulate for capital estimates. Using our scaled and pooled dataset from Problem I, we obtain results confirming the general range of between

approximately 30% and 50% for reduction in VaR. However, when extending a two-factor Gaussian copula to a two-factor (Student's) t -copula for more conservative capital estimation, we clearly note how tailedness affects capital estimates when examining for expected loss and VaR estimates of various t -copulas.

Keywords: *operational risk, scaling, external data, internal data, quantile regression, dependence structure, factor copula, high dimensions, capital estimation*

Afrikaanse opsomming

In hierdie studie poog ons om 'n kwantitatiewe respons te gee tot die twee onderskeie operasionele risikoprobleme: i) eksterne dataskalering, en ii) die optimisering van afhanklikheidsstrukture vir kapitaaldiversifikasie. Hierdie studie is gebaseer by 'n finansiële instituut in Suid-Afrika wat gebruik maak van die 'Advanced Measurement Approach (AMA)' - Gevorderde Metingsbenadering om die vereiste kapitaal vir operasionele risiko te beraam.

Met probleem I rakende eksterne dataskalering, volg ons studie die gebruik van magswettransformasie soos gebruik om die proporsionele effek van operasionele risiko te bepaal. Ons bestudeer 'n uitgebreide tegniek wat gebruik maak van verhoudingsgewys-skalering met sy oorsprong in die basiese magswettransformasie. Ons aanbevole oplossing wyk dan af na kwantielregressie. Ons pas hierdie teorie toe op die regressieprobleem om 'n verhoudingsgewys-skaleringsmeganisme te konstrueer.

Ons sluit ons studie van probleem I af deur 'n skaleringsmeganisme aan te beveel wat eksterne verliesdata, soos van die buiteland ontgin, afskaal tot Suid-Afrikaansgebaseerde interne verliesdata wat direkte vermenging tot 'n saamgevoegde datastel moontlik maak. Wanneer ons hierdie gemengde datastel gebruik in 'n impakstudie, neem ons 'n toename waar in ongediversifiseerde kapitaal van ongeveer 9% wanneer dit met 'n soortgelyke kapitaalberaming vergelyk word van slegs interne verliesdata.

Vir probleem II aangaande die optimisering van afhanklikheidsstrukture vir kapitaaldiversifikasie, bestudeer ons studie die gebruik van kopulas om die afhanklikheidsstrukture van operasionele verliese oor tyd uit te druk. Ons kyk spesifiek na die toepassing van faktorkopulas soos afgelei vanuit faktoranalise. Die gebruik van faktorgebaseerde kopulas gee aanleiding tot aansienlike vermindering in die dimensies van afhanklikheidsstrukture wat 'n beduidende probleem is gegewe die hoë dimensionaliteit soos verbonde aan operasionele risikokategorieë. Ons aanbevole oplossing gee die moontlikheid om afhanklikheidsstrukture te konstrueer vir 'n groot aantal operasionele risikokategorieë deur slegs van twee faktore te gebruik. Ons bou ons studie rondom elliptiese kopulas en ondersoek afhanklikheidsstrukture en diversifikasievoordele soos ons aanpassings maak vir die teenwoordigheid asook grootte van hoë-waardeafhanklikheid - 'tail dependence'.

Ons sluit ons studie van probleem II af met 'n tweeledige metode om afhanklikheidsstrukture deur middel van faktoranalise saam te stel en gebruik dan hierdie resultaat van die lae-dimensionele afhanklikheidsstrukture om sodoende moeiteloos 'n hoë-dimensionele kopula te konstrueer waaruit ons kapitaalberamings kan simuleer. Met die gebruik van die geskaleerde, gemengde datastel van probleem I bevestig ons die verlaging in *Waarde-op-Risiko* (Value-at-Risk op Engels) binne die industrie-algemene omvang van tussen 30% en 50%. Alhoewel, met die oorgang vanaf 'n twee-faktor Gaussiese kopula na 'n twee-faktor t -kopula vir meer konserwatiewe kapitaalberaming, merk ons duidelik op hoe beïnvloed hoë-waardeafhanklikheid die kapitaalberaming wanneer ons vergelykings tref ten opsigte van verwagte verliese (expected loss) en *Waarde-op-Risiko* vir verskillende t -kopulas.

Sleutelwoorde: *operasionele risiko, skalering, eksterne data, interne data, kwantielregressie, afhanklikheidsstrukture, faktorkopula, hoë dimensies, kapitaalberaming*

Contents

Acknowledgements	i
Abstract	ii
Afrikaanse opsomming	iv
Glossary	xiii
1 Overview of the study	2
1.1 Problem statement and substantiation	2
1.2 Research aims and objectives	4
1.3 Methods of investigation	5
1.4 Chapter division	5
2 Operational risk modelling practices	7
2.1 Introduction	7
2.2 The Basel II Capital Accord	7
2.3 The four types of data	8
2.4 Loss Distribution Approach (LDA)	11
2.5 Chapter conclusion	12
I Problem I - External data scaling with quantile regression	13
3 Review of academic literature	14
3.1 Introduction	14
3.2 The scaling concept	15
3.3 Na study on data scaling	18
3.4 A multivariate take on data scaling	22
3.5 Chapter conclusion	24
4 Advanced literature topics	26
4.1 Introduction	26
4.2 The theory and application of quantile regression	26
4.3 Practical overview with graphical diagnostics	28
4.4 Chapter conclusion	31

5	The data environment	32
5.1	Introduction	32
5.2	Data overview	32
5.3	Data preparations	34
5.4	Data analysis	36
5.5	Chapter conclusion	53
6	Proposed solution	57
6.1	Introduction	57
6.2	Literature in practice	58
6.3	The scaling mechanism	73
6.4	Impact analysis	78
6.5	Chapter conclusion	85
7	Conclusion on external data scaling	87
7.1	Literature study	87
7.2	The data environment	88
7.3	Proposed solution	89
II	Problem II - Dependence optimization with factor copulas	91
8	Review of academic literature	92
8.1	Introduction	92
8.2	Background of dependence structures	92
8.3	A core theory of copulas	93
8.4	Copula families and their attributes	95
8.5	Goodness-of-fit techniques	99
8.6	Dependence: definition and measurement	101
8.7	Chapter conclusion	104
9	Advanced literature topics	106
9.1	Introduction	106
9.2	Conditional probability	107
9.3	The one-factor Gaussian copula	108
9.4	A two-factor Gaussian copula model	109
9.5	The two-factor t -copula	110
9.6	Chapter conclusion	112
10	Proposed solution	114
10.1	Introduction	114
10.2	The data environment	114
10.3	Two-factor copulas	116
10.4	Graphical analysis of a bivariate case	126
10.5	Impact analysis	128
10.6	Chapter conclusion	132

11 Conclusion on dependence optimization	133
11.1 Literature study	133
11.2 Proposed solution	135
 III Conclusion	 137
12 Overall summary	138
12.1 Quantitative response to two operational risk problems	138
12.2 Future perspectives	140
 IV Appendices	 144
A	145
A.1 Results of ELD regression modelling process	145
B	151
B.1 Proof of deriving Kendall's τ from copula function	151
B.2 Proof that tail indices are nonparametric and only depend on the underlying variables' copula(s)	152
B.3 Proof that $\frac{\partial}{\partial u}C(u, v)$ and $\frac{\partial}{\partial v}C(u, v)$ exist for almost all u and v	152
C	154
C.1 Results of factor analysis	154
C.2 Factor dependence structures	155
 References	 159

List of Figures

4.1	Cumulative distribution functions vs. quantile functions	27
4.2	Quantile regression in homoskedastic and heteroskedastic environments	29
4.3	An example of quantile regression for individual explanatory variables	30
5.1	Loss density distributions of Internal Loss Data (ILD) vs. External Loss Data (ELD)	36
5.2	Tail of the ILD distribution vs. ELD distribution	37
5.3	Bar charts of the factor variables - ILD number of observations	40
5.4	Bar charts of the factor variables - ILD loss amounts	40
5.5	Analysis of missing values for ILD independent variables	41
5.6	Bar charts of the factor variables - ELD number of observations and loss amounts	43
5.7	Analysis of missing values for ELD independent variables	44
5.8	Correlograms of the ILD numeric proxies	46
5.9	Correlograms of the ELD numeric proxies	48
5.10	Distributional forms of the Business Lines variable	50
5.11	Distributional forms of the Event Types variable	52
5.12	Distributional forms of the Regions variable	54
6.1	Quantile regression on the ILD models' terms	66
6.2	Quantile regression on the ELD model's terms	72
6.3	ELD - actuals vs. predicted	75
6.4	ILD - actuals vs. predicted	76
6.5	Boxplots per quantile of component B's ranges	78
6.6	Densities of the scaled data vs. those of the ILD and ELD	79
6.7	Tail densities of the ILD vs. full pooled dataset	80
8.1	Defining probability on a plane	94
8.2	Illustration of the minimum (W), product (Π), and maximum (M) copulas on the unit square	96
9.1	Illustration of elliptical copulas on spread and tailedness: Gaussian vs. t -copulas	111
10.1	Correlation structure of the pooled dataset	116
10.2	Scree plot and proportion of variance for factor analysis	119
10.3	qq-plots of the fitting of the Gaussian and three t_ν -copulas	123
10.4	Measures of association for the t_ν -copulas	125

10.5 3d-scatterplots of the copula densities	127
10.6 Contour plots of the copula densities	129

List of Tables

4.1	Illustration of quantile regression for homoskedastic and heteroskedastic environments	29
5.1	Guide to ORCs	34
5.2	Listing of filtering rules for ILD	35
5.3	Listing of filtering rules for ELD	35
5.4	Summary statistics of the numeric variables in the ILD - (* in USD '000)	39
5.5	Summary statistics of the factor variables - ILD number and amount of losses	39
5.6	Summary statistics of the numeric variables in the ELD - (* in USD millions)	42
5.7	Summary statistics of the factor variables - ELD number of losses . .	42
5.8	Summary statistics of the factor variables - ELD loss amounts	44
5.9	ILD Kendall's τ rank correlation	46
5.10	ILD Spearman's ρ rank correlation	47
5.11	ILD Kendall's τ rank correlation hypothesis test	47
5.12	ELD Kendall's τ rank correlation	48
5.13	ELD Spearman's ρ rank correlation	49
5.14	ELD Kendall's τ rank correlation hypothesis test	49
6.1	Exploratory Ordinary Least Squares (OLS) regression	60
6.2	Reduced OLS regression for ILD	61
6.3	Quantile regression: reporting on deciles of ILD	63
6.4	Quantile regression: reporting on deciles of ILD	64
6.5	Final model for ILD	68
6.6	Reduced OLS regression for ELD	70
6.7	Final model for ELD	74
6.8	Quantiles across the different datasets	81
6.9	Function attributes across different datasets	81
6.10	Estimation of distribution parameters	83
6.11	Capital estimates for selected ORCs (in USD '000)	84
10.1	Proportion of variance - normal marginals transformation	118
10.2	Dependence generator parameter estimates	120
10.3	Factor dependence structure - Gaussian copula	122
10.4	Goodness-of-fit test results	124
10.5	Tail index vs. degrees of freedom	126

10.6	Capital and provisioning estimation - undiversified ('000)	130
10.7	Capital and provisioning estimation - copula diversification ('000) . . .	131
A.1	Exploratory OLS regression for ELD - part i	146
A.2	Exploratory OLS regression for ELD - part ii	147
A.3	Full OLS regression for ELD	148
A.4	Quantile regression across deciles for ELD	149
A.5	Quantile regression across high quantiles for ELD	150
C.1	Proportion of variance - t_3 -marginals transformation	154
C.2	Proportion of variance - t_4 -marginals transformation	154
C.3	Proportion of variance - t_5 -marginals transformation	155
C.4	Factor dependence structure - t_3 copula	156
C.5	Factor dependence structure - t_4 copula	157
C.6	Factor dependence structure - t_5 copula	158

Glossary

AMA Advanced Measurement Approach.

BCBS Basel Committee on Banking Supervision.

BEICFs Business Environment and Internal Financial Controls.

DD-model Regression scaling model by Dahan and Dionne (2010).

EL Expected Loss.

ELD External Loss Data.

ILD Internal Loss Data.

LDA Loss Distribution Approach.

OLS Ordinary Least Squares.

ORC Operational Risk Category.

ORX Operational Riskdata eXchange (Association).

SA Scenario Analysis.

SARB South African Reserve Bank.

SMA Standardized Measurement Approach.

VaR Value-at-Risk.

A quantitative operational risk study

Chapter 1

Overview of the study

1.1 Problem statement and substantiation

1.1.1 Two quantitative problems

Introduction

The Basel Committee on Banking Supervision (2006:144), as world leader on banking supervision, defines operational risk as ‘...*the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. This definition includes legal risk, but excludes strategic and reputational risk.*’

Our quantitative response presented in this study proposes practical techniques to be used in such operational risk environments. We consider the two specific problems of i) a high degree of scale divergence between internal and external loss data; and ii) developing tractable dependence structures in environments with high dimensionality.

We consider these problems in more detail below, along with the motivation for solving said problems.

Problem I

The scaling issue arises when data from sources which are noncomparable to the financial institution which is modelling the data, are used. The problem lays therein that magnitudes (i.e. operational loss severity) and especially incidence (i.e. operational loss frequency) may be completely unmatched. It is therefore necessary to scale the external data or scenario analysis data (where external data are used for scenario generation) to align general expectations and / or realized values for magnitudes and incidence.

Problem I is therefore defined as:

- *How can we scale external loss data in such a way as to allow for subsequent direct combination / pooling of internal and external loss data to model capital estimates which reflect a financial institute’s operational risk profile which is also informed by external operational loss tendencies?*

Problem II

An overview of dependence modelling is provided by Embrechts and Hofert (2011:190) who mention three methodologies applied for correlation modelling: namely those of copulas, correlation matrices and other *ad hoc* techniques. Significant research has been completed on the use of copulas to analyze, quantify and apply correlation (dependence) issues in the underlying data as well as possible optimization when choosing between copulas.

The question of such dependence modelling approaches is identified as the second problem. This problem may be subject to the following complications: firstly, there is no consistency for correlation calculation and application practices, and secondly, there is no consistency in the modelling practices either.

Problem II is therefore defined as:

- *How can we create an optimal dependence structure which facilitates the high-dimensional dependence analysis associated with the taxonomy of operational risk loss data, whilst still allowing for undemanding diversified capital estimation using our scaled and pooled dataset?*

Motivation

Motivation for the intended study is provided here respectively in relation to the two problems identified.

With respect to Problem I, the successful analysis and conclusion of the scaling problem may incorporate a deleterious aspect in terms of the over-complicating nature of dependence studies by easing the model flow. However, the main focus of data scaling is not modelling facilitation, but rather the accurate representation of the underlying factors of external data (Basel Committee on Banking Supervision, 2006:152). Using external data such as ORX Association¹ data leads to highly unmatched sets. Since much external loss data derives from institutions based in North America and Europe not only exchange rates but actual nominal scales are not comparable to the South African environment.

Torresetti and Nordio (2014:59) obtained proof confirming that external data which are not appropriately adjusted prior to usage in the modelling process may increase systemic risk. Taking into account the position of the host financial institution within the limited South African banking environment, such a consideration is essential for capital calculation as well as governance issues which may arise with the SARB.

In conclusion, the benefits for the hosting (and any South African) financial institution are evident from the preceding argument. The possibility of greater efficiency and accuracy in scaling data will facilitate modelling practices, whilst still preparing a conservative operational risk profile referencing foreign loss experience.

Concerning Problem II, it is evident that a successful investigation and conclusion of the dependence procedure during the modelling phase may in fact complicate the modelling methodology with possible consequences expected even in terms of model governance; i.e. monitoring and validation. However, the positive impact given the negative complicating nature of such a conclusion is the possibility of a reduction

¹<https://managingrisktogether.orx.org/>

in capital held by the financial institution. This expectation is also confirmed by Giacometti et al. (2008:17).

Since operational risk capital is currently the second largest amount held by most banks after credit risk, and has minimal possibility of implementing a ‘run-down’ strategy unlike other risk types, it is imperative that alternative reductive strategies be investigated by the financial institution.

1.1.2 The operational risk environment

International environment

Operational risk modelling has seen dramatic changes, or at least, proposals since 2016 originating from its very source, Basel. A consultative document was issued early in 2016 explaining a new methodology - the Standardized Measurement Approach (SMA). For this methodology operational risk modelling would purely consist of a revision of the Basic Indicator Approach (see section 2.2.3), a proxy of the financial institute’s financial statements and bank-specific operational risk losses (Basel Committee on Banking Supervision, 2016:8). The motivation cited centres on the Basic Indicator approach adequately incorporating a proxy for business volume and being complementary to the financial statement information. Further BCBS adds that including bank-specific losses extends the SMA with great sensitivity.

The stakeholder response has been fairly mixed, but there is significant opposition to the proposal. This is mainly due to the expected large increases in capital estimates according to Osborn and Haritonova (2017). Later amendments to the proposed methodology that countries’ respective regulating bodies will be able to waive the bank-specific loss component in the calculation led to great dissatisfaction due to the playing field now being *unlevel* (Woodall, 2017). As explained, the basic reason for the SMA was to allow for greater comparability which would thus be defeated with such a concession.

Host environment

Given the international environment, we confirm that the SARB has not provided any formal opinion on the usage of the SMA. The financial institution which hosts this study has investigated and modelled a crossover to the SMA, but retains the Advanced Measurement Approach (AMA) for modelling operational risk capital for its primary business functions; both regulatory and economic capital. The host institution suggested that even if regulatory requirements were to favour the SMA, it would retain the AMA for economic capital modelling.

1.2 Research aims and objectives

The research objectives for this study are defined here in a generalized sense in terms of both of the aforementioned problems. The generalized objectives are proposed as the following:

- define the modelling and governance environment where the problem is rooted;

- determine the extent to which the financial institute is subject to the problems;
- investigate possible mitigation procedures; and
- validate successful mitigation procedures with an impact study.

1.3 Methods of investigation

Empirical research concerning Problem I, external data scaling, considered the environment at the financial institution at the time of the study as indicated by Chernobai et al. (2007:198) for the possibility of scaling incoming external data. This process took into account the practices of obtaining a loss distribution by investigating modelling practices surrounding the underlying frequency and severity distributions. The parameters of the frequency and severity distributions were considered along with the convolution methodology to obtain the final loss distribution. The extended use of the principle of change of measure as completed by Na (2004:71) was investigated for its inclusion of the Power-Law transform.

In addition, special attention was given to the possibility of applying location shift or location-scale shift models. These models, as analyzed by Cope and Labbi (2008:8) focus on the use of quantile regression. The resulting quantile regression provides a perspective in terms of whether the loss data need to be scaled for location only (i.e. a drift or progression in the mean of the loss distribution) or location and scale (i.e. the mean and variance of the loss distribution).

Problem II required empirical research which addresses specific issues relating to dependence modelling in operational risk modelling as found in the paper ‘Operational Risk – Supervisory Guidelines for Advanced Measurement Approaches’ Basel Committee on Banking Supervision (2011:8). An overview of the issues investigated includes i) testing dependence assumptions loss data; ii) ensuring sufficient levels of conservatism whilst incorporating technical correctness, integrity and appropriateness; iii) investigating the assumption that dependence within operational risk categories is zero – i.e. losses within categories are independent as opposed to the dependent relationship among categories; iv) performing impact analyses.

Our approach for solving the two operational risk problems is given more tangible structure in the following section where we provide the layout of our study.

1.4 Chapter division

Taking into account the objectives listed for this study, the subsequent listing provides a perspective on the chapters we included for each of the respective problems of the study in order of their performance:

- general literature study;
- a review of specialized academic literature;
- analysis of the data environment;
- proposed solution and an impact analysis aimed at risk capital estimation; and

- summary and conclusion.

After providing an overview of basic operational risk modelling techniques in Chapter 2, we follow the aforementioned chapter division in full for Problem I before restarting the division for Problem II. We take this approach since the methodologies are similar, and the results of Problem I are used in Problem II.

For both Problems I and II, we consider *basic literature* studies where we review concepts which we use in our proposed solution. In Chapters 3 and 8 respectively, we provide theoretical definitions, derivations and cross-referencing between sources to track our development of the two proposed solutions.

Subsequently we build on this general knowledge by considering *advanced literature* topics and their practical applications in detailed case studies in Chapters 4 (external data scaling) and 9 (dependence structure modelling). In the separate studies we evaluate the techniques used in the case studies which we then use as base structures for compiling our own proposed solutions.

We then discuss the data environments for Problems I and II, drawing on the (transformed) data of the host institution - Problem I - or the adapted dataset - Problem II. These analyses on the data environment provide us with a measurement as to the extent the problems are embedded in a real-world industry's operations and are given in Chapters 5 and 10 (Section 10.2).

Finally we provide the proposed solution to our two identified problems in Chapters 6 and 10. We discuss the theoretical groundwork and then focus on the practical aspects; challenges during modelling, intermediate mitigation steps and a final defined model in each case. Using the proposed solution we perform impact studies by performing capital estimation in the two distinct environments to gauge the effect of implementing our solutions.

We summarize our theoretical ideas, the proposed models, and capital results before ultimately reflecting on how our proposed solution solved our problems as delineated here.

Chapter 2

Operational risk modelling practices

2.1 Introduction

Since the basis of the study is a quantitative response to problems faced by the operational risk industry, it is necessary to consider the basic quantitative aspects of measuring operational risk. Hence, this chapter provides an overview of the quantitative practices when measuring operational risk leading to the ultimate result of regulatory and / or economic capital.

We first consult the guidance given as part of Basel II. This includes different modelling approaches adhering to Basel II principles. We then continue to reviewing the practices for operational risk (loss) data collection and usage. In addition, we study the ultimate goal of quantifying the measurement of operational risk to estimate a capital charge for the financial institution. Finally, we summarize the practices as noted in the aforementioned sections.

2.2 The Basel II Capital Accord

2.2.1 Nomenclature

In this study the Basel Capital Accord(s) forms a core directive in our approaches to address capital modelling. We therefore note that the terminology associated with the capital accord, the actual naming convention of ‘Basel II / 2.5 / III’ and its publishing entity – Basel Committee on Banking Supervision (BCBS) – is used interchangeably. The interchangeable use of the terminology thus refers to the documentation, the mandate it is endowed with by the BCBS, or the committee itself.

2.2.2 Pillar I and modelling approaches

In terms of the three pillars of the Basel II Capital Accord, Chernobai et al. (2007:38) point out that modelling operational risk forms part of Pillar I; i.e. the so-called ‘minimum capital requirements’. This is what the focus of our two quantitative studies

centres on. However, Panjer (2006:7) states that minimum capital requirements are significantly supported by both Pillars II and III. We therefore must note here that our focus did not directly consider effects of Pillars II and III, other than noting the explicit prerequisites imposed by our own supervisor, the South African Reserve Bank (SARB) (Pillar II – Supervisory review process), or calculating the final regulatory capital amounts (Pillar III – Market discipline).

Chernobai et al. (2007:40-47) discuss the three modelling techniques of the Basic Indicator Approach (BIA), the Standardized Approach (TSA), and the Advanced Measurement Approach (AMA) according to the guidelines presented by the Basel Committee on Banking Supervision (2006:144-151). As stated in Chapter 1, the focus of the study follows the process of the AMA model.

2.2.3 The three modelling approaches

As defined by the Basel Committee on Banking Supervision (2006:144-145) the BIA focuses on the financial institution's gross income as an indicator of its potential risk profile for operational risk losses. With the BIA the financial institution is required to hold a fixed percentage of the average of three years' gross income as operational risk capital.

For TSA, Basel Committee on Banking Supervision (2006:146-147) follows a similar approach to that of the BIA. However, the gross income element now receives a more granular treatment in that the financial institution's respective business lines are isolated and specific percentage charges are multiplied by the gross income for those individual business lines.

Basel Committee on Banking Supervision (2006) allows greater leeway for modelling practice when a financial institution uses the AMA. Subject to regulatory approval, the AMA permits the financial institution to directly analyze and model in great detail its individual operational risk profile. A significant input to the AMA modelling procedure is the four types of loss data used for operational risk modelling. In the following section we discuss these data types.

2.3 The four types of data

2.3.1 Introduction

Basel II identifies the following four data inputs for capital modelling (Basel Committee on Banking Supervision, 2006:152-154):

- Internal Loss Data (ILD);
- External Loss Data (ELD);
- Scenario Analysis (SA); and
- Business Environment and Internal Financial Controls (BEICFs).

Below we consider short overviews of these data types which are specifically relevant to our study. For further detailed reading on the topics, the reader is referred to the Basel II paper (Basel Committee on Banking Supervision, 2006).

2.3.2 Quantitative data - Internal and External Loss Data

Internal Loss Data (ILD)

Concerning ILD, we note from Basel II that individual financial institutions (banks in the case of our study) ought to track their internal losses within the AMA framework's guidelines. Clear guidelines are provided on the metadata¹ for this exercise and the financial institution should capture as much information as possible on operational losses.

Banks should capture inter alia the business lines, event types, possible credit loss impacts, and loss date information. For business lines and event types granularity is encouraged when capturing the loss data, but banks should at a minimum be able to map back to the categories supplied by the Accord (Basel Committee on Banking Supervision, 2006:302-307).

External Loss Data (ELD)

With ELD we confirm that the above-mentioned metadata is captured in the same manner. ELD may be sourced from public databases for free, or tracked from private sources paying a subscription fee. The discriminating element between ILD and ELD arises from the fact that ELD may not always contain the same characterizing metadata, and indeed whole data fields, as captured for ILD. This can be seen in the instance of the SAS[®] Global (SAS[®]-G) dataset.

The SAS[®]-G dataset is based on public data; i.e. only losses which had significant magnitude and therefore became known publicly. In the case of the Operational Riskdata eXchange (Association) (ORX), the data are submitted by participating financial institutions, made anonymous by ORX, and then sold as a pooled set back to the financial institutions. Our study focuses on how we can directly use such data sources with our own; i.e. how to use ILD and ELD pooled as a single loss dataset.

Operational Risk Category (ORC)

We noted in our review of ILD that Basel II requires the banks to capture data on inter alia the business lines and event types of the operational risk losses. For the purpose of internal taxonomy of such data Basel II provides definitions specifically for the above-mentioned metrics of business lines and event types (Basel Committee on Banking Supervision, 2006:302-307).

In this Basel II specification, we see eight possible classes for business lines and seven for event types. Given the possibility of this two-way classification we can compile an $m \times n$ matrix with a total of 56 cells. We can see that some banks may compile more granular taxonomy should the data be available to them, whilst others may opt to 'collapse' data classification across cells.

Link to Problem I

To recapitulate our problem setting in Chapter 1; the first focus of this study was to combine the two quantitative data elements of ILD and ELD. As illustrated in

¹Metadata refers to the classifying attributes of data which allows for more descriptive taxonomy of data without actually providing subset examples to explain the data characteristics.

Chapter 5, many of the metadata aspects of the respective ILD and ELD sets differ markedly. We therefore cannot combine the data sets directly and must construct a scaling method with which to adapt one of the sets to be aligned to the other whilst still retaining its inherent risk profile and loss behaviour.

2.3.3 Scenario Analysis

In our study we only note that Scenario Analysis (SA) consists of expert opinions which are sought by the central operational risk entity – e.g. a modelling team – within a financial institution. The expert opinion attempts to compile a new ‘dataset’ where business experience is forecasted in terms of operational risk losses. The output of this process is usually in the form of parameters of ‘an assumed statistical loss distribution’ (Basel Committee on Banking Supervision, 2006:154). This expert judging procedure will often be informed by loss data histories.

This expert opinion has naturally evolved over time in the respective business units and can therefore provide expert judgement on possible operational risk outcomes. As mentioned, we note that the expert opinion can be informed by both ILD and ELD as part of the consensus reaching process. The consensus is mostly reached by means of a customized form of the Delphi technique (Fletcher & Marchildon, 2014:2), however, the element of presenting anonymous summaries is often excluded.

2.3.4 Business Environment and Internal Financial Controls

With these data, Basel Committee on Banking Supervision (2006:302-307) encourages financial institutions to utilize the results of enterprise risk management in assessing its operational risk profile with the purpose of including such outcomes in the capital modelling. Such results may include an overview of key risk indicators for business units or alternatively, internal audit assessments of the business line in question.

A particular aspect of the Business Environment and Internal Financial Controls (BEICFs) data is that the data (or at least its modelling interpretation) should be forward-looking. This is also confirmed in the suggestion that the process should be tracked over time by means of a comparison between predicted and actual internal loss experience.

2.3.5 Overview

The four data types can be used in capital modelling to varying degrees of execution feasibility. In the case of BEICFs there is no clear guidance or formal precedent of its application. The most tangible example of using BEICFs in capital modelling, is how the expert opinions which construct scenarios for modelling, will indirectly be influenced by their respective business lines’ environment and internal controls. In contrast, SA can have direct quantitative interpretation with modelling application, albeit to a lesser degree.

As became apparent from our investigation of ILD and ELD, we can see that their direct quantitative nature makes modelling usage fairly easy. When considering ILD and ELD the field becomes more level; both data sources are mostly quantitative in nature and their attributes (metadata) are usually comparable. However, the scale

difference between ILD and ELD can be an inconvenient matter, as was pointed out in Chapter 1. We therefore note the sensitive nature of the scaling element between the two datasets and direct our focus on it whilst we consider other data input as *ceteris paribus*. We also note the basic grouping measure of quantitative data; i.e. the practice of categorizing data into ORCs.

2.4 Loss Distribution Approach (LDA)

2.4.1 Introduction

Once the data have been identified and cleaned, various options are available to model operational risk. Firstly, the frequency of losses is modelled by means of a discrete distribution. Subsequently we model the severity of our losses. The two distinct sets of results are then combined through a compound distribution to form an aggregate loss distribution.

The most common technique to obtain such an aggregate loss distribution, which we also use in our study, is that of the Loss Distribution Approach (LDA). The LDA originates from actuarial science where it was constructed to combine the results of the number of claims and the magnitude of these claims to calculate the overall effect on the portfolio of an insurer.

2.4.2 LDA

Overview of the theory

From our abovementioned data sources, we can empirically determine a frequency distribution to be used in our (aggregate) loss distribution. Similarly, we can compute our severity distribution and we are now ready to combine the two distributions. Note that pure parametric approximations are available and are often fitted to delineate the information of our operational risk profile.

Consulting Chernobai et al. (2007:223), we define our aggregate loss as

$$S = X_1 + X_2 + \cdots + X_N \quad (2.1)$$

$$= \sum_{i=1}^N X_i, \quad (2.2)$$

where

S = sum of the loss distribution;

N = number of losses; and

X_i = random loss amounts.

From the resulting aggregate loss S , we can now see where we are incorporating our expected number of losses (the frequency distribution – N), along with our expected magnitude of said losses (the severity distribution – X). When this aggregation process is repeated in large simulations (e.g. 1 million times), the resulting distribution

of the aggregate losses, S , forms our loss distribution. Very high quantiles of this distribution – Value-at-Risk (VaR) estimates – form the capital estimate which the bank is aiming to determine.

Link to Problem II

Our second problem lies within the context of the loss distribution approach. Since the bank has already identified the numerous ORCs where it incurs losses and these losses have now been aggregated to estimate high-end losses, the bank now needs to compile an overall figure for operational risk losses.

Merely summing the VaR estimates across all the ORCs results in an extremely large total capital which is not only be infeasible for the bank to hold, but may also be a poor reflection of the risk profile and loss behaviour. As explained in the motivation of Chapter 1, in Problem II we consider how losses across ORCs may be dependent and exhibit linked behaviour. The result of such a study implies that a pure summation of VaR estimates overestimates the possible losses and that dependence in loss behaviour needs to be taken into account.

2.5 Chapter conclusion

In this chapter we have considered the guidance provided by Basel II on operational risk. The four data elements of ILD, ELD, SA and BEICFs were briefly reviewed. The possible sources for the extensively quantitative ILD and ELD were seen, whilst we noted the process of sourcing the more qualitative SA and BEICFs from subjective internal processes.

We closed the chapter by studying the basics of the actuarial technique of the loss distribution approach. This process forms the backbone of most operational risk models and is also be the main element in our study's quantification process.

This chapter provided an overarching context for our problem and set the scene for where Problem I is located within the operational risk universe. In the subsequent chapter we provide the reader with the direct quantification tools which we utilize to solve the first problem.

Part I

Problem I - External data scaling with quantile regression

Chapter 3

Review of academic literature

3.1 Introduction

Our first review of academic literature attempts to develop background from which we are sufficiently informed to develop a theoretical approach. This theoretical approach is ultimately aimed at achieving our goal of constructing a scaling mechanism using quantile regression. Apart from the ‘quantile’ approach, the actual regression and scaling mechanism themselves, however, were developed from distinct techniques. It is said techniques which we delineate and review in their continuously transforming structures.

We firstly consider how power-law concepts assist us in developing the scaling mechanism. The power-law studies explain how such patterns are natural phenomena in financial and physical processes (Bouchaud, 2001) and therefore we intuitively expect to see comparable effects in our data. An extensive investigation into detecting and measuring this power-law effect (Na et al., 2006) assists in defining an integral part of our scaling mechanism by assuming that variation in operational losses are composed of common components universal to all financial entities, and idiosyncratic components unique to each financial entity.

Subsequently we review the regression process by considering the work of Dahlen and Dionne (2010) which builds on the earlier scaling mechanism work, but then significantly extends the regression process into a multivariate setup. In this work we see that many issues from previous studies are addressed in the process: i.e. the problem of left-censoring bias during capturing of ELD and the improvement of the coefficient of determination when more variables – which are inherently part of the operational risk world (e.g. business lines, event types) – are included in the regression.

For each of the aforementioned we respectively consider the theoretical background in detail and then provide a summary of results obtained by the various authors. The chapter concludes when we summarize the theories studied here to equip our further steps of empirical and impact analyses once we continue onto the quantile regression approach.

3.2 The scaling concept

3.2.1 Overview of data scaling studies

Our study of the scaling concept can be seen as a staggered approach of existing scaling methods¹. To track our study's path we firstly set a background scene by investigating how the current scaling themes developed from well-established power-law techniques. Here we note the power-law relationships may especially appropriate for operational risk studies since it seeks to define nonlinear relationships as seen in the tails of operational loss distributions.

Secondly, we review the groundwork for scaling theory and applications as undertaken by Shih et al. (2000). We investigate the nature of the scaling relationships and see how well univariate regression on firm size measures fare. Building on this, we consider the extensive work completed by Na et al. (2006). Further applications of the regression of operational losses on firm size are seen, whilst we also assess whether to build a study based on aggregate losses as opposed to considering frequency and severity in isolation.

Finally, we conclude the introductory literature study with a review of the data scaling work by Dahlen and Dionne (2010). Here we see further improvements in the coefficient of determination when modelling the scaling relationships, but we also see a significantly extended approach to the regression theory by including further explanatory variables.

An excellent summary is also provided by Galloppo and Previati (2014) of the various methods of combining ELD and ILD. Specifically the section on scaling methods (Galloppo & Previati, 2014:90) provides an overarching view of scaling techniques and the ground already covered to theoretically define and expand this technique, but also how to apply it successfully to real-world data.

3.2.2 Overview of power-law studies

Bouchaud (2001:105) provides a meaningful introduction to power laws as detected in the economic and financial world. As Bouchaud (2001) points out, we are specifically interested in the attribute of scale invariance for power laws. Notwithstanding their extensive use in the field of physics, there is direct mention of the Pareto distribution of wealth. Given a population size of N , Bouchaud (2001:106) asserts that the ratio of largest to median wealth growth can be expressed by $N^{1/\mu}$. When $\mu < 1$, we see a distribution where large wealth is concentrated with a few individuals, whereas $\mu > 1$ indicates a wider, more even distribution.

Kühn and Neu (2003:652) revisit the physics inspiration in their study of phase transitions and their analogy to operational risk loss. We see that so-called 'collective phenomena' may be the reason that the overarching result of a system (in our study – an incurred loss) is inherently resilient to the primary unique attributes (in our

¹An excellent summary of the various methods of combining ELD and ILD is also provided by GALLOPPO G. & PREVIATI, D. 2014. A review of methods for combining internal and external data. *Journal of Operational Risk*, 9:22. Specifically the section on scaling methods *ibid.* provides an overarching view of scaling techniques and the ground already covered to theoretically define and expand this technique, but also how to apply it successfully to real-world data.

study – what unique attributes of the loss may prove to be causal). We note the same observation here that a loss incidence (i.e. a phase transition in physics) often exhibits power law trends with scale invariance.

Further definition is provided by Clauset et al. (2009:662) where it is mentioned that power laws often hold true only on censored intervals; i.e. for some minimum value of the exponent’s base number. The definition then follows that the distribution’s *tail* follows a power law. This idea we also noted in Bouchaud (2001:106), where the Pareto wealth distribution is generalized for individual wealth, ‘in its asymptotic tail’

$$P(W) \simeq \frac{W_0^\mu}{W^{1+\mu}} , \quad (3.1)$$

where $W \gg W_0$ and μ is defined as the distribution decay as W increases. Na et al. (2006:3) also references this concept and we observe the same conclusion that the scaling practice holds true independent of underlying situation-specific information. We note however, that the aforementioned study focuses on a more tangible interpretation of the scaling concept which we investigate further in the next section.

In conclusion, we have seen two noteworthy aspects of power laws illustrated; namely those scale invariance, and the power law occurring in the tails of distribution. We required this context of power laws since the first and most major assumption made by Na (2004:64) in his scaling study, is that there exists ‘*a universal power-law relationship between the operational risk loss amount within a certain time period and an indicator of size $\mathcal{E}$ exposure towards operational risk within a certain time period of different financial institutions.*’

3.2.3 Shih prelude to scaling concept

Overview of the study

In a study conducted by Shih et al. (2000) the query is compiled whether there exists some relationship between the size of a financial entity and the size (‘magnitude’) of operational losses. We see that the most important findings of the study are that minimal information is captured by the size of an entity, and that the relationship between firm size and operational losses is not linear.

Constructing the investigation

The study started off by investigating what measure of an entity’s size reflect best with respect to the operational losses. Shih et al. (2000:1) firstly consider what measure to use for the size of a financial entity. Using a simple logarithmic correlation study, it was found that there is high correlations for the operational losses and three size measures; i.e. the firm’s gross income, (total) assets, and the number of staff. It was found that income has the highest correlation with operational losses. Furthermore, we specifically note that the result of the logarithmic transformation of variables provided better results than those of original-format variables.

Shih et al. (2000:1) constructed their relationship model such that

$$L = R^\alpha . F(\theta) , \quad (3.2)$$

where

L = actual loss amount;
 R = revenue amount (i.e. gross income); and
 α = scaling factor.

The term θ functions as the error term (i.e. all variation not explained by the independent variable). Taking the logarithm of the equation, an Ordinary Least Squares (OLS) regression can be performed. We follow the notation of Shih et al. (2000:2), rewriting the equation as

$$\ln L = \alpha \cdot \ln R + \ln F(\theta) \quad (3.3)$$

$$l = \alpha \cdot r + \beta + \varepsilon \quad (3.4)$$

where

$r = \ln R$;
 $\beta = E[\ln F(\theta)]$; and
 $\varepsilon = \ln F(\theta) - \beta$.

Details of the investigation

The regression results showed that the intercept β and scaling coefficient α are both statistically significant – the respective t-tests are 10.51 and 10.31. We note that the coefficient for the scaling factor is given as 0.151. As mentioned before, we see that minimal information is captured in the size measure in that the coefficient of determination obtained is very low – 5.4%.

As expected for our given data, Shih et al. (2000:2) acknowledge the presence of heteroskedasticity. This problem is then addressed by reworking the regression principles and using generalized least squares regression in the form of weighted least squares. The weighting technique is not expressly stated, but the results of the generalized least squares regression show that a larger amount of variation is explained by the independent variable – the coefficient of determination is now 9.1%. Also, the scaling factor coefficient now obtains a value of 0.232 with a prominent t-statistic of 24.86.

Conclusion of the study

The study concludes that the size measure has low explanation value in terms of operational losses. Shih et al. (2000:2) then suggest that the remaining variation could be derived from intrinsic disparity in risk environments among financial entities, but also qualitative elements; e.g. the BEICFs of a firm.

3.2.4 Summary of scaling introduction

In our overview of power-law concepts (Bouchaud, 2001), we saw that such relationships are often naturally occurring in natural processes, but specifically also in the financial world. We additionally noted that such relationships are most prominent when considering the extreme values of our outcome space; i.e. clear power-law relationships exist in the tail of distribution.

When combining this principle with the work done by (Shih et al., 2000), we can confirm that the power-law relationship is present when considering the scaling aspects between a firm's operational losses and said firm's size. In this instance we learned that the relationship is not linear, and that firm size when considered in isolation is insufficient to describe the aforementioned nonlinear relationship.

In the subsequent section we build further on the concepts of power-law relationships and the applicability of firm size as a proxy for expected operational losses (Na et al., 2006). We see comparable treatment of the loss equation postulated in this section, but find a different viewpoint of so-called 'intrinsic disparity of risk environments' when the loss environment is divided into unique ('idiosyncratic') and general ('common') components. The idiosyncratic component is assumed to capture information directly relating to the loss environment of a specific entity whereas the common component captures information relating to the general loss environment such as geo-political and macroeconomic aspects.

3.3 Na study on data scaling

Overview

The Na (2004) study examines the BCBS suggestion that ELD can be combined with ILD by means of scaling. We note that Na (2004:65) suspects that minimal usage of such a method would be necessary for his environment which is a large bank with core operations in the Netherlands, and therefore reporting operational risk losses at the ORX threshold of €20,000.

However, it is noted that despite comparable threshold reporting, ILD and ELD may still follow different distributions (Na et al., 2006:2). The sources consulted for this practice indeed confirm that such an assumption may incorporate major model risk when applied to the scaling mechanism (Frachot & Roncalli, 2002:1). In order to counter this element of unique attributes captured in ELD (each observation is bank-specific), Na makes his assumption of the universal power law which we now study in more detail.

3.3.1 Na's model and power law transform

Building on the Basel taxonomy of the different Business Lines (BL) within a bank, Na et al. (2006:4) asserts that each business line can be viewed as an independent financial entity. We then see that an aggregate loss is defined per BL. This loss, L_b , is assumed to be the result of some function $u(\cdot)$ of two components; entitled the 'common component' R^{com} , and the 'idiosyncratic component', r_b^{idio} , such that

$$L_b = u(r_b^{idio}, R^{com}), \quad (3.5)$$

where $b = 1, \dots, 8$ for the eight BLs.

Na et al. (2006) indicated that the identified common component, R^{com} , can be ascribed to general circumstances which are faced by any financial institution; e.g. all such institutions face operational challenges related to general external conditions: macroeconomic, geopolitical and cultural situations. These are stated to be stochastic in nature.

We note that r_b^{idio} , the idiosyncratic component is notated in terms of the b BLs; i.e. this element is unique to the respective BLs. This is then elaborated on by Na et al. (2006), who explains that this component is deterministic in nature and reflects particular characteristics of the BL in question – possible examples include the size and exposure of the BL.

Na et al. (2006:4) then makes the assumption that the total effect of the common and idiosyncratic components as indicated by the function $u(\cdot)$ can be decomposed into a combination of two distinct functions such that

$$L_b = u(r_b^{idio}, R^{com}) = g(r_b^{idio}) \cdot h(R^{com}) , \quad (3.6)$$

where $g(\cdot)$ and $h(\cdot)$ respectively represent functions expressing the idiosyncratic and common components. Na et al. (2006) reworks the function $g(\cdot)$ to a type of ‘scaling factor’, given by (s_b) . This function is further defined to reflect the power law we are expecting in the idiosyncratic component of the aggregate losses, by denoting the function as

$$L_b = (s_b)^\lambda \cdot h(R^{com}) , \quad (3.7)$$

where the parameter λ then indicates a so-called ‘universal exponent’. This implies that the number is expected to be equal for all BLs and therefore the expression can be re-written as

$$\frac{L_b}{(s_b)^\lambda} = h(R_{com}) = L_{st} , \quad (3.8)$$

which in turn can be simplified to

$$L_b \cdot (s_b)^{-\lambda} = L_{st} , \quad (3.9)$$

where L_{st} then equates to a *standard* aggregate loss for a business line of unitary size. Note that the equation is then solved for λ to obtain scaling mechanism – a multiplying factor. Based on this study we are thus assuming that the data, whether the original set or the re-scaled result, are solely derived from a single distribution – in this instance that of L_{st} .

3.3.2 Theoretical aspects of the scaling exercise

Na et al. (2006:5) provides a brief overview of the scaling technique apparent from the aforementioned equation where one stochastic variable is transformed into another. The technique is based on us knowing the probability density function of one of the variables and then changing the variable – which translates to a scaling mechanism.

Assuming that the probability density function of L_{st} is given by $f(l_{st})$, then the probability density function $f'(l_b)$ of L_b is expressed as:

$$f'(l_b) = f((s_b)^{-\lambda} \cdot l_b) \left| \frac{dl_{st}}{dl_b} \right| \quad (3.10)$$

$$= f((s_b)^{-\lambda} \cdot l_b) \times (s_b)^{-\lambda} \quad (3.11)$$

The practical application of the derived results is explained by Na et al. (2006) in terms of a uniform distribution. If $f(l_{st})$ follows a uniform distribution then $f(l_{st}) = 1$ where $l_{st} \in [0, 1]$. Similarly then, $f'(l_b)$ also follows a uniform distribution such that $f'(l_b) = (s_b)^{-\lambda}$ where $l_b \in [0, (s_b)^\lambda]$. Relying on standard practices for transforming stochastic variables, Na et al. (2006) indicates that the transformational expression as given in Equation 3.9, also holds true for the mean and standard deviation of the variable(s). This can thus be illustrated by:

$$\mu_{L_b} \cdot (s_b)^{-\lambda} = \mu_{L_{st}} \quad (3.12)$$

and

$$\sigma_{L_b} \cdot (s_b)^{-\lambda} = \sigma_{L_{st}} \quad (3.13)$$

Rewriting the expression and applying a logarithm throughout results in a purely linear equation where λ is now the gradient of a straight-line function. Here we see the derivation of this regression equation for the case of the mean of the distribution:

$$\ln(\mu_{L_b}) = \lambda \ln(s_b) + \ln(\mu_{L_{st}}) \quad (3.14)$$

We point this out as an important extension in this study (Na et al., 2006), since the empirical investigation undertaken centred on determining whether the λ coefficient was equal between the three respective solutions; i.e. for the mean, standard deviation and aggregate loss.

3.3.3 Study results

Background

As mentioned, the study was performed for three different elements of the distribution mean, standard deviation and aggregate loss. Separate regressions were performed for the ELD, the ILD and the combined (scaled) data. The results of the study are summarized by Na et al. (2006:10).

Overview of regression for aggregate losses

For the mean regression on ELD, the coefficient and intercept for the abovementioned equation were both statistically significant at 90%. For ILD regression comparable results were obtained; the intercept was not found to be statistically different from 0 and the remaining coefficient did not reject the null hypothesis at 90%. The combined data regression reflected that the coefficients are highly significant; both passing at significance level $\alpha = 5\%$.

The regression for the standard deviation of the distribution had a clear conclusion across all three datasets' regression; none of the intercepts were found to be statistically different from 0. The coefficients of the scaling factors, however, all passed at 90%; the scaling factor for the ELD and combined datasets passed at 95%.

Overview of regression for frequency of losses

As with the preceding regression, the mean regression for both ELD and combined data did not reject the null hypotheses for either the intercept or scaling coefficient at a significance level of $\alpha = 1\%$. This was not the case, however, for the ILD which rejected the null hypothesis completely.

The standard deviation regression provided the same results; the same pronounced rejection of the null hypothesis when regressing on ILD, but no rejection for ELD or combined data regression.

Overview of regression for severity of losses

When regression on the mean, the only null hypothesis not rejected was that of the intercept for ELD. All the other coefficients were rejected; i.e. ELD scaling coefficients, and intercepts and scaling coefficients for ILD and the combined dataset.

For the regression on the standard deviation, there was a full rejection of all null hypotheses for all the various datasets.

3.3.4 Conclusion

With the study by Na et al. (2006) we see that there is a clear scaling relationship between the losses experienced by different financial institutions. More specifically, we note that there is a possibility that this relationship can be expressed as a power law.

The unexpected part of the study results is that the severity of losses did not show a clearly scaling relationship between comparing entities. Intuitively, our basis for data scaling analysis would exclude a review of the frequency element (and therefore the aggregate losses) given that this element would be assumed to be more unique to the base scaling entity.

Furthermore, our argument can find basis in the recommended usage of ELD as laid out by Basel (Basel Committee on Banking Supervision, 2006:153):

“A bank’s operational risk measurement system must use relevant external data (either public data and / or pooled industry data), especially when there is reason to believe that the bank is exposed to infrequent, yet potentially severe, losses. These external data should include data on actual loss amounts, information on the scale of business operations where the event occurred, information on the causes and circumstances of the loss events, or other information that would help in assessing the relevance of the loss event for other banks.”

Reflecting on this citation from the Basel II core document, our view can centre on the fact the severe losses are expected to be infrequent. Therefore, an internal loss database which does not have a high frequency of such losses ought to be informed and expanded with more severe external losses.

We thus continue on the path set up by Shih et al. (2000) that a scaling relationship exists between the magnitude of expected operational losses a financial institute will experience and the size of said institute. We also note the confirmation in Na et al. (2006) of the existence of such a relationship and incorporate the theory provided on the idiosyncratic and common components that explain this loss relationship.

In the next section we review the expansion undertaken by Dahlen and Dionne (2010) where it becomes clear the size (indicator) of a financial entity remains a determining factor for operational losses. However, we now see that other aspects – inherent to operational risk – may provide more information on the scaling relationship as already proposed earlier in Shih et al. (2000).

3.4 A multivariate take on data scaling

3.4.1 Extending the loss relationship scaling

Introduction

The Dahlen and Dionne (2010) study builds further on the preceding work by extending the univariate regression to a multivariate setup. As noted in the foregoing studies, a firm's size modelled in isolation is insufficient to explain the variation seen in operational losses. In this study we see how other data elements which are inherent to the operational risk field – e.g. business line and event type as defined for ORCs – are incorporated in a multivariate regression.

We specifically observe the commentary on the biases associated with reported data; i.e. ELD as set up with qualifying criteria for loss capturing, or the publicly availability of ELD (Dahlen & Dionne, 2010:1486). It is explicitly noted here that there was no mention of it in Shih et al. (2000) and minor attention is given to it in Na et al. (2006). In this overview, we see how truncation² is utilized during the regression process to 'correct' for the biases which arise when data are collected.

We do not review the process for scaling frequency of operational losses. Continuing from our previous reasoning, our study only focuses on the scaling of operational loss severity. However, we provide some points on the frequency scaling with respect the treatment of censoring, since we are intending to use a concomitant censoring treatment in our suggested solution.

Modelling assumptions

Relating to the aforementioned issue on data collection bias, we also see significant awareness of the issue at hand when Dahlen and Dionne (2010:1486) state their assumptions for the model. These assumptions include publicly available loss information is considered to be correct and that each loss has the same probability to be

²We note the confusion in terminology here: i.e. the different scenarios defined by the respective terms of truncation and censoring. It is understood that for censoring the data points greater or less than a threshold value were excluded for statistical analysis for some reason. For truncation it is understood that the aforementioned data points were never observed / recorded in the first place for some reason; thus leading to their inherent exclusion. The distinction is somewhat vague as seen in the source EVERITT, B. S. & SKRONDAL, A. 2010. *The Cambridge Dictionary of Statistics*, New York, USA, Cambridge University Press. A further conceptual distinction is introduced by DAHEN, H. & DIONNE, G. 2010. Scaling models for the severity and frequency of external operational loss data. *Journal of Banking & Finance*, 34:13., where it is indicated that censoring is the instance where the dependent variable cannot be observed above / below a certain threshold level whilst the explanatory variables have a full set observations available at all levels. In our study the terms are used interchangeably, but with a more pronounced focus on the term 'censoring' to align to software titles.

captured in the loss dataset. In addition, it is assumed that there is no correlation between the magnitude of the loss and the probability of it being captured in the loss dataset.

In terms of the data collection bias and its effect on regression, it is mentioned that the collection threshold of \$1 million will naturally influence the regression results. Dahlen and Dionne (2010:1485) indicated that they would correct for it, but provided little further information on the censoring treatment applied. However, for the scaling of the loss frequencies, we see the application of so-called ‘zero-inflated count data models’; i.e. Dahlen and Dionne (2010:1492) use zero as their threshold point when regressing the frequency data by means of Poisson or negative binomial models.

The scaling model in theory

For their severity scaling model, Dahlen and Dionne (2010) note an assumption where loss is deconstructed into a variable and constant component, comparable to the idiosyncratic and common components of Na et al. (2006) - refer to Section 3.3.1. The previously established principle of using a nonlinear relationship is continued here.

The model of the loss, L , is defined as follows³:

$$L = Size^\alpha \cdot F(\omega, \theta) \cdot \psi , \quad (3.15)$$

where ψ functions as the constant component; i.e. independent of the respective financial entities’ unique attributes. The variable component is divided between the $Size$ variable with α being a scaling factor, and the function $F(\omega, \theta)$. This function specifies a set of variables θ , with variables specific to the financial entity being investigated, whilst ω is a vector signifying the set of variables’ scaling information.

In this analysis, as with previous ones, Dahlen and Dionne (2010:1487) use the logarithm of the equation to obtain the result

$$\ln L = \alpha \cdot \ln Size + \ln (F(\omega, \theta)) + \ln \psi . \quad (3.16)$$

Note the term $\ln (F(\omega, \theta))$ will be expanded since we are now seeing a multivariate treatment of the regression process where

$$\ln (F(\omega, \theta)) = \sum_j \beta_j BL_j + \sum_k \delta_k ET_k , \quad (3.17)$$

given that the factor BL refers to the business lines of a financial entity and β_j is the parameter up for estimation for the specific Business Line. The same principle applies to the factor variable ET with its parameter δ_k . Henceforth we refer to this equation as the ‘DD-model’ after the mentioned authors.

Note the term ET was originally defined as RT; i.e. the risk type. We changed this to ET for ‘event type’ – one of the core categorizations of ORCs. Subsequent to

³The notation used here is slightly amended from the original to align to that of NA, H. S., MIRANDA, L., VAN DEN BERG, J. & LEIPOLDT, M. 2006. Data scaling for operational risk modelling. *ERIM Report Series Reference No. ERS-2005-092-LIS* , as well as SHIH, J., SAMAD-KHAN, A. & MEDAPA, P. 2000. Is the size of an operational loss related to firm size? : OpRisk Advisory - Stamford Risk Analytics. - see **References**.

this step the expected statistical inference techniques are applied to select a final model. We do not review these points in detail, since we are following the same steps as in Section 6.2.4 when we construct our own extension of the DD-model.

Once this regression process has completed for ILD and ELD, Dahlen and Dionne (2010) define A as the observed loss in the ELD and B as the observed loss in the ILD. In the following ratio we are calculating a scaled external loss; i.e. the external severity is scaled to the proportions of the internal losses. Using the same principles as Na et al. (2006) that the common / constant component are inherently equal between losses A and B, we see derived from Equation 3.15 that

$$L_B = L_A \frac{Size_B^\alpha \ln(F(\omega, \theta_B))}{Size_A^\alpha \ln(F(\omega, \theta_A))} . \quad (3.18)$$

Results of the study

The regressions were tested various confidence levels; i.e. 90%, 95% and 99%. In the regression we see that the Size and BL variables mostly remain statistically significant with a confidence level of 99%. For the ET variables most are statistically significant with a confidence level of either 90% or 95%. We specifically note that the only BL and ET variables (in this instance their corresponding factor levels) not to be concluded as statistically significant, are those of BL level DPA (damage to physical assets) and ET level IF (internal fraud).

The final regression model with parameter estimates is not provided, nor is the associated coefficient of determination. However, we do see that during the initial stages of building up the regression model from the version suggested by Shih et al. (2000) up to full extended multivariate model, there is a marked increase in coefficient of determination. In this regression the original model by Shih et al. (2000) obtained an R^2 of 7%. With the full model regression (nonsignificant variables not yet eliminated), the R^2 had reached a value of 29.61%.

Overview of the study

From Equation 3.18 we can see that we now have a mechanism that scales an external loss to the proportions associated with a set of internal losses; i.e. once we apply the scaling mechanism across all our external losses, we can directly combine the resulting dataset with our ILD. This then leads to a pooled dataset which is ‘re-informed’ with external aspects thereby providing a possible solution to the Basel II proposal of using ELD in the calculation of the operational risk capital charge (Basel Committee on Banking Supervision, 2006:153).

Another noteworthy point is the degree to which the coefficient of determination has increased by including more variables. As reiterated by Dahlen and Dionne (2010), the extended model clearly explains a lot more variation in the dependent variable than the original starting point of the Shih et al. (2000) model did.

3.5 Chapter conclusion

In this chapter we introduced the basic academic literature for possible scaling methods. This chapter above all allowed us to the introductory mathematical

concepts of scaling data. We also considered what variables to include when regressing for operational losses. Other points we evaluated during regression, included data issues such as data capturing above thresholds; and heteroskedasticity in the data. When comparing the different scaling models we noted similarity in defining distinct common / constant vs. idiosyncratic / variable components in the data. When expanding the models to multivariate environments we saw the best results for the coefficient of determination.

We firstly reviewed power-law concepts and how these relate to patterns observed in physics and finance. After defining the basic mathematical form of the power-law technique we considered its usage in the study by Shih et al. (2000). We saw that gross income is often the most Dahlen and Dionne (2010) appropriate variable to use as an indicator of a financial entity's size and therefore also as proxy when regressing for operational losses. In this study we noted no mention was made of handling the data capturing bias due to very high threshold value of the dataset: \$1 million. However, we saw how the coefficient of determination can improve when treating for heteroskedasticity during the regression.

In the Na et al. (2006) study we see a unique treatment of the loss variable when assuming that its variation is due to 'common' components as present in all geopolitical and macroeconomic spheres along with an 'idiosyncratic' component which is unique to each financial institute from the operational losses originate. This study stated awareness of the data bias in terms of the threshold issues, but did not provide much detail on its treatment during regression. A unique perspective of this study was how the regressions were repeated for the aggregate losses, and the severity and frequency of losses respectively. The results concluded that severity regression may in fact be pointless – i.e. most variation is explained by the loss frequency. However, as intuitively drawn from the study by Shih et al. (2000) and explicitly seen in the study Dahlen and Dionne (2010), there is much regression value to be obtained in regressing the loss severity.

Dahlen and Dionne (2010) continued with the concept of common and idiosyncratic elements of losses, renaming them as the constant and variable attributes of losses. Their study significantly extended the regression process by including variables inherently associated with operational risk; i.e. the business lines and event types as factor variables, culminating the in DD-model. An earlier working paper even ventured out to including the region variable Dahlen and Dionne (2008). The multivariate extension led to much higher coefficients of determination. We also noted said authors' awareness and direct treatment of bias in terms of censoring thresholds.

In overview we have learned that multivariate regression using inherent operational risk variables provide the best results. In this process we assume that there is always an inexplicable part of the loss distribution that our regression cannot explain, and that we assume this part to be a universal constant among financial entities. We especially saw that heteroskedasticity must be tested for in the data and addressed during the regression process.

In the next chapter we retain these learnings as points of reference when we consider the statistical theory behind our selected solution: quantile regression. We focus on the quantile regression theory associated with heteroskedasticity. Our main sources are Koenker (2005) and Davino et al. (2014).

Chapter 4

Advanced literature topics

4.1 Introduction

As quantile regression is the chosen the method for solving Problem I of external scaling, we track the developments of quantile regression in academic literature in this chapter. The oversight compiled and inter-linked here, seeks to improve theoretical knowledge of the field of quantile regression whilst allowing for illustrations of practical application. This forms a highlighting element to the outline we have once reviewing inference on quantile regression techniques as presented in inter alia the extensive theoretical works of the Koenker (2005) and the practical applications Davino et al. (2014). In our core section we focus on fitting techniques for quantile regression. We review the classic literature by Powell (1984) as supported by Koenker (2005) and Davino et al. (2014).

4.2 The theory and application of quantile regression

4.2.1 Introduction

Davino et al. (2014) sets the background for understanding quantile regression by revisiting our view of so-called ‘classical’ regression. For classical regression we consider the values of a dependent response variable as defined by an independent explanatory variable(s). The base idea though is that the dependent variable is therefore conditional on the values of the explanatory variables(s).

The terminology used by Davino et al. (2014:1) is that classical regression focuses on a specific *location* of the response variable. A quantile regression approach aims to extend this focus to *various locations*. In summary we agree that quantile regression versus classical regression is analogous to a distribution’s quantiles versus its mean.

4.2.2 Overview of the statistical interpretation

For this section we follow the same definition given in Davino et al. (2014:2) by first considering how the median (itself the 50th percentile and thus central quantile) of a distribution can be defined as a minimization process. We are interested in regressing

variable Y conditionally on a set of independent variables $\mathbf{X}$, classically expressed as $E(Y|\mathbf{X})$.

Considering firstly for classic OLS, we conceptualize the mean as the minimizing distance between Y and some centre c - squaring to avoid offsetting positive and negative distances. A distribution's median, however, aims to minimize the (sum of) absolute deviations from said centre c :

$$Me = \underset{c}{\operatorname{argmin}} E|Y - c| . \quad (4.1)$$

We then see the definition of the common cumulative distribution function as

$$F_Y(y) = F(y) = P(Y \leq y) . \quad (4.2)$$

Subsequently we define the concomitant quantile function as the inverse of said function

$$Q_Y(\theta) = Q(\theta) = F_Y^{-1}(\theta) = \inf \{y : F(y) > \theta\} , \quad (4.3)$$

with $\theta \in [0, 1]$.

We know that if $F(\cdot)$ is strictly increasing and continuous, we have the definition that $F^{-1}(\theta)$ is a unique value y so that $F(y) = \theta$. We review a graphical illustration provided of the concept (Davino et al., 2014:3).

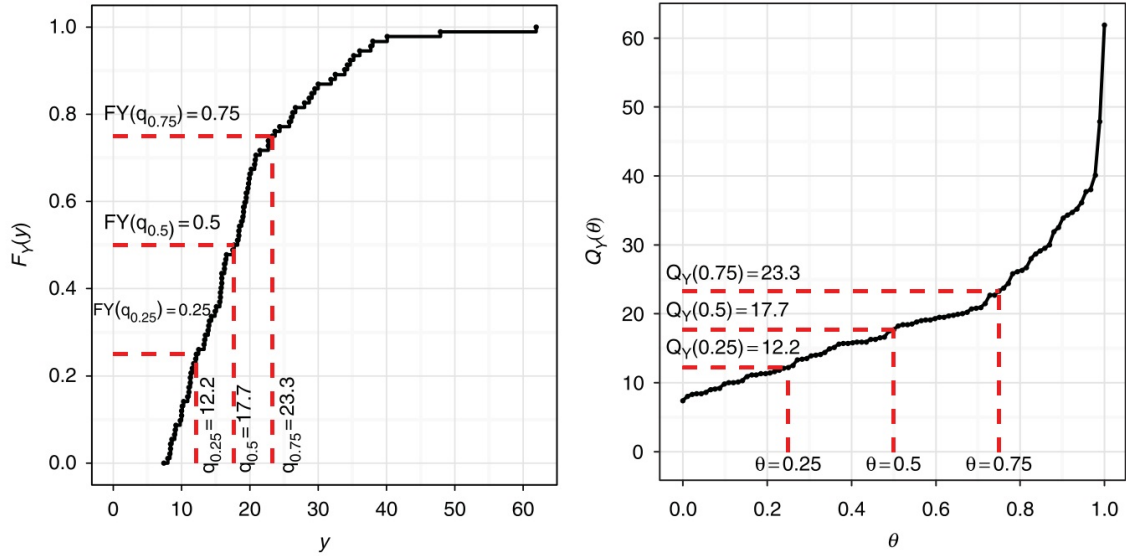


Figure 4.1: Cumulative distribution functions vs. quantile functions

SOURCE: Davino et al. (2014:3)

When reviewing the minimization process, we see the quantile redefined in terms of the process

$$\begin{aligned} q_\theta &= \underset{c}{\operatorname{argmin}} E|\rho_\theta(Y - c)| \\ &= \underset{c}{\operatorname{argmin}} \left\{ (1 - \theta) \int_{-\infty}^c |y - c| f(y) d(y) + \theta \int_c^{+\infty} |y - c| f(y) d(y) \right\} , \end{aligned} \quad (4.4)$$

where $f(y)$ indicates the probability density function for the stochastic variable Y .

Equivalent to our definition of the median above, we see the argument that the minimization problem can be reinterpreted for the estimation of a mean function (i.e. a conditional mean function)

$$\hat{\mu}(\mathbf{x}_i, \boldsymbol{\beta}) = \underset{\mu}{\operatorname{argmin}} E[Y - \mu(\mathbf{x}_i, \boldsymbol{\beta})]^2, \quad (4.5)$$

where $\mu(\mathbf{x}_i, \boldsymbol{\beta}) = E[Y|\mathbf{X} = \mathbf{x}_i]$ is then considered the *conditional* mean function. Davino et al. (2014:7) subsequently consider this result for the case of a linear mean function $\mu(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i^\top \boldsymbol{\beta}$. The estimation function is then defined as a known least squares regression

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} E[Y - (\mathbf{x}_i^\top, \boldsymbol{\beta})]^2. \quad (4.6)$$

We then see the natural extension of this result to our quantile definition

$$\hat{q}_Y(\theta, \mathbf{X}) = \underset{Q_Y(\theta, \mathbf{X})}{\operatorname{argmin}} E[\rho_\theta(Y - Q_Y(\theta, \mathbf{X}))], \quad (4.7)$$

where $Q_Y(\theta, \mathbf{X}) = Q_\theta[Y|\mathbf{X} = \mathbf{x}]$ is defined as a generic conditional quantile function. We then finally the linear version defined as

$$\hat{\beta}(\theta) = \underset{\beta}{\operatorname{argmin}} E[\rho_\theta(Y - \mathbf{X}\beta)], \quad (4.8)$$

where (θ) indicates the specific quantile for which we are performing the regression.

We conclude this section by noting that the parameter estimates obtained from quantile regression can be considered in the same light as those resulting from OLS. We know that the parameters are interpreted as the rates of change of the dependent variable and can therefore agree with the parameters defined as follows

$$\beta_i(\theta) = \frac{\partial Q_\theta(Y|\mathbf{X})}{\partial \mathbf{x}_i}. \quad (4.9)$$

4.3 Practical overview with graphical diagnostics

4.3.1 Introduction

In this section we see how quantile regression offers a direct solution to dealing with heteroskedasticity, which OLS techniques cannot. That being said, we also observe the value of using quantile regression even when homoskedasticity is guaranteed. To illustrate these concepts we generate our own hypothetical models following the ideas presented by Davino et al. (2014:38) and Koenker (2005:100-104).

4.3.2 Practical review

To illustrate the different capabilities of quantile regression within both homoskedasticity and heteroskedasticity we construct our own hypothetical models which are subsequently delineated graphically.

Recall that heteroskedasticity is present if the variance of a variable function is functionally related to its mean. We therefore incorporate this practice in our model as

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad (4.10)$$

where we have the following tabulated inputs listed in Table 4.1.

Table 4.1: Illustration of quantile regression for homoskedastic and heteroskedastic environments

Model element	Homoskedastic model	Heteroskedastic model
$\hat{\beta}_0$	7.5	7.5
$\hat{\beta}_1$	0.25	0.25
x_i	1, ..., 100	1, ..., 100
e_i	$\sim \mathcal{N}(0, 2^2)$	$\sim \mathcal{N}(0, (0.02 + 0.04x_i)^2)$

We take $\hat{\beta}_0$ and $\hat{\beta}_1$ as fixed inputs. The term x_i on which the quantile regression is applied conditionally is defined as a sequence from 1 through 100. Using a basic quantile formulation of the first to ninth quantiles q_θ and simulating according to the error term, e_i , as defined in Table 4.1 we obtain the graphical results given in Figure 4.2.

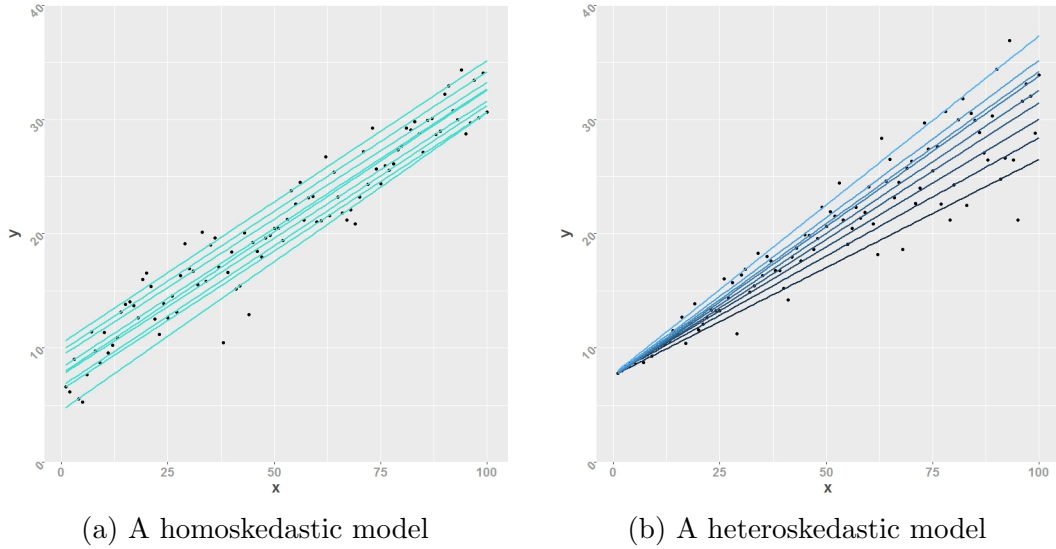


Figure 4.2: Quantile regression in homoskedastic and heteroskedastic environments

We can see the clear difference quantile regression can make for the respective environments. In Figure 4.2a one can see the virtually constant variance as an oscillating pattern around the central quantile line. We also note that the gradients of the respective quantiles seem to be equal mostly. Should confidence intervals be constructed around this fluctuation we would expect our subsequent inference to be

fairly successful. This would imply that a single quantile gradient can be calculated and the merely the intercept would require updating as we move across the outcome space of the quantiles.

However, in the case of Figure 4.2b, it is clear that there is significant deviation away from the central quantile as the value of x increases. The graph illustrates that the quantiles are not parallel to each other and thus have unique gradients. Using static confidence bands for this regression would result in inappropriate inference conclusions. The result is that distinct gradients and intercepts ought to be calculated for all quantiles when performing the quantile regression.

In Koenker (2005:100-104) we see a practical example of these concepts explained in the form of quantile regression graphs for individual explanatory variables. We retain the explanation here, but generated our own graphs using the R package ‘quantreg’ code example with its associated dataset (Koenker, 2017).

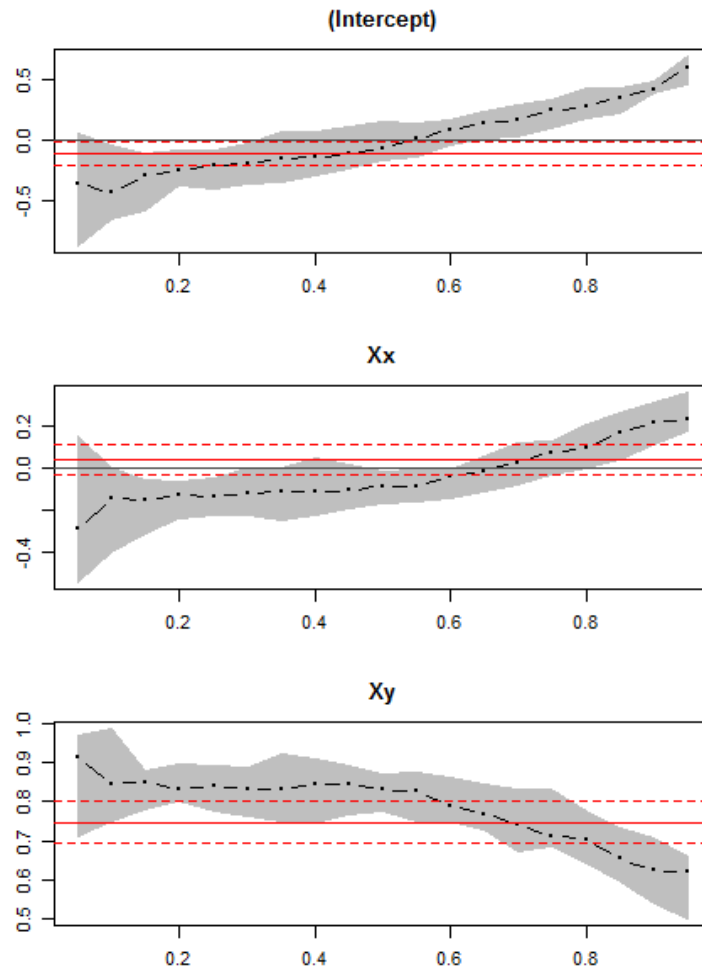


Figure 4.3: An example of quantile regression for individual explanatory variables

We consider the dataset ‘CobarOre’ from said R package which has three variables; the dependent variable of ore thickness and two independent variables respectively indicating the x - and y -coordinates of the ore’s physical location. We perform full scale quantile regression; i.e. over a wide set of quantiles $\theta \in [0.05, 0.95]$ in increments

of 5%. We obtain the graphical results in Figure 4.3 by following the illustration in (Koenker, 2005:102-103).

We see the results of quantile regression for the model specification of two independent variables and an intercept. Note that the quantile regression is the dashed-dotted black line with confidence intervals in grey, whilst the red solid line is normal OLS regression with its confidence bands. It is therefore clear, that quantile regression is significant as the values of the conditioning variable(s) change.

4.3.3 Summary

In our review of the practical aspects of quantile regression, we firstly saw how quantile regression can directly address regression issue when heteroskedasticity is present in the data. For such an instance it is advised that a full-scale quantile regression be performed where quantiles are modelled to have distinct intercepts and gradients to capture the variance profile. Secondly, we noted the quantile regression practice of first determining goodness-of-fit of a quantile regression model to confirm whether a homoskedastic or heteroskedastic quantile model would be more appropriate for the data, prior to actually estimating our parameters. Finally, we also saw different graphical diagnostics which can be performed on quantile regression with the purpose of better understanding empirical results, but also in terms of variance profile, to guide our further steps in the quantile regression process.

An important closing point we reflect on is a conclusion drawn in Cope and Labbi (2008:33) on the use of quantile regression specifically for scaling. It is explained that a successful regression can only be measured in terms of how well it reflects the original distribution in its scaled versions. It is explicitly stated that the amount of variation explained by a quantile regression is meaningless. This is confirmed by an earlier study of Koenker and Machado (1999:1297), where it is explicated that the coefficient of determination measuring the explanation of variation in the dependent variable is a *local* measure; i.e. should such a measure be considered for interpretation, there should be an R^2 for each quantile.

4.4 Chapter conclusion

In this chapter we firstly saw the theoretical derivation of quantile regression as an extension of classical OLS. We learned that the regression process can be performed for as many quantiles as desired. We therefore saw that viewing the locus of OLS as a point, the locus of quantile regression is a line equation. Secondly, we observed how this practice of an extended locus resulting from quantile regression can assist in defining and capturing heteroskedasticity when regressing for our response variable.

This chapter has allowed us to familiarize ourselves with the techniques we can consider as supplementary to the quantile regression process. These techniques are specifically aimed at our study of *inter alia* heteroskedastic data, data behaviour in tail regions, and possibly large coefficients. These tools for quantile regression are now applied in our practical work of the suggested solution for external scaling in the following chapter.

Chapter 5

The data environment

5.1 Introduction

In this chapter we investigate the data environment within which we intend to construct our solution for the problem external data scaling. We reflect on the data sources available to us, the structures of said sources in terms of their metadata, and finally review their basic descriptive statistics.

We inspect the ILD and ELD which we are using for our study. We note the anonymity handling for the ILD. For ELD we consider how the external database is compiled and maintained. In terms of metadata, we discuss the defining characteristics of the datasets by reviewing data totals, noting the periodicity and points on data preparation such as the inflation adjustments. We delve deeper into data preparation by focusing on the significant detail of the filtering processes of both ILD and ELD.

We then commence our data analysis of the ILD and ELD with a specific focus on defining our response variable in terms of the proposed explanatory variables. We firstly consider the form of the response variables, operational losses, and make a comparison between ILD and ELD to understand the empirical context of the problem of external data scaling. Subsequently we focus on the respective aspects of numeric and factor level explanatory variables. We provide graphical diagnostics of the data structures in terms of the magnitude of numeric variables and the associated factor levels. We then analyze the dependence composition of the numeric explanatory variables which we intend to use as possible proxies for the response variable.

Finally, we focus on the distributional forms of the factor level variables for the response variable. We discuss unexpected / counter-intuitive attributes of the distributional forms noting aspects we carry over to our regression modelling.

5.2 Data overview

5.2.1 Data sources

Internal Loss Data

The ILD we are utilizing for our study originates from a financial institution hosting the study. The host institution is a Big 4 bank in South Africa using the AMA

approach¹ for modelling operational risk capital. The loss data are captured by operational risk managers within the business units, who are in general only qualitatively trained in operational risk management.

Basic sense checks to establish a meaningful level of data quality are conducted firstly by a quantitative-orientated *data management* team. The *modelling team* forms a subsequent layer in this data quality assessment. In addition, data quality assessments are normally conducted by independent third-party model vetters as part of annual model validations. We therefore assume the raw data to be a true and correct reflection of the institution's operational risk profile.

External Loss Data (ELD)

The ELD we are using in our study consist only of the SAS[®] OpRisk Global Data. These data were obtained from the author's academic entity, the North-West University (Potchefstroom campus) who is an active on-site user of the SAS[®] software. No other ELD served as input to this study.

5.2.2 Metadata

Internal Loss Data

In total there are just over 82,000 observations captured as loss events spanning a period of eight years which includes the major world recession of 2007 / 2008 and its aftermath (Adebambo et al., 2015:648). These observations each have seventeen defining fields (variables) including a single unique identifier key. The fields each have a single attribute of being an identifier, character, numeric, factor, or date variable.

Among the fields are attributes providing the details of which area a loss affected, along with the nature of the event. Further details are provided about the dates concerning the loss, the amount of the loss, its final financial status, and a description of the loss. Certain calculated fields are also included in this base dataset; i.e. the losses are flagged for being associated with aggregate losses if necessary, as well as whether there is a credit (risk) impact related to the loss.

In Section 5.3.1 we see how using our knowledge of the field attributes of in combination with filtering rules are prescribed by the Basel Committee on Banking Supervision (2006), allows us to filter the data to a subset deemed appropriate for modelling purposes.

External Loss Data

In total there are more than 28,000 observations captured as loss events in the period from March 1971 through December 2013. Recall from Section 3.4.1 that the basic lower bound for inclusion of these observations is set at \$ 1 million. These observations each have 49 defining fields (variables) including a single unique identifier

¹At present, the South African regulator is of the opinion that banks in SA using the AMA approach should continue to do so until further notice from the regulator. The new Standardized Measurement Approach (SMA) is being investigated, but it is opined that the AMA will still be applied for calculating economic capital in the operational risk environment.

key. The fields each have one attribute of being an identifier, character, numeric, factor, or date variable.

The fields include *inter alia* attributes which provide the details of which area a loss affected, along with the nature of the event. These fields specifically are extended across further fields to a very granular taxonomy. These data also contain fields indicating the region in the world where the loss occurred as well as the specific institute where it occurred, and its revenue, assets under management and number of employees. Further details are provided about the dates concerning the loss, the amount of the loss, its final financial status, and a description of the loss. Certain calculated fields are also included in this base dataset; i.e. the losses are flagged for being associated with aggregate losses if necessary and here we also see the loss amount adjusted for cumulative inflation².

As mentioned before, in Section 5.3.1 we combine the filtering rules are prescribed by Basel Committee on Banking Supervision (2006) with the dataset attributes reviewed here to filter the data to a subset deemed appropriate for modelling purposes.

Operational Risk Categories (ORCs)

We provide a user guide here for the reader on the ORCs. As we mostly use abbreviations for the cell descriptions and dimensions, this tabulation aids in tracking the different components. Refer to Table 5.1.

Table 5.1: Guide to ORCs

Business Lines (BL)	abbr.	Event Types (ET)	abbr.
Agency Services	AS	Business Disruption and System Failures	BDSF
Asset Management	AM	Clients, Products and Business Practices	CPBP
Commercial Banking	CB	Damage to Physical Assets	DPA
Corporate Finance	CF	Employment Practices and Workplace Safety	EPWS
Payment and Settlement	PS	Execution, Delivery and Process Management	EDPM
Retail Banking	RB	External Fraud	EF
Retail Brokerage	RBR	Internal Fraud	IF
Trading and Sales	TS		

5.3 Data preparations

5.3.1 Filtering of data

Internal Loss Data

The filtering rules were obtained from the data management and modelling teams at the host institution. These are tabulated below and it is clear that the filtering rules are adherent to Basel II requirements (Basel Committee on Banking Supervision, 2006).

²The adjustment calculation was provided in detail as part of the dataset and is confirmed as correct by the author.

In Table 5.2 we see the variables on which the filtering is applied with the specific variable outcomes which are *excluded* prior to modelling. We point out that the threshold for inclusion is R 10,000. From personal experience this seems to be a general choice across South African banks. This is naturally not aligned to international standards as the pure magnitude of international activity and exchange rates are not directly comparable to those we face in our South African environment.

A final step in the filtering process is the expected step of enforcing unique IDs across observations; i.e. to delete duplicated observations.

Table 5.2: Listing of filtering rules for ILD

Variable	Filter attributes
Aggregate losses	Yes
Impact type	Not equal to loss
Financial status	Not equal to settled, cost, provisioned
Credit loss	Yes
Loss amount	Less than 10,000

External Loss Data

The filtering rules were obtained from the data management and modelling teams at the host institution. These are tabulated below and it is clear that the filtering rules are well-aligned to the filtering applied for ILD.

Table 5.3: Listing of filtering rules for ELD

Variable	Filter attributes
Industry Sector Name	Equal to ‘Financial Services’
Basel Business Line - Level 1	Not equal to ‘Insurance’
Financial Loss Status	Equal to ‘Final’

In Table 5.3 we see the variables on which the filtering is applied with the specific variable outcomes which are *included* prior to modelling.

An additional filter we applied relates to the variables of the firms’ assets, revenue and shareholder equity. Given the common practice of logarithmic transformation of these substantially large variables, it becomes necessary to remove observations where the variables assets and / or revenue are less than or equal to zero since logarithmic transformations cannot be obtained for these values. Note the exclusion of the numeric proxy, ‘number of employees’; an apparent capturing error in the variable relating to the number of employees has rendered this variable unusable for our study.

Any other filtering which may arise due to the actual modelling process is mentioned where such occurs in the modelling approach.

5.4 Data analysis

5.4.1 Direct comparison of the response variable in ILD vs. ELD

Overview

Given the isolated analyses we have performed on the ILD and ELD respectively, we now consider a comparison between the two datasets in terms of our response variable. This comparison functions firstly as an informative measure to contrast the scope (both frequency and severity) and nature (distributional forms as resulting from the frequency and severity aspects) of the respective operational risk environments - i.e. the global environment vs. the South African environment.

Secondly, and most importantly, this comparison illustrates the core motivation for our study. From this comparison we directly observe that it is infeasible and could potentially be negligent risk management to use the two datasets together directly without applying a suitable scaling mechanism.

Initial diagnostics

Before we illustrate our comparison of the two loss distributions of the ILD and ELD respectively, we applied the most basic scaling technique to the ILD: i.e. we apply the exchange rate to the ILD to allow for a level comparison between the ELD in USD and the ILD originally in ZAR. Since we have already applied cumulative inflation to level all the losses to the same point in time (i.e. end-2013), we therefore use the USD - ZAR exchange rate as applicable at that time.

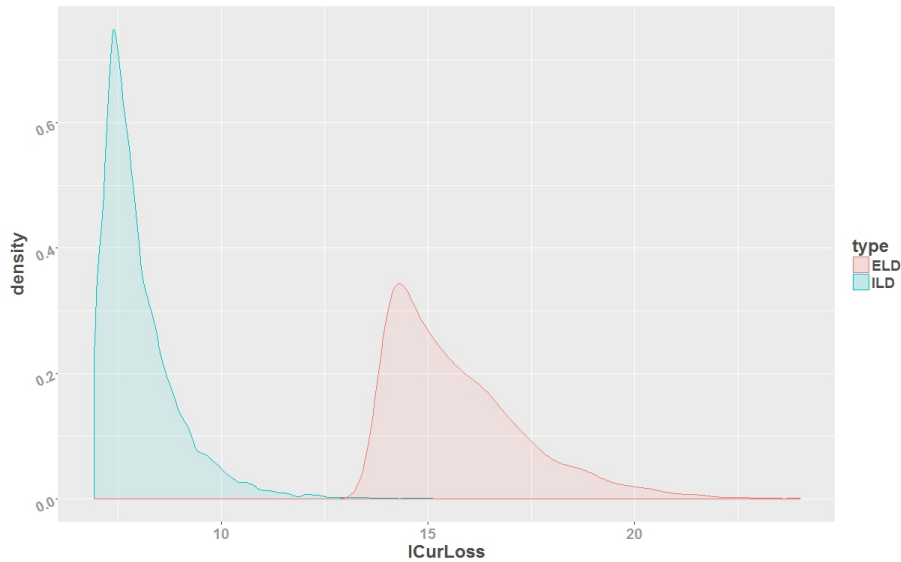


Figure 5.1: Loss density distributions of ILD vs. ELD

In Figure 5.1 we therefore see the ILD of our host institution in comparison to the global environment of the SAS[®] ELD. The graph illustrates the logarithmic transformation of the respective loss distributions.

We can observe in Figure 5.1 the significant distributional differences, but also the similarities. In terms of the latter note we specifically note that the two distributions are comparably skewed to the right. This functional form is considered the general form for loss distribution in operational risk literature (Chernobai et al., 2007:82).

The differences between the two distributions are however, more pronounced. Firstly, the locations of the distributions are far apart. Along with the locations, the scales of the distributions do not appear alike; we observe that the ILD has a very narrow scale, whereas the ELD has a much wider scale.

A further point of difference is the kurtosis of each distribution. For the ILD distribution, we see a very high frequency of smaller losses clustered around the mode of the distribution. There is a clear long tail in the distribution. However, in comparison to that of the ELD loss distribution, the ILD distribution is much more platykurtic; i.e. the ELD distribution are significantly more leptokurtic.

It is specifically this attribute which makes the inclusion of ELD necessary since it informs our loss distribution for modelling with *more* and *larger* operational losses. Refer to Figure 5.2 where we see a sectioned version of the density distributions zoomed in on the tail of the ILD and containing the full ELD distribution - note the changes in scales of the respective axes³. One should specifically note that the lower bound was set equal to the mean of the ILD distribution, thus re-affirming the highly platykurtic nature of the ILD when considering in the context of Figure 5.1.

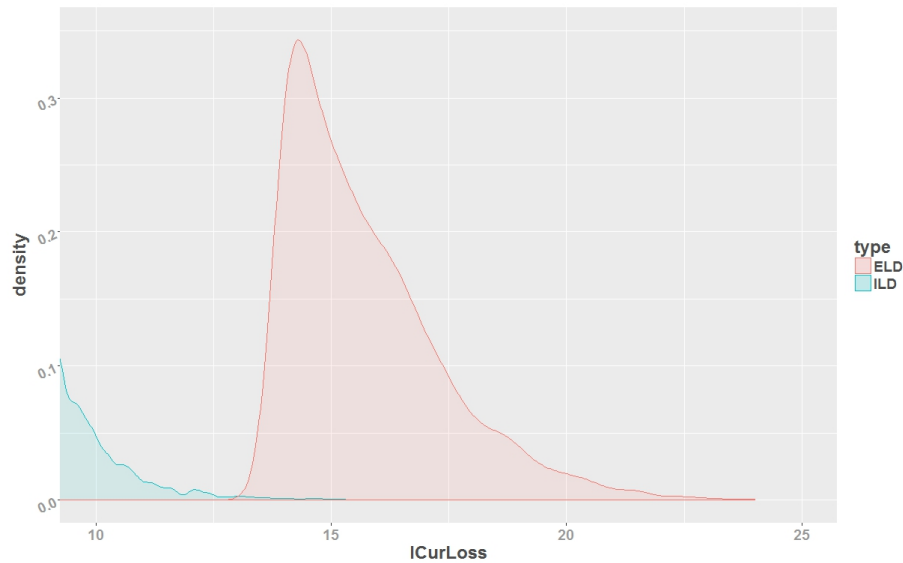


Figure 5.2: Tail of the ILD distribution vs. ELD distribution

Link to our problem study

The analysis and conclusion we performed on the response variable, operational losses, serves to delineate our problem of ILD and ELD being incompatible for direct combination. In answering the problem we propose a scaling of the ELD downwards -

³Note that the scales of Figures 5.1 and 5.2 are not adjusted equally for both axes. Therefore a direct comparison between the two would not be appropriate. The axes were explicitly exaggerated to illustrate the difference in tails in terms of the x -axis; i.e. the logarithm of operational losses.

i.e. applying a modelled reduction to the ELD - so that it is more comparable to the ILD. Intuitively we expect a similar positively-skewed distribution located between the ILD and ELD. In terms of the high frequency seen for lower losses in the ELD, this may influence the body of the resulting scaled distribution to be centred close to the ILD. However, the significantly larger losses of the ELD, even when scaled, we expect to elongate the tail of our resulting scaled distribution.

Conclusion

In this section we considered a direct comparison between ILD (adjusted only for exchange rates) and ELD. In this analysis we saw that the two respective distributions have comparable distributional forms - comparable skewness - and therefore a combination can be considered as appropriate in principle. We also noted that a scaling mechanism is necessary to better quantify and subsequently relate the location, scale and kurtosis of the two distinct distributions. Specifically in terms of the kurtosis, we saw a clear illustration that our ILD does not contain sufficient nor sufficiently large losses; whereas such losses are contained in the ELD. This is therefore a further motivation to construct a scaling mechanism to effectively combine the two datasets.

5.4.2 General analysis - ILD

Introduction

According to our literature review in Chapters 3 and 4, we investigate a wider set of explanatory variables to model our response variable, the loss amounts. We intend to firstly consider a size indicator which is naturally also a numeric variable like the response variable. As possible proxies for the numeric variables we consider a financial institute's revenue (gross income), assets under management, shareholder equity. The additional variables we include are the factor variables which we consider inherent to the operational risk environment; i.e. the BLs, ETs and regions of operational losses.

In our analysis we therefore use data techniques applicable to the two respective types of variables. Therefore, for numeric variables we perform descriptive statistics to obtain a better understanding of our datasets. For the factor variables we analyze the datasets for our response variables (i.e. operational losses) in terms of the factor levels; e.g. consider how loss amounts cluster in a specific BL or ET. The factor variables are also analyzed within the two perspectives of general operational risk modelling; i.e. in terms of the underlying severity vs. the frequency.

Overview - Numeric variables

To open our analysis of the data, we consider the summary statistics of our datasets. For the numeric variables we review the descriptive statistics of the datasets such as number of observations, mean, standard deviations and lower and upper bounds. We tabulate the number of observations, the mean, standard deviation and lower and upper bounds for the variables. The results are presented in Table 5.4. As expected we see that the variable 'Assets' shows much larger magnitude in general in

comparison to the other variables. The variables ‘Revenue’ and ‘Equity’ show a close likeness in their magnitude for all measures in Table 5.4. Specifically, we note here that the losses are significantly smaller in magnitude when compared to the possible size indicators. This may indicate that using size indicators is not appropriate for ILD. We consider this point during our modelling.

Table 5.4: Summary statistics of the numeric variables in the ILD - (* in USD ’000)

Proxy	N	Mean*	St. Dev.*	Min*	Max*
Loss	15,684	11	123	1	10,354
Revenue	15,684	3,675,555	189,672	3,142,192	4,062,162
Assets	15,684	73,090,728	3,138,172	65,349,476	78,282,105
Equity	15,684	5,607,623	403,833	4,519,737	6,440,040

Overview - Factor variables

We first concentrate on the frequency of operational losses among the factor levels of the BLs and ETs. The results of the operational losses’ clustering for firstly BLs and then ETs are noted in Table 5.5. We observe a major clustering of loss event frequency in the BL ‘RB’ - Retail Banking.

Table 5.5: Summary statistics of the factor variables - ILD number and amount of losses

BLs	number	amount	ETs	number	amount
AS	< 1%	< 1%	BDSF	3%	4%
AM	1%	1%	CPBP	2%	14%
CB	7%	15%	DPA	1%	1%
CF	< 1%	1%	EPWS	1%	2%
PS	3%	16%	EDPM	16%	36%
RB	85%	49%	EF	76%	39%
RBr	1%	1%	IF	2%	5%
TS	2%	17%			

This result is not completely unexpected in terms of the South African context; given that the major focus of the banking system is still centred on banking the average citizen. We subsequently see very low frequencies in the remainder of the BLs. Especially for the factor levels ‘AS’ - Agency Services - and ‘CF’ - Corporate Finance the frequencies are so low they classify as zero when using no decimals.

With respect to the ETs, the greatest clustering is seen for ‘EF’ - External Fraud with three-quarters of the operational losses. This is followed by ‘EDPM’ - Execution, Delivery and Process Management. Further we see the same very low frequencies for the remainder of ETs.

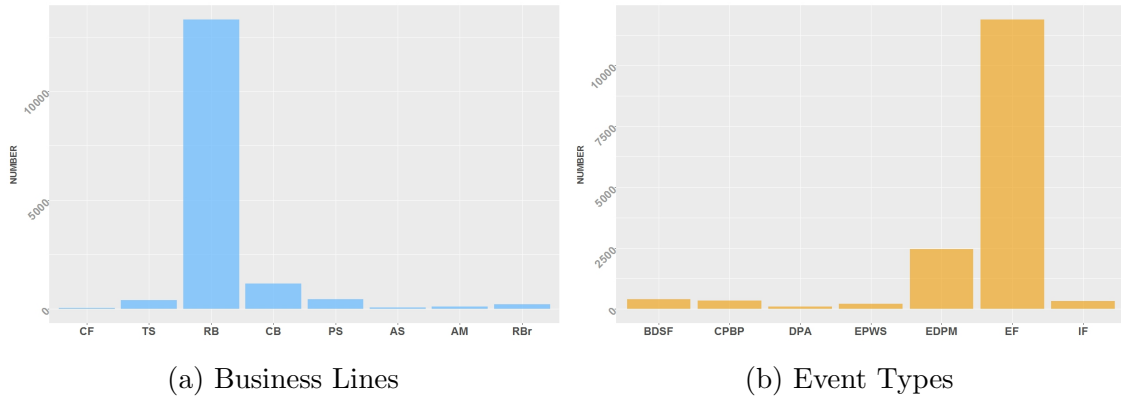


Figure 5.3: Bar charts of the factor variables - ILD number of observations

We now review the graphical illustration of the tabulated results in Figure 5.3. Here the tabulated results appear even more pronounced in the bar charts. However, this illustration allows us to observe that BL ‘CB’ - Commercial Banking - in Figure 5.3a should receive proper attention during modelling.

In terms of loss frequency, we intuitively expect that our delineated factor levels require the most attention during parameter fitting.

Continuing our analysis of the numeric variables we consider the operational loss per factor level in terms of the actual loss amounts; the severities. In reference to the loss amounts in Table 5.5 we specifically note that a very different situation is seen for both the BLs and ETs. Even though the core concentrations seen for the number frequencies are still apparent, in this instance the concentrations have lowered and there is a more wide-spread distribution of the operational losses.

For the BLs there is almost equal clustering of loss amounts under ‘CB’, ‘PS’ - Payment and Settlement - and ‘TS’ - Trading and Sales. In terms of the ETs we see comparable clustering between ‘EF’ and ‘EDPM’ with a further noteworthy clustering in ‘CPBP’ - Clients, Products, and Business Practices.

In Figure 5.4 we see our conclusion confirmed from the loss amounts in Table 5.5. In Figure 5.4b is it moreover visible that this distribution is considered the most *even*. We note how these graphical diagnostics in the bar charts are reflected in our parameter inference during modelling.

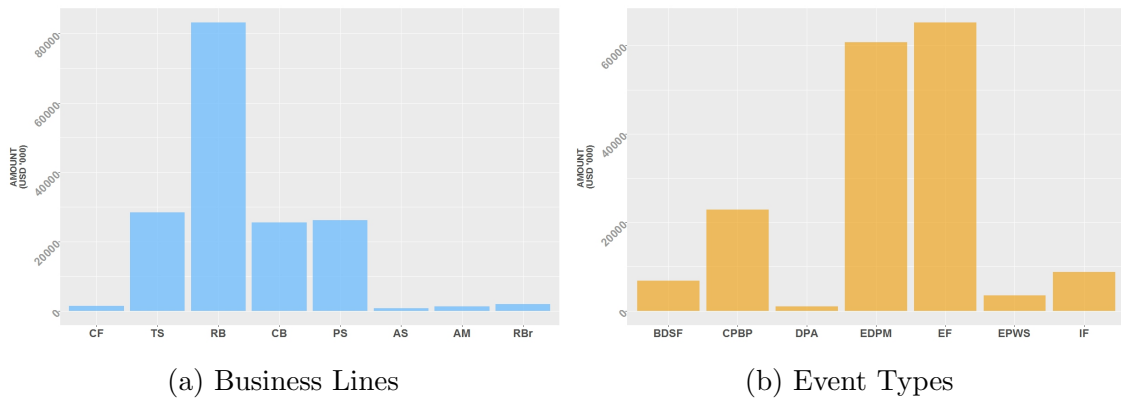


Figure 5.4: Bar charts of the factor variables - ILD loss amounts

Missing values

In this section we analyze the missing values on our filtered ILD set. The set we use for modelling includes our independent variables; i.e. the possible numeric proxies for the loss variable as well as the factor variables for our extended analysis as guided by Dahlen and Dionne (2010).

Missing values may have a significant effect on the modelling process. Variables chosen in principle for having appropriate business value or theoretical insight may become unusable due to too many missing values. This may result in either discarding certain observations with missing values, or discarding the entire variable for modelling purposes. Another possibility is reinvestigating the underlying issues for missing values. In certain cases it may be appropriate to use imputation to estimate missing values, or when the distribution forms are known, maximum-likelihood estimation of missing values may be possible (Kabacoff, 2011:355).

Our analysis starts with counting the amount of missing values per independent variable and then checks for combinations of missing values. The results are illustrated in Figure 5.5. We conclude from the graphic diagnostics that there are no missing values in any of our variables to be used for modelling. Therefore no treatment of missing values is required.

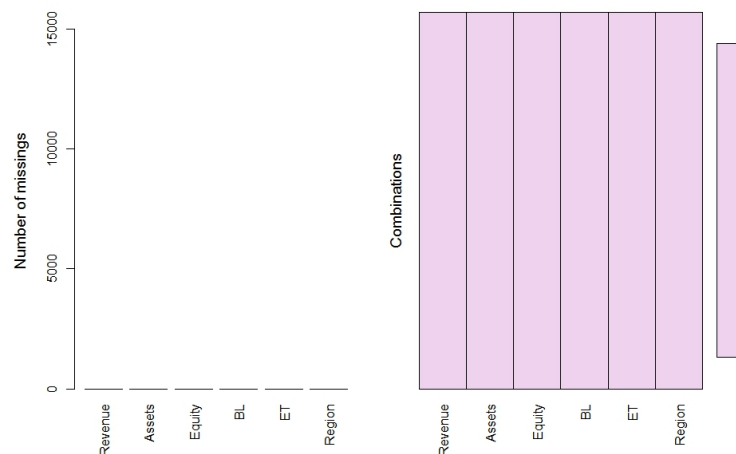


Figure 5.5: Analysis of missing values for ILD independent variables

5.4.3 General analysis - ELD

Overview - Numeric variables

In this first part of the analysis we compile summary statistics for the numeric variables of the ELD. The summary statistics consists of the number of observations, and basic statistical measures; e.g. the mean and standard deviation. The results are provided here. It can be seen in Table 5.6 that the variable ‘Assets’ has the largest scale with its mean and maximum value the highest across the testing sample. The variables ‘Revenue’ and ‘Equity’ show fairly comparable results. Our response variable, ‘Loss’, seems to be the *smallest* variable in terms of scale. Note the minimum value reflects the threshold for data capturing by SAS®.

Table 5.6: Summary statistics of the numeric variables in the ELD - (* in USD millions)

Proxy	N	Mean*	St. Dev.*	Min*	Max*
Loss	1,924	77	739	1	26,449
Revenue	1,924	31,914	43,345	< 1	196,449
Assets	1,924	575,449	798,365	< 1	3,783,173
Equity	1,924	36,727	56,323	< 1	236,956

Overview - Factor variables

Our second part of the analysis consists of inspecting the factor variables of the ELD. For this analysis we firstly define the distinct factor levels for each of the factor variables and then determine the proportions which each of those factor levels make of the variable as a whole; i.e. the *number* of losses classified as per the relevant category. These results are tabulated in Table 5.7.

From Table 5.7 it can be concluded that in terms of BLs, the ELD losses seem to be significantly concentrated under ‘Retail Banking’ with 42% of the losses. The second highest business line for losses is ‘Commercial Banking’; with more than half less; 25%. With respect to ETs, the clear majority is focused in category CPBP with 35% of the overall losses. Almost tied in second place, is the respective fraud levels; EF and IF. The leader for losses on an international scale is as expected North America with 53% of the losses. Europe is the second highest with less than half of that at 24%⁴.

Table 5.7: Summary statistics of the factor variables - ELD number of losses

BLs	number	ETs	number	Region	number
AS	< 1%	BDSF	1%	Africa	2%
AM	6%	CPBP	35%	Asia	17%
CB	25%	DPA	1%	Europe	24%
CF	5%	EPWS	2%	North America	53%
PS	2%	EDPM	5%	Other Americas	1%
RB	42%	EF	30%	Other	3%
RBr	10%	IF	26%		
TS	10%				

In order to better gauge the relative aspects of these values, we have compiled bar charts for the respective factor variables to allow direct comparison of the magnitudes of loss taxonomies. The graphical results are presented in Figure 5.6.

We can conclude from Figure 5.6 that a minimal number of losses arise from respectively Agency Services (BL), both BDSF and DPA (ETs), and Other Americas (Regions including Central and South America).

⁴Note that all the factor variables utilized in this study are nominal (as opposed to ordinal).

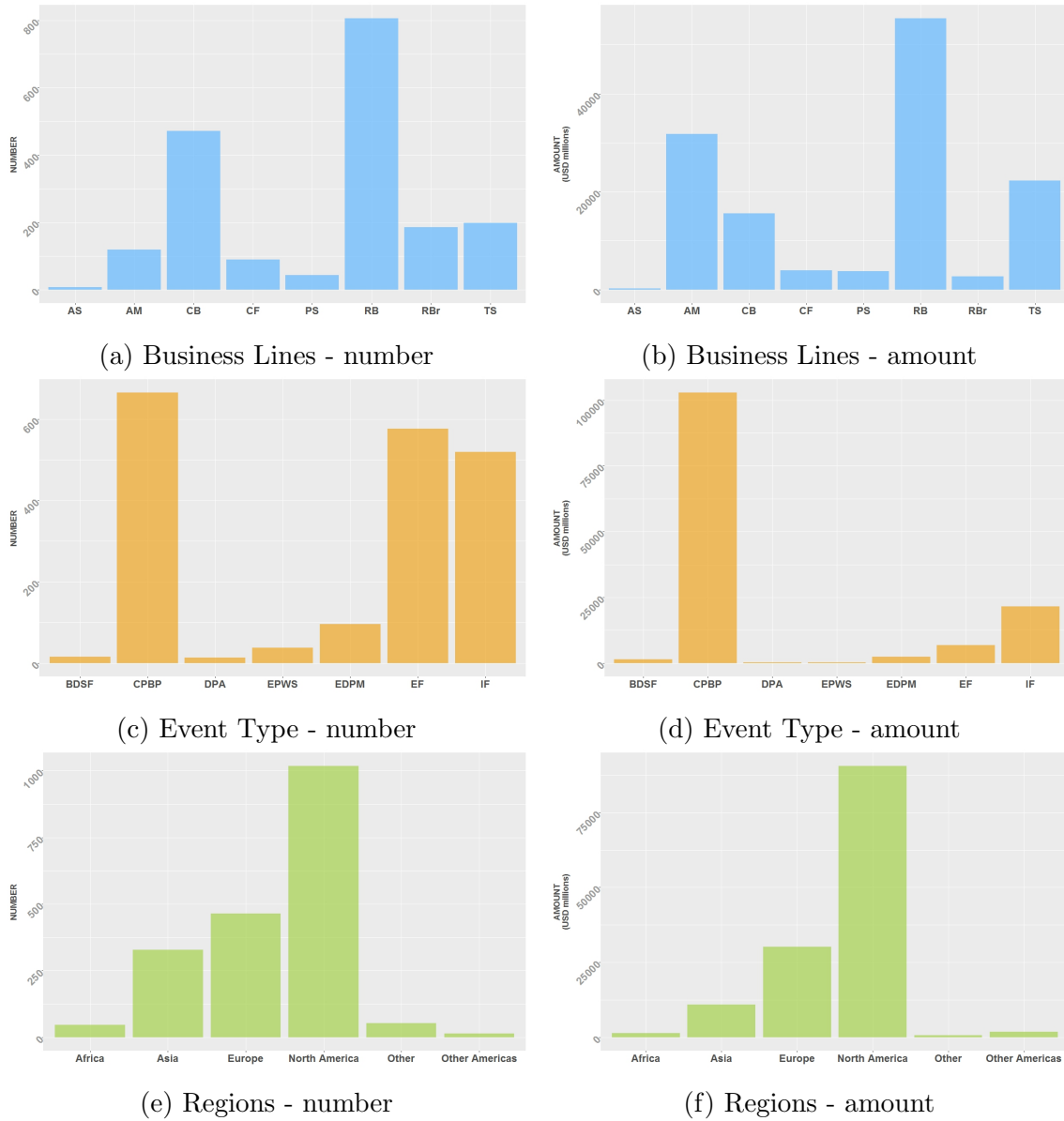


Figure 5.6: Bar charts of the factor variables - ELD number of observations and loss amounts

However, we now extend our analysis to inspect the taxonomy of the factor variables in terms of the actual *loss amounts* associated with the categorized losses. We follow the same approach as above and provide the tabulated percentages along with the concomitant bar charts.

Table 5.8 when viewed in conjunction with Table 5.7, provides interesting results. We see that the greatest loss amounts occur under the BL TS; however, this is a category where there is not a noteworthy amount of losses. A comparable situation is noted for BL AM. For ETs, the results are intuitive; the highest loss amounts are associated with the category seeing the highest number of losses; CPBP. A similar situation is seen for the factor variable Region where loss amounts are fairly strongly correlated with the number of losses. Note that there are numerous instances where the loss amounts were negligible in comparison to other proportions.

The bar charts in Figure 5.6 confirm our conclusion.

Table 5.8: Summary statistics of the factor variables - ELD loss amounts

BLs	amount	ETs	amount	Region	amount
AS	< 1%	BDSF	1%	Africa	1%
AM	26%	CPBP	76%	Asia	8%
CB	11%	DPA	< 1%	Europe	22%
CF	3%	EPWS	< 1%	North America	67%
PS	3%	EDPM	2%	Other Americas	1%
RB	39%	EF	5%	Other	1%
RBr	2%	IF	16%		
TS	16%				

Missing values

In this section we analyze the missing values on our filtered ELD set. The set we use for modelling includes our independent variables; i.e. the possible numeric proxies for the loss variable as well as the factor variables for our extended analysis as guided by Dahlen and Dionne (2010).

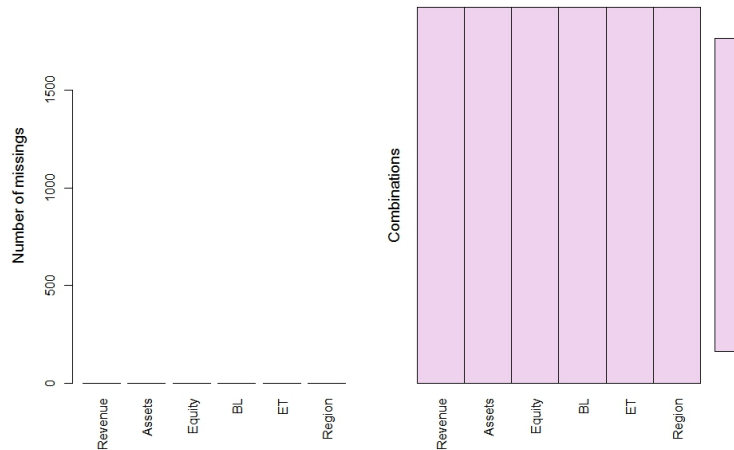


Figure 5.7: Analysis of missing values for ELD independent variables

Missing values may have a significant effect on the modelling process. Variables chosen in principle for having appropriate business value or theoretical insight may become unusable due to too many missing values. This may result in either discarding certain observations with missing values, or discarding the entire variable for modelling purposes. Another possibility is reinvestigating the underlying issues for missing values. In certain cases, it may be appropriate to use imputation to estimate missing values, or when the distribution forms are known, maximum likelihood estimation of missing values may be possible.

Our analysis starts with counting the amount of missing values per independent variable and then checks for combinations of missing values. The results are illustrated in Figure 5.7. From this figure we conclude that there are no variables with missing values.

5.4.4 The explanatory variables - size indicator

Overview

In this section we are completing the analysis to make a significant decision of our quantile regression model. If we recall the first literature relevant to this question: Shih et al. (2000) suggested that a proxy variable for losses in the form of a firm size indicator be used to forecast estimate future losses. The variables proposed in that study were revenue, assets, and number of employees. As mentioned, in our ELD set the variable ‘number of employees’ is not unusable and we therefore substitute it with another possible indicator of firm size; i.e. shareholder equity.

Our analysis to determine the best proxy is to investigate the association relationship between the possible proxy variable and our response variable. For this purpose we consider the correlation structures between the numeric size indicators – namely that of revenue, assets, equity – and our response variable, the operational losses. We therefore want a proxy which is appropriately indicative of the response variable, but is also representative of the other numeric variables.

We acknowledge the shortcoming of correlation as measure of association, but we also see that this was the method used by Shih et al. (2000:1), whereas Dahlen and Dionne (2010:1487) provide no indication of the choice and it seemingly remains an arbitrary choice.

Our comparison process consists of compiling a so-called ‘correlogram’ to visually analyze how well the firm size indicators correlate with our response variable. In a correlogram the two colours are used to indicate the nature of the association (whether the relationship is positive or negative); whilst the hue – intenseness of the colour – serves as an indicator of the strength of that associative relationship. The colour setting here is blue for positive correlation to red for negative correlation. A further addition to the correlogram is to include a modified pie chart which provides some measurable indication of the strength of association. To conclude our decision-making process, we incorporate two correlation matrices; namely those of Kendall’s τ and Spearman’s ρ rank correlation (Chernobai et al., 2007:265).

Internal Loss Data

On inspection of Figure 5.8 we notice a counter-intuitive result. As explained in Section 5.4.2 we are investigating the numeric variables analyzed in this correlogram for their appropriateness to serve as proxies for the response variable, operational losses. We therefore intuitively expect that the proxies would be positively related to the operational losses.

However, in the correlogram in Figure 5.8 all the proxy variables show negative correlations as indicated by reddish colouring of the correlation disks and the bottom triangle showing backward-slanting lines. We note though that these are very *small* correlations; i.e. low relational behaviour. We recall the linear nature and inherent

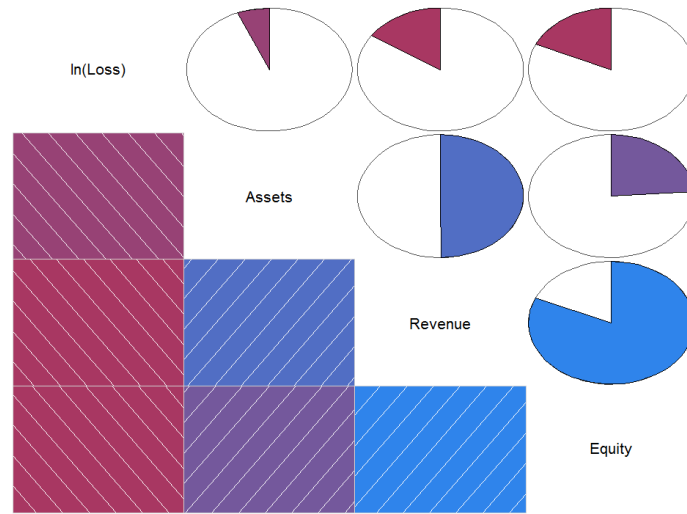


Figure 5.8: Correlograms of the ILD numeric proxies

shortcoming of correlation calculations as mentioned Section 5.4.4. Also, we reflect on the results obtained by Na et al. (2006) that there is no concrete relation between size indicators and the severity of operational losses - refer to Section 3.3.3. We note these points when reviewing our parameter estimation during quantile regression.

A point that should be highlighted here to expand our contextual understanding of the results, is the nature of the size indicator proxy variables. Given that operational losses for ILD are obviously sourced from a single firm, many losses are associated with single value for a proxy explanatory variable since financial results are only posted annually. These attributes of the financial process may be compounding factor in distorting the data analysis for proxy variables.

Table 5.9: ILD Kendall's τ rank correlation

	ln(Loss)	Revenue	Assets	Equity
ln(Loss)	1	-0.131	-0.052	-0.194
Revenue	-0.131	1	0.670	0.534
Assets	-0.052	0.670	1	0.204
Equity	-0.194	0.534	0.204	1

Moving on from our main focus of the relationship between the response variable and loss proxies, we observe that all the numeric proxies are highly correlated to each other, save for the relationship between the proxy variables 'Assets' and 'Equity'. These results are numerically illustrated in Tables 5.9 and 5.10. Our investigation into the dependence structures of the numeric proxies thus continues.

We perform a test of association between the sampled pairs; i.e. the logarithmic transformation of the losses vs. the proxies Revenue, Assets, and Equity respectively. We perform this test at a confidence interval of 99% and specify the calculation method of Kendall's τ for a two-sided test. This test thus performs inference on an

Table 5.10: ILD Spearman's ρ rank correlation

	ln(Loss)	Revenue	Assets	Equity
ln(Loss)	1	-0.180	-0.073	-0.262
Revenue	-0.180	1	0.770	0.602
Assets	-0.073	0.770	1	0.177
Equity	-0.262	0.602	0.177	1

interval to determine whether we can reject the null hypothesis that the dependence measure - i.e. our correlation values - that said interval contains the value 0; thereby rendering the concomitant parameter estimation useless. Having used the hypothesis test:

$$H_0 : \text{true } \tau = 0$$

$$H_A : \text{true } \tau \neq 0 \quad ,$$

the results of the testing are presented in Figure 5.11. It is clear from all tests' minute p -values that the null hypotheses cannot be rejected. Also note that the τ estimates are negative for all pairwise testing.

This results links to the output obtained by Na et al. (2006) that the numeric proxy as size indicator may not be an appropriate decision for inclusion in a regression on operational losses. However, we note that no such results were observed by the extended regression results obtained by Dahlen and Dionne (2010). We complete a similar test for ELD to compare the results between data sources.

Table 5.11: ILD Kendall's τ rank correlation hypothesis test

ln(Loss)	test statistic	p -value	τ -estimate
Assets	-09.032	< 2.20e-16	-0.05165
Revenue	-22.937	< 2.20e-16	-0.13117
Equity	-33.846	< 2.20e-16	-0.19356

External Loss Data

For our ELD we see positive correlation between all numeric size indicators and our response variable. Specifically we note lower-left panels all have a positive slope indicating positive correlations. In terms of the colouring and the pie charts' shading, we can see that the explanatory variables are all highly correlated with each other; i.e. we can therefore only use one of the variables as representative of the others.

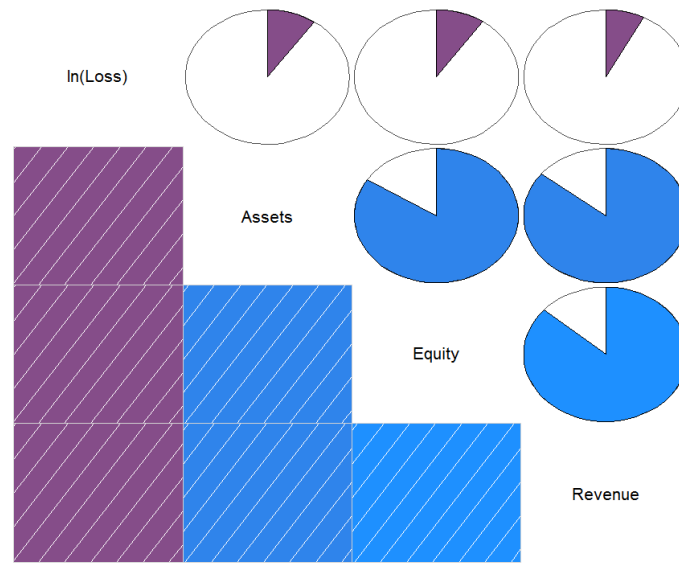


Figure 5.9: Correlograms of the ELD numeric proxies

The association between our response variable and the ‘Revenue’ explanatory variable appears to be the weakest - see Figure 5.9. The respective relationships for ‘Assets’ and ‘Equity’ appear to be equally strong.

We also include the correlation matrices for Kendall’s τ and Spearman’s ρ rank correlation below. From both Tables 5.12 and 5.13 we conclude that the strongest relationship between the response variables ‘ln(Loss)’ appears to be with ‘Assets’.

Nonetheless, the relationship as evidenced from the graphical diagnostics and the correlation matrices is clearly not a strong one, indicated by the purple hue. We especially reflect on this analysis once we start reviewing our regression results and further testing is done on the coefficients for the size indicator.

Table 5.12: ELD Kendall’s τ rank correlation

	ln(Loss)	Revenue	Assets	Equity
ln(Loss)	1	0.063	0.063	0.062
CurRev	0.063	1	0.831	0.837
Assets	0.063	0.831	1	0.847
Equity	0.062	0.837	0.847	1

As mentioned in the preceding section we conduct further inspection into the hypothesis testing around our dependence structure estimation. We perform a test of association between the sampled pairs. We again perform this test at a confidence interval of 99% and specify the calculation method of Kendall’s τ for a two-sided test. This test thus performs inference on an interval to determine whether we can reject the null hypothesis that the dependence measure - i.e. our correlation values - that said interval contains the value 0; thereby rendering the concomitant parameter estimation useless.

Table 5.13: ELD Spearman's ρ rank correlation

	ln(Loss)	Revenue	Assets	Equity
ln(Loss)	1	0.092	0.093	0.092
CurRev	0.092	1	0.958	0.960
Assets	0.093	0.958	1	0.967
Equity	0.092	0.960	0.967	1

The results of the testing are presented in Figure 5.14. It is clear from all tests' small p -value that the null hypotheses cannot be rejected. Note that the τ estimates are now positive for all pairwise testing which mirrors ours expectation for this proxy relationship.

Table 5.14: ELD Kendall's τ rank correlation hypothesis test

ln(Loss)	test statistic	p -value	τ -estimate
Assets	4.1277	3.67e-05	0.062835
Revenue	4.1891	2.80e-05	0.063771
Equity	4.0962	4.20e-05	0.062357

5.4.5 Analysis of the distributional forms

For the analysis of the factor variables in the subsequent sections we consider our response variable, operational losses, for each level of the factor variables. We perform this analysis for both the ILD and ELD. On face value factor variables provide minimal information prior to modelling. For this reason we inspect the distributional forms of the response variable per factor level.

In order to investigate these distributional forms of the responses variable we make use of *violin plots* in combination with box plots. Violin plots when rotated in an oblong manner clearly delineate the distributional form of a variable; i.e. one can develop a clear view of the location, scale, skewness and kurtosis of the variable under investigation. We combine these with box plots to explicitly map the main quartiles of the variable which are not included in the normal format of violin plots.

Note that in our analysis, we retain the logarithmic transformation of the response variable. Also, the overall ranges of the between the respective ILD and ELD graphs differ dramatically as expected.

5.4.6 The explanatory variables - business line (BL)

Internal Loss Data

In the violin plots of the logarithm of the operational losses in Figure 5.10a, we note comparable skewness and distributional location - given that we are using a

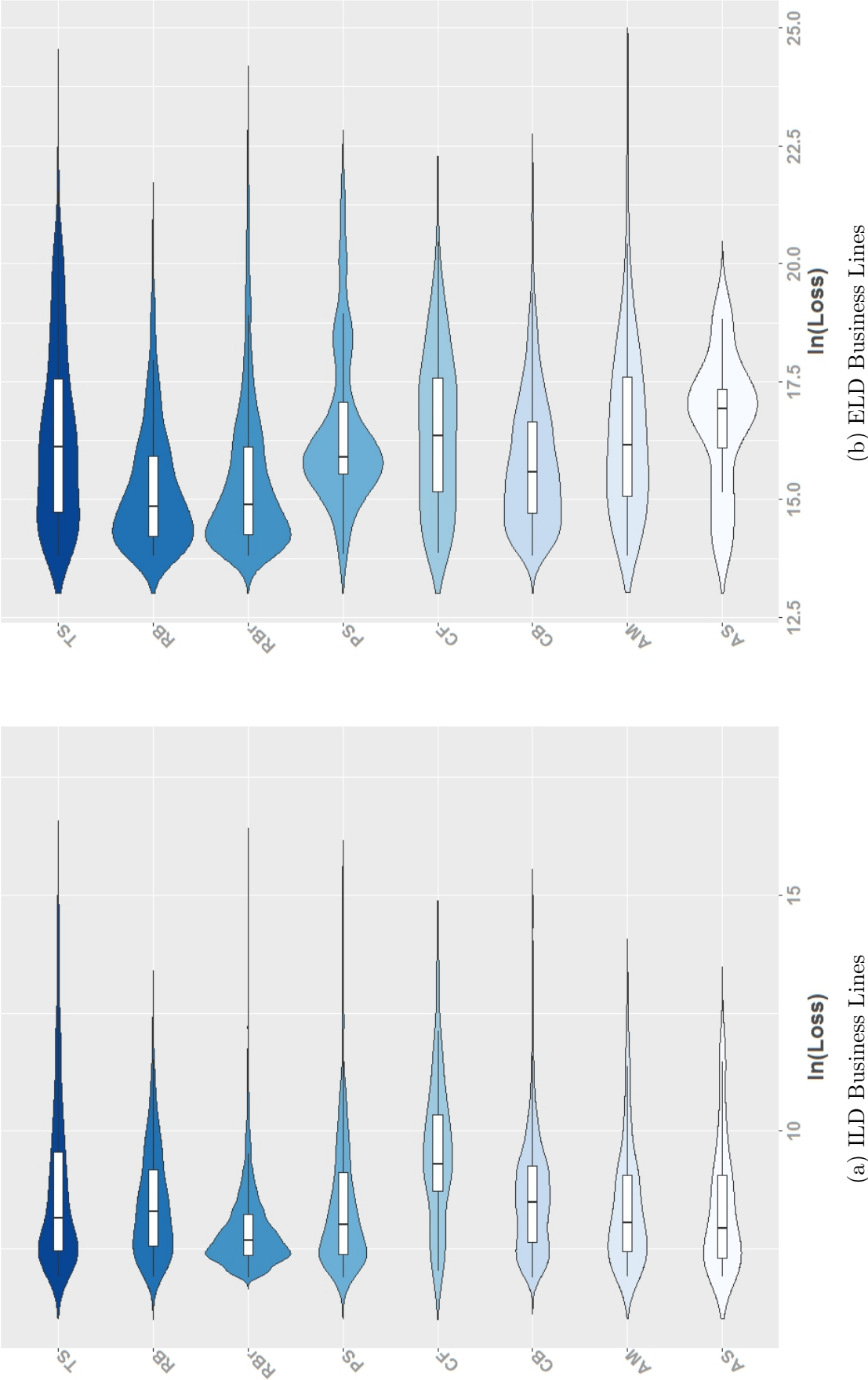


Figure 5.10: Distributional forms of the Business Lines variable

logarithmic transformation. For the BL levels of CF and CB we note a deviation from the expected distributional forms; i.e. the clear positive skewness is still present, but significant kurtosis makes for leptokurtic appearances. Specifically for the case of CB it appears as though the distribution is bimodal with peaks at both 7.5 and 8.75.

For levels TS and RB we see similar leptokurtic tailedness, with the latter being more pronounced. Level PS appears significantly more platykurtic, however, closer inspection of the tail indicated clear losses much further into the upper ranges than the other levels.

Lastly, we note the clear peakedness of level RBR - Retail Brokerage. However, relating this to our earlier inspection of the number of loss amounts and the actual loss amounts, we recall that this level has very few observations.

External Loss Data

In Figure 5.10b we see the same complex combination of distributional attributes. Save for level AS the expected skewness is seen for the distributional forms. Again, we see bimodal forms among the distributions; namely those of PS and AS. Level AS has a highly counter-intuitive form, however, from earlier analysis we noted that this level has very low representation in terms of its frequency and severity.

Among the distributional forms for the ELD leptokurtic tailedness seems to be the norm, especially CF and AM - Asset Management - have very gradual slopes in the higher ranges of the response variable. Also, the locations of the distributions per factor level see a much wider range; from just below 15 through almost 17.5.

5.4.7 The explanatory variables - event type (ET)

Internal Loss Data

When viewing the violin plots of the response variable per factor level - i.e. the ET levels - we see in Figure 5.11a that these distributional forms mostly conform to our expectations of positively skewed distributions. Also for this factor variable we see the narrow range of distribution means as with the BLs' distributions.

Levels IF - Internal Fraud - and EPWS - Employment Practices and Workplace Safety - have very low peakedness. For EF we note a core concentration around the first quartile of the distribution. Given the lower range peakedness, closer inspection confirmed that the thin tail extends far into the higher ranges of the response variable. Similar, but shorter, thin tails are also noted for CPBP and BDSF - Business Disruption and System Failure.

External Loss Data

Most of the ET level distributional forms of the ELD as illustrated in Figure 5.11b are aligned to our intuition of positive skewness. Level BDSF almost appears to be a trimodal distribution and for most distributions we see significant leptokurtic tailedness of the response variable.

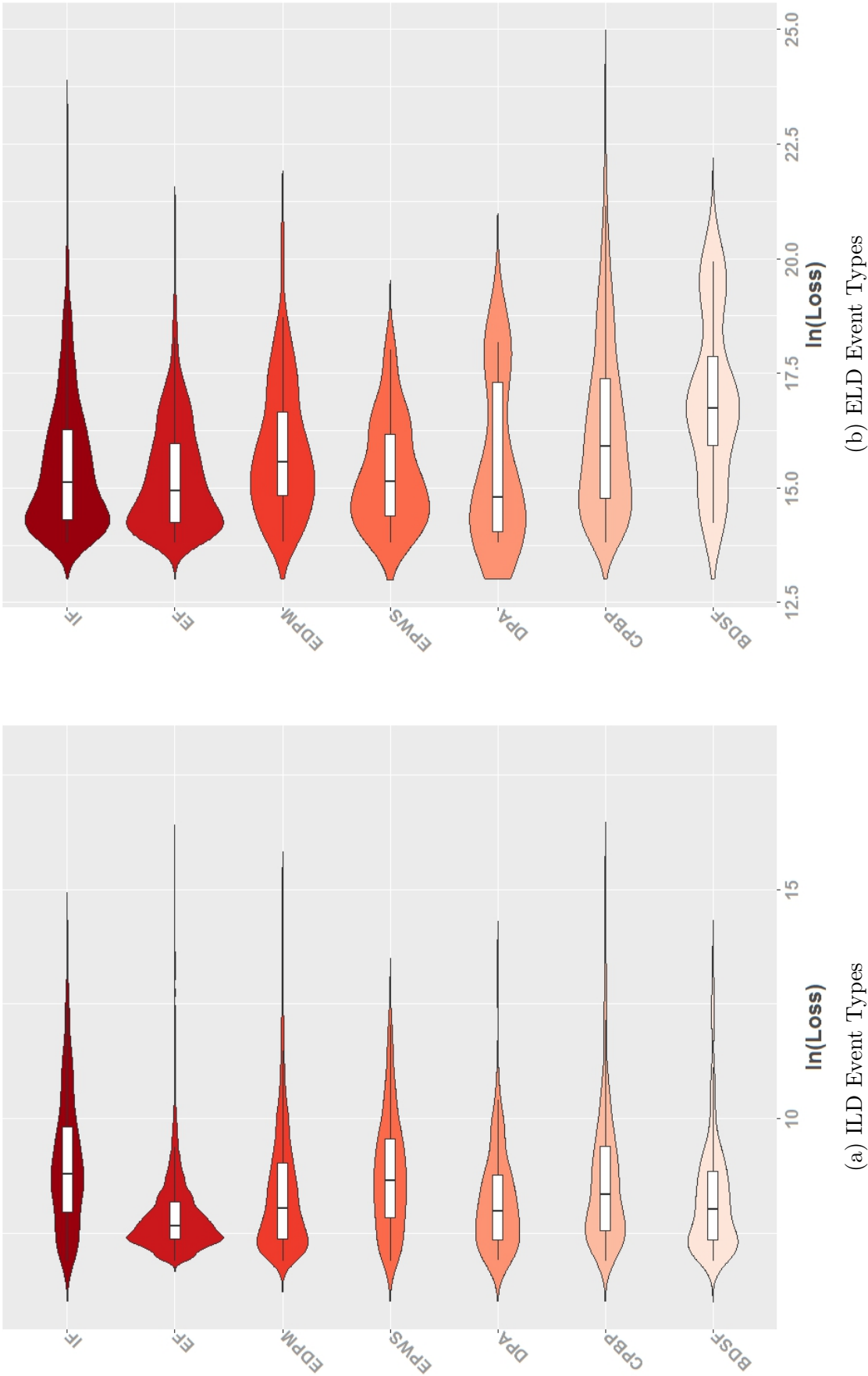


Figure 5.11: Distributional forms of the Event Types variable

We also observe especially long tails for CPBP, DPA - Damage to Physical Assets, and IF. Despite said tail structures, the distribution means exhibit behaviour comparable to that seen for the BL levels.

5.4.8 The explanatory variables - region

External Loss Data

Given that our ILD only originates from a single region - Africa - we only analyze the ELD for regional variation in terms of our response variable - see Figure 5.12.

Before analysing the distributional forms for the factor levels we consider a few background points on the definitions of the actual factor levels. Firstly, the level North America consists only of the USA, Mexico and Canada. Secondly, the factor level Other Americas therefore include Central (excluding Mexico) and South America into a single factor level. Note that this includes Caribbean island territories of sovereigns listed as North America or Europe. Finally, factor level Other therefore includes Australia, New Zealand, Oceania and Pacific Ocean island territories of sovereigns listed under North America, Europe, or Asia.

For the region factor level distributional forms of the response variable we see a high level of conformance in terms of our experience of positively skewed loss distributions. The only exception is factor level Other Americas. We almost see a type of bimodal form towards the end stretch of the distribution's tail, however, in context of the other distribution we assume this to be a very leptokurtic tail which is not clearly delineated for the period in question.

This distribution is also unique in terms of the distribution mean. The other factor level distributions show high correspondence with respect to their means save for Other Americas. In combination with the tailedness this factor level should receive closer inspection during our regression modelling to see its effects on the overall model process.

5.5 Chapter conclusion

5.5.1 Overview

In this chapter we focused on our data environment. We analyzed the structure of the ILD and ELD available to us. We noted special restrictions on data and how to filter data for general operational risk capital modelling. Our focus then centred on the response variable, the operational losses, which is the variable we attempt model and thus predict when calculating for operational risk capital. The comparative nature of the analysis of the response variable - i.e. comparing the variable's structure between the ILD and ELD respectively - set the scene for the continued comparative type analyses we performed on the explanatory variables.

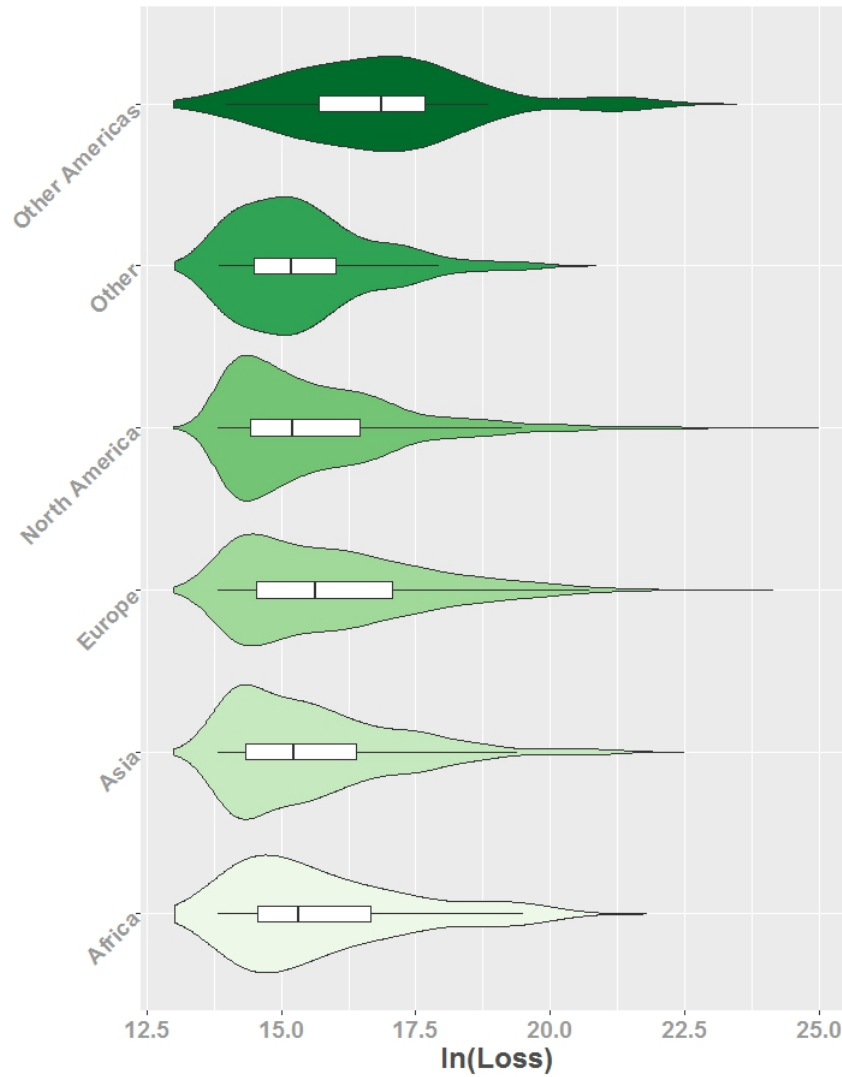


Figure 5.12: Distributional forms of the Regions variable

In this section we recap and summarize our conclusion for the aforementioned points. We retrace the analyses' results and relate these to Problem I which we subsequently attempt to solve.

5.5.2 Data analysis

General overview

We firstly discussed our sources of ILD and ELD. Our hosting institution shared its ILD provided that anonymity of the data be retained. This we ensured by means of applying an undeclared linear transformation; thus keeping most of our variables' distributional attributes. Our ELD is sourced from the general public database as compiled and maintained by SAS®. In finalizing our data preparation we adjusted the numeric variables for cumulative inflation, and in the case of the ILD also the exchange rate to allow for direct comparison. We subsequently performed separate filtering of two datasets and noticed that the filtering are not only aligned

to international best practise and regulatory standards, but also compare favourably to each other in structure.

Next we reviewed the basic statistical characteristics of our response variable, operational losses, for both the ILD and ELD. Prior to our investigation we already knew of the scale issues (at a minimum) in combining ILD and ELD. This is also our core definition of Problem I as set out in Section 1.1.1. When finally illustrating the response variable and determining basic descriptive statistics it became clear just how *incomparable* the two datasets are; i.e. we could see that merely in terms of location and scale the two datasets are highly unique. Our graphic analysis confirmed comparable skewness, but further graphical diagnostics confirmed that kurtosis elements are also highly different. This analysis allowed us to confirm that a solution to the scaling problem will address the issues of the divergent distributional attributes.

In the remaining section we continue with the practice of analysing variable, whether numeric or factor level type, in terms of the response variable. Specifically in Sections 5.4.2 and 5.4.3 we note the aforementioned descriptive statistics of the respective datasets. In order to better gauge the response variable's resulting behaviour due to our expected explanatory, we consider bar charts of the ILD and ELD. These bar charts delineate the frequency and severity of the operational losses respectively. When comparing the bar charts of the ILD against those of the ELD the scale difference is naturally an aspect to consider, however there are clear difference in which factor levels of the explanatory drive the response variable in terms of frequency and severity. As we see in our subsequent solution attempt (Section 6.2) these unique attributes of the two distinct datasets are certainly not a deleterious characteristic. Our intuition is that this unique characteristic allows us to not only scale losses down (or up), but more specifically scale the loss as appropriate for the entity in question; i.e. our host institution. This implies that a similar exercise carried out at another financial institution in South Africa will result in scaled losses reflecting the South African environment but also the institution scaling the losses.

As mentioned, we expect the risk environment for South Africa to be different; our overall South African business models' resulting business practices are closely aligned to those seen in North America and Europe. This becomes apparent when reviewing the results for the factor levels for the different explanatory variables. With respect to the loss frequencies, we see that for BL levels the results between ILD and ELD are comparable. The leading driver is RB followed by CB. For ET levels ELD has a very high concentration under CPBP which is absent for ILD. Otherwise, ILD and ELD both see a noteworthy number of losses for EF and EDPM. We can therefore conclude that in terms of loss frequency, the South African environment is similar to the international one. For the ET driver of loss frequency the absence of CPBP points to a clear difference between business models. When considering loss severity the comparison become more divergent for both BL and ET levels. In terms of BL levels, ELD has greater concentration under TS whereas this does not feature in ILD at all. ILD sees high concentration for RB which is also present in ELD. However, in terms of ET levels, CPBP is the most significant level for ELD yet it ranks much lower for ILD.

When reviewing the possible proxies for size indicators - the numeric variables - the ILD and ELD provided highly divergent results. For the ILD, the proxies do

not seem to be appropriate measures owing to the negative correlations. However, the ELD results follow our intuition that increasing size indicators is associated with increasing losses - in both frequency and severity. This forms an integral part of our inspection on parameter estimation during regression modelling. The explanatory variables' distributional forms showed unexpected results in terms of bimodal forms. However, it was clear from many of the empirical distributions that the leptokurtic attribute is a core underlying factor especially in the of ELD; this follows our expectation that ELD have more losses concentrated in the higher ranges. We specifically desire this as we want to also scale this attribute down to our ILD.

Data in the regression context

Our next step is to use the ILD and ELD in our suggested solution to Problem I. We see how a scaling mechanism becomes possible when combining the conceptual idea of the loss ratio as in Equation 3.18 with the extended explanatory variable set from Dahlen and Dionne (2008; 2010), and then reformulating this model for quantile regression. Once the basic regression model is underway, we include the refinements as discussed in Chapter 4.

Chapter 6

Proposed solution

6.1 Introduction

In this chapter we can finally incorporate our literature study and subsequent conclusions, with the data analytics we completed on our ILD and ELD. By performing various forms of statistical regression we attempt to produce a scaling mechanism whereby internationally-driven and derived ELD can be scaled down to be used *directly* with the more modest - i.e. small scale - ILD.

We firstly revisit our literature studies and adapt inter alia definitions, concepts, and model terms to suit our data environment; and more importantly to align our initial regression work to the later quantile regression work. This overview is concluded with a discussion on the encoding of our factor variables for the DD-model and our proposed extension of it.

The remainder of the chapter is divided into a discussion of the underlying regression processes and the subsequent impact study on capital estimation of our resulting regression solution.

Our initial foray in the regression processes is graphical diagnostics where now see a real-world situation, and specifically, a clear illustration of the benefits of using quantile regression rather than OLS where appropriate. The diagnostics allow us to visualize the distinctive approaches of the respective OLS and quantile regression techniques. We then provide detailed discussions on the regression of our ILD and ELD; both OLS and quantile regression. These are assisted by tabulated results of the regressions.

After having defining our final models for ILD and ELD respectively, we apply the scaling mechanism to our ELD. In this process there are many unique application challenges and these are discussed with suggested solutions in detail. Once the scaling has been completed we perform further diagnostics and data analytics on the new dataset to illustrate our proposed solution.

We close our proposed solution by investigating its effects on capital estimation. We consider the differentiating trend of pooling our ILD with the scaled ELD and its implications when calculating undiversified operational risk capital.

6.2 Literature in practice

6.2.1 Conceptual overview of the scaling process

Recall the scaling expression as derived in Equation 3.18:

$$Loss_B = Loss_A \frac{Size_B^\alpha \ln(F(\omega, \theta_B))}{Size_A^\alpha \ln(F(\omega, \theta_A))} . \quad (6.1)$$

The subscript A and B indicate the type of loss data used in the scaling process. Subscript A is ELD, whilst subscript B indicates ILD. In terms of the scaling formula we identify that we already have the information to populate the terms $Loss_A$ and $Loss_B$; i.e. respectively with the loss information ELD and ILD. The formula term $Size_B^\alpha \ln(F(\omega, \theta_B))$ is the regression model using ILD, whilst $Size_A^\alpha \ln(F(\omega, \theta_A))$ is analogous for ELD.

Once we slot in the loss data information, along with the regression estimates are obtained from our quantile regression (see Section 6.2.4), we can redefine our scaling formula as

$$Loss_B = Loss_A \frac{\hat{Y}_B}{\hat{Y}_A} . \quad (6.2)$$

As the final step of our scaling process we use the respective loss estimates for ILD and ELD, and apply it as a ratio to the known ELD value ($Loss_A$) to scale it down to a value to be included in the known ILD loss dataset ($Loss_B$). We define the ratio components as component A - ($\hat{Y}_A$) for ELD, and component B - ($\hat{Y}_B$) for ILD.

6.2.2 Components

We consider the generic formula structure for the loss model estimate $\hat{Y}_i$. The loss estimate model is defined as

$$\hat{Y}_i = Size_i^\alpha \ln(F(\omega, \theta_i)) . \quad (6.3)$$

Following from the definitions provided by Dahen and Dionne (2010) in equations 3.16 and 3.17, we consider the loss estimate model as

$$\ln(\hat{Y}_i) = \alpha \cdot \ln(Size_i) + \sum_j \beta_{ij} BL_{ij} + \sum_k \delta_{ik} ET_{ik} . \quad (6.4)$$

6.2.3 Variable encoding

From the aforementioned model we specifically note the model terms of $\sum_j \beta_{ij} BL_{ij}$ and $\sum_k \delta_{ik} ET_{ik}$, with β_{ij} and δ_{ik} indicating the coefficients to be estimated. We reserve some attention for these model terms since they form a pivotal point for our later matching of quantile regression models.

Our initial regression of these factor variables consisted of including the variable as a single term and then regressing with the single variable per factor level. In order to gauge the most important factor levels we re-programmed the factor levels using

the general dummy variable treatment, omitting the base case of one factor level when performing the regression.

We provide an example of a single level of each respective factor variable, BL and ET, to illustrate our re-programming of the factor variable Business Line

$$\text{BL.AM} = \begin{cases} 1 & \text{if Business Line} = \text{Asset Management} \\ 0 & \text{if Business Line} \neq \text{Asset Management}, \end{cases} \quad (6.5)$$

whilst considering the factor variable Event Type

$$\text{ET.DPA} = \begin{cases} 1 & \text{if Event Type} = \text{Distruction to Physical Assets} \\ 0 & \text{if Event Type} \neq \text{Distruction to Physical Assets}. \end{cases} \quad (6.6)$$

6.2.4 Regression process - ILD

Introduction

For the regression procedure we follow the model setup as defined in Dahan and Dionne (2010). We subsequently use this regression setup for our quantile regression specifying even quantiles of 10% increments. As part of the variable selection during the iterative regression process we base our decision making on common p -values, conditioning on a 95% confidence interval.

The regression process

We now perform the regression on our ILD. Firstly, we perform a normal OLS regression; i.e. regressing for the expected value, mean value. Secondly, we perform the quantile regression for the range 5% through 95%; i.e. q_θ where $\theta \in [0.05; 0.95]$, in intervals of 5% increments¹.

With respect to the notation we follow for the reporting of our results, we see that the model terms are the factor levels of the two factor variables we use in the ILD OLS model. Therefore we note the respective levels of the factor variables ‘Business Lines’ (BL) and ‘Event Type’ (ET). Following from this notation principle, we note that one level of the factor variables each, is missing.

The missing, or rather, omitted level follows the encoding principle when isolating the levels of a factor variable for a regression model. This omitted level is normally referred to as the *base case*. For the base cases of both factor variables we use the omission as specified in the earlier work of Dahan and Dionne (2008:20). We therefore use the level ‘Payment and Settlements’ under BLs, and ‘Business Disruptions and System Failures’ under ETs.

Finally, we also point out that we do not include the factor variable ‘Region’ for the ILD regression since it consists of a single value across all observations, thereby leading to a singular matrix of parameter estimates in the regression model.

The results are presented in Tables 6.1 and 6.2. Under model (1) we see results of an OLS regression based on the original regression model as suggested by Shih

¹Note that we are initially considering equidistant intervals among the quantiles. As part of the further regression process we consider more detailed analyses across the higher quantiles.

Table 6.1: Exploratory OLS regression

	<i>Dependent variable: lCurLoss</i>		
	(1)	(2)	(3)
log(CurRev)	−2.949*** (0.146)	−2.694*** (0.142)	−2.674*** (0.140)
BL.CF		1.064*** (0.161)	
BL.TS		0.287*** (0.064)	
BL.RB		−0.549*** (0.044)	
BL.CB		0.109** (0.051)	
BL.AS		−0.147 (0.126)	
BL.RBR		0.006 (0.078)	
BL.AM		0.013 (0.105)	
ET.CPBP			0.403*** (0.068)
ET.DPA			−0.064 (0.105)
ET.EPWS			0.531*** (0.077)
ET.EDPM			0.097** (0.049)
ET.EF			−0.409*** (0.046)
ET.IF			0.741*** (0.069)
Constant	72.951*** (3.217)	67.798*** (3.126)	67.176*** (3.084)
Observations	15,684	15,684	15,684
R ²	0.025	0.090	0.107
Adjusted R ²	0.025	0.090	0.107
Residual Std. Error	0.950 (df = 15682)	0.918 (df = 15675)	0.909 (df = 15676)
F Statistic	407.604*** (df = 1; 15682)	193.962*** (df = 8; 15675)	269.496*** (df = 7; 15676)

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 6.2: Reduced OLS regression for ILD

	<i>Dependent variable: lCurLoss</i>	
	(4)	(5)
log(CurRev)	−2.491*** (0.138)	−2.498*** (0.138)
BL.CF	0.979*** (0.157)	0.989*** (0.156)
BL.TS	0.174*** (0.063)	0.181*** (0.060)
BL.RB	−0.388*** (0.044)	−0.377*** (0.041)
BL.CB	0.178*** (0.050)	0.190*** (0.047)
BL.AS	−0.321*** (0.123)	−0.312** (0.122)
BL.RBR	−0.335*** (0.079)	−0.322*** (0.077)
BL.AM	−0.086 (0.102)	
ET.CPBP	0.465*** (0.069)	0.484*** (0.066)
ET.DPA	−0.114 (0.104)	
ET.EPWS	0.515*** (0.077)	0.536*** (0.074)
ET.EDPM	0.097** (0.049)	0.117*** (0.044)
ET.EF	−0.255*** (0.047)	−0.235*** (0.042)
ET.IF	0.894*** (0.068)	0.915*** (0.065)
Constant	63.330*** (3.045)	63.451*** (3.041)
Observations	15,684	15,684
R ²	0.141	0.141
Adjusted R ²	0.140	0.140
Residual Std. Error	0.892	0.892
	(df = 15669)	(df = 15671)
F Statistic	183.813*** (df = 14; 15669)	214.287*** (df = 12; 15671)

Note:

*p<0.1; **p<0.05; ***p<0.01

et al. (2000); i.e. a model consisting exclusively of a size indicator - in our case the logarithmic transformation of the firm's current revenue. Shih et al. (2000) obtained an adjusted R^2 of 5.4%. Our regression obtained a value of 2.5%. Models (2) and (3) are not formally part of our regression procession reasoning, but we include these here to illustrate the predictive ability contained in the respective factor variables of BL and ET. Note the BL terms do not perform as well as the ET terms overall; neither the statistical significance of parameter estimates nor the adjusted coefficient of determination.

Adjusting for regression results

Our regression process for the proposed solution continues in model (4) where we see the results of the initial OLS regression. This model is the *full* model containing all the possible model terms, given the omission of the base cases. We note that the model terms BL.AM and ET.DPA are not statistically significant at our lowest level of inclusion - 95% confidence level. An interesting point to mention here is that the statistical significance of BL factor levels has not markedly increased in comparison to their regression in isolation - see model (2).

We therefore exclude these terms in a second regression as reported under model (5). In the aforementioned results we note that all the model terms are statistically significant at levels of $\alpha = 1\%$. Interestingly we note that there was no change in the standard or adjusted coefficients of determination. In this instance we focus on the results handling as an instance of parsimony; it is desirable to utilize a reduced model when the interpretation thereof is reasonably sound.

We thus conclude model (5) as our final OLS regression model for the ILD. This model conclusion is noteworthy since we are using it as a guide for our quantile regression. This means that the model terms combination as set out in model (5) is the base regression form for our quantile regression.

Furthermore, in the incidence of statistically nonsignificant parameter estimates (i.e. at a confidence level of 95%) in our quantile regression we are slotting in the OLS regression results for the parameter estimate in question; effectively creating our version of *hybrid* quantile regression models. We follow this adjustment since a parameter estimate originating from the OLS regression is naturally based on the full set of data; i.e. not the subsetting inherently part of quantile regression.

Quantile regression

In Table 6.3 we see the quantile regression results for our loss estimate regression model. We include the OLS regression results for comparison.

When reviewing the regression results of the parameter estimates it is clear that the vector form for estimates of OLS regression is now translated to a matrix form for quantile regression. We see the same model terms and their coefficients, but these differ across the various quantiles for which we had performed a regression. The quantile regression estimates show the confidence level at which the coefficients pass for statistical significance. We retain the formal cut-off of a 95% confidence level.

Within our context of focusing on the tail of the loss distribution, the next step in our proposed solution is to perform a regression for higher quantiles in greater

Table 6.3: Quantile regression: reporting on deciles of ILD

$\theta =$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	OLS
log(CurRev)	-1.704*** (0.088)	-2.054*** (0.084)	-2.229*** (0.098)	-2.568*** (0.111)	-2.951*** (0.129)	-3.099*** (0.158)	-3.357*** (0.202)	-3.680*** (0.234)	-3.548*** (0.345)	-2.498*** (0.138)
BL.CF	0.481*** (0.029)	0.476 (0.325)	1.432*** (0.272)	1.475*** (0.161)	1.495*** (0.244)	1.441*** (0.086)	1.365*** (0.340)	0.883 (0.540)	1.090* (0.561)	0.989*** (0.156)
BL.TS	-0.022 (0.051)	-0.037 (0.044)	0.021 (0.056)	-0.043 (0.084)	0.050 (0.113)	0.256* (0.137)	0.288 (0.192)	0.419** (0.199)	0.684** (0.290)	0.181*** (0.060)
BL.RB	-0.080*** (0.028)	-0.060*** (0.013)	-0.090*** (0.035)	-0.127** (0.054)	-0.147*** (0.053)	-0.184*** (0.069)	-0.350*** (0.118)	-0.609*** (0.114)	-0.789*** (0.202)	-0.377*** (0.041)
BL.CB	-0.002 (0.032)	0.114*** (0.031)	0.249*** (0.062)	0.447*** (0.071)	0.509*** (0.064)	0.561*** (0.082)	0.461*** (0.125)	0.274** (0.123)	0.112 (0.216)	0.190*** (0.047)
BL.AS	-0.128*** (0.039)	-0.158 (0.097)	-0.274*** (0.069)	-0.272 (0.198)	-0.205 (0.126)	-0.235 (0.356)	-0.117 (0.378)	-0.426 (0.513)	0.173 (0.660)	-0.312** (0.122)
BL.RBR	0.075 (0.065)	0.063 (0.067)	0.015 (0.080)	-0.003 (0.123)	0.098 (0.092)	0.003 (0.088)	-0.277 (0.251)	-0.569*** (0.173)	-0.998*** (0.247)	-0.322*** (0.077)
ET.CPBP	0.074 (0.066)	0.138** (0.060)	0.226*** (0.073)	0.257** (0.113)	0.250** (0.108)	0.480*** (0.130)	0.696*** (0.155)	0.940*** (0.172)	1.210*** (0.277)	0.484*** (0.066)
ET.EPWS	0.103** (0.052)	0.329** (0.149)	0.601*** (0.095)	0.638*** (0.125)	0.658*** (0.127)	0.753*** (0.119)	0.764*** (0.156)	0.879*** (0.163)	0.795*** (0.201)	0.536*** (0.074)
ET.EDPM	-0.019 (0.024)	-0.002 (0.029)	0.043 (0.043)	0.064 (0.079)	0.010 (0.053)	0.100 (0.074)	0.191*** (0.072)	0.322*** (0.087)	0.486*** (0.154)	0.117*** (0.044)
ET.EF	0.062*** (0.024)	0.055** (0.027)	0.019 (0.039)	-0.072 (0.076)	-0.213*** (0.048)	-0.241*** (0.068)	-0.293*** (0.061)	-0.332*** (0.077)	-0.494*** (0.133)	-0.235*** (0.042)
ET.IF	0.170** (0.066)	0.467*** (0.108)	0.852*** (0.121)	0.863*** (0.092)	0.893*** (0.097)	1.043*** (0.119)	1.248*** (0.150)	1.514*** (0.227)	1.767*** (0.195)	0.915*** (0.065)
Constant	44.708*** (1.942)	52.546*** (1.859)	56.583*** (2.160)	64.286*** (2.447)	73.019*** (2.841)	76.484*** (3.486)	82.592*** (4.451)	90.262*** (5.151)	88.117*** (7.595)	63.451*** (3.041)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6.4: Quantile regression: reporting on deciles of ILD

$\theta =$	0.95	0.96	0.97	0.98	0.99	OLS
log(CurRev)	-3.523*** (0.497)	-3.700*** (0.642)	-3.719*** (0.619)	-3.502*** (0.768)	-3.368*** (0.737)	-2.498*** (0.138)
BL.CF	0.914 (1.670)	0.424 (0.576)	-0.236 (1.037)	-0.411 (0.636)	-1.171** (0.531)	0.989*** (0.156)
BL.TS	0.660* (0.396)	0.413 (0.644)	0.205 (0.623)	0.578 (0.744)	0.065 (0.521)	0.181*** (0.060)
BL.RB	-1.015*** (0.357)	-1.312** (0.568)	-1.569*** (0.417)	-1.760*** (0.580)	-1.966*** (0.515)	-0.377*** (0.041)
BL.CB	0.014 (0.362)	-0.342 (0.586)	-0.551 (0.425)	-0.800 (0.617)	-0.743 (0.617)	0.190*** (0.047)
BL.AS	-0.115 (0.546)	-0.420 (0.669)	-0.920* (0.512)	-1.436** (0.587)	-2.196 (1.437)	-0.312** (0.122)
BL.RBR	-1.395*** (0.442)	-1.796*** (0.662)	-2.239*** (0.456)	-2.601*** (0.611)	-3.329*** (0.519)	-0.322*** (0.077)
ET.CPBP	1.319*** (0.458)	1.263* (0.697)	1.506*** (0.495)	1.684** (0.797)	0.848*** (0.288)	0.484*** (0.066)
ET.EPWS	0.452 (0.349)	0.045 (0.517)	-0.351 (0.222)	-0.141 (0.837)	-1.007** (0.441)	0.536*** (0.074)
ET.EDPM	0.418 (0.311)	0.186 (0.533)	0.173 (0.232)	0.354 (0.781)	-0.415 (0.306)	0.117*** (0.044)
ET.EF	-0.897*** (0.294)	-1.192** (0.516)	-1.363*** (0.223)	-1.239 (0.766)	-2.119*** (0.304)	-0.235*** (0.042)
ET.IF	1.368*** (0.328)	1.049* (0.563)	0.880** (0.378)	1.138 (0.790)	0.267 (0.330)	0.915*** (0.065)
Constant	88.617*** (10.970)	93.246*** (14.166)	94.341*** (13.640)	89.905*** (16.953)	88.506*** (16.227)	63.451*** (3.041)

Note: *p<0.1; **p<0.05; ***p<0.01

detail. We therefore specifically regress on percentiles between the quantiles of 0.95 and 0.99; i.e. $\theta \in [0.95; 0.99]$. We report these coefficient estimates in Table 6.4.

Recall that our quantile regression is on every fifth percentile up to 95%. Our underlying procedures strictly aligns to this process, but for our interpretative discussion here we only report on deciles; i.e. every second 5% percentile. When considering the regression results of the initial quantile regression only, we note different behaviour for most of the respective variables. We review these in summary here.

Concluding from Table 6.3, for the variables BL.RB, ET.IF and ET.EPWS we note that the parameter estimates are statistically significant across all the quantiles. The other variables have varying results. Variables BL.TS, BL.RBR and ET.EDPM see statistical significance in their parameters estimates on the higher end of the quantile ranges; i.e. at least greater than 80%. The remained of the variables show clear statistical significance across most of their quantile range with only a few quantile estimates not being statistically significant.

If we consider $q_{0.95}$ as the lower threshold for so-called *high* quantiles taxonomy - see Table 6.4, it is only variables BL.CB and ET.EDPM which have statistically nonsignificant parameters estimates for any of the high quantiles. Once we consider these quantile regression results at the very high quantiles specifically, a clearer conclusion forms on the usefulness of quantile regression for operational loss modelling. When regressing for the percentiles from 0.95 through 0.99 we see definite statistical significance for the parameter estimates across the variables.

Following our reasoning of slotting in OLS regression estimates when we do not have statistically significant quantile regression estimates, we compile our ILDD quantile regression model - i.e. our extended DD-model - as follow from Equation 6.4

$$\ln(\hat{Y}_{\theta i}) = \alpha_{\theta i} \cdot \ln(Size_{\theta i}) + \sum_j \beta_{\theta ij} BL_{\theta ij} + \sum_k \delta_{\theta ik} ET_{\theta ik}, \quad (6.7)$$

where $\theta \in [0.05, \dots, 0.95] \cup [0.96, \dots, 0.99]$; and $i \in 1, \dots, n$ represents our dataset of losses. Recall that j and β sums over the factor levels of the Business Lines variables and it follows analogously for k and δ for the Event Type factor levels.

Graphical diagnostics

For the graphical diagnostics of our regression we plot the parameter estimates across quantiles for each of the model variables in isolation. The quantile configuration which we are using for these plots consist of equidistant 5% quantiles starting at 5% and ending on 95%. The quantiles are indicated as a dot-and-dash line with the greyed out sections around the quantiles representing standard error bands around the quantile parameter estimates as we see in Figure 6.1. Along with the plotted quantiles we also include the OLS regression. This line is indicated as the solid, horizontal red line.

The individual model variables exhibit various behaviours. Some variables have gradual changing curves either increasing or decreasing across the quantiles - e.g. the intercept, ET.CBPB. Other show sharp changes in trend in higher quantiles - e.g. BL.RBR, ET.EF. For variables BL.CF and BL.AS we note very stable trends

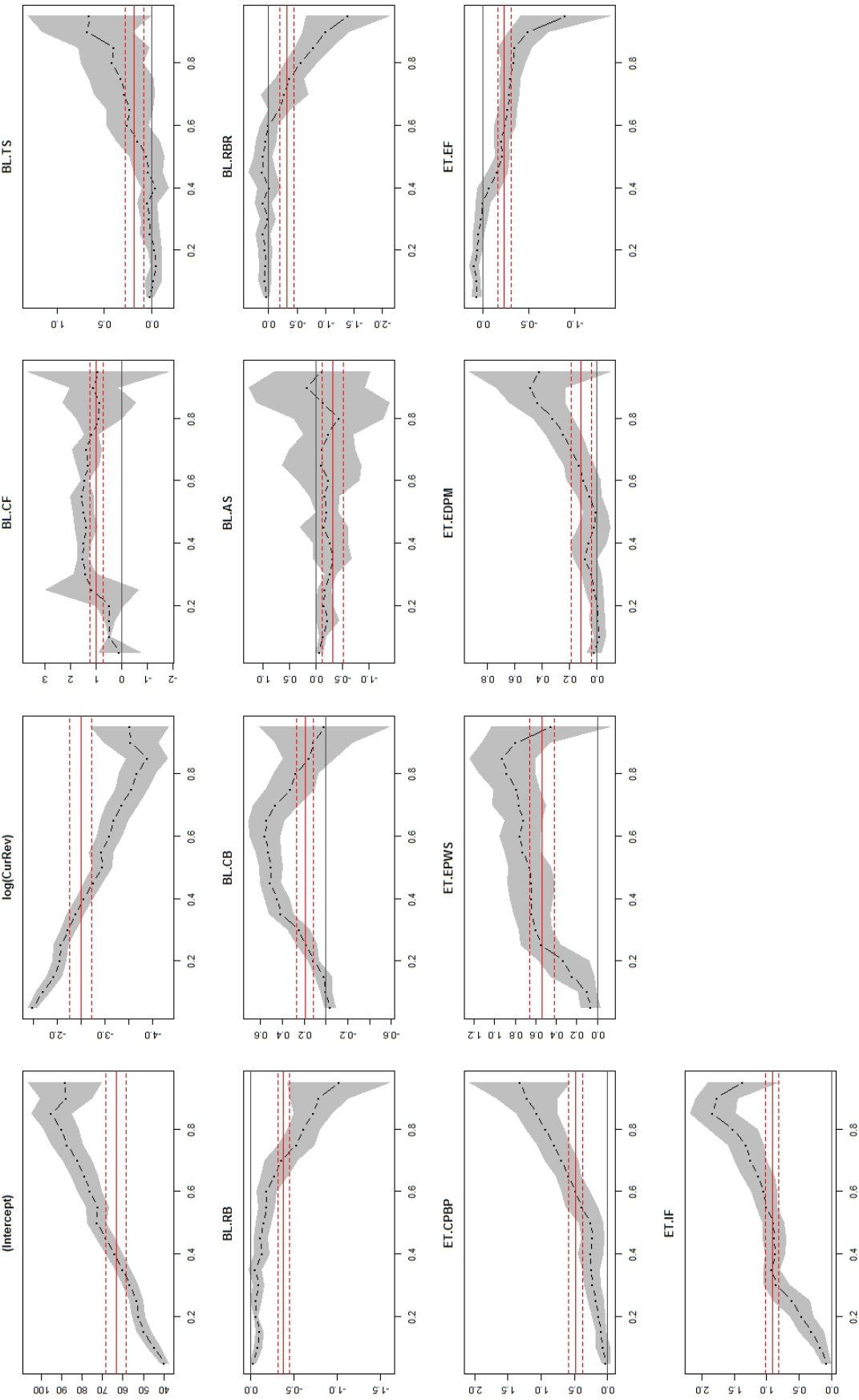


Figure 6.1: Quantile regression on the ILD models' terms

around the OLD estimates, however with sharp, isolated deviation respectively at the beginning and ending of the graphs.

These graphical diagnostics generally follow our preceding statistical analyses. When we specifically look at ET.EDPM, the graph indicates that quantile regression estimates do not seem to be statistically different from the OLS estimates. We compared this view with the parameters estimates in Table 6.3 for the variable. The estimates are only confirmed as statistically significant - in this context *different from the OLS estimates* - from quantile q_{70} onwards.

Overall, our motivation for using quantile regression is consolidated here. Recalling our inspection of the frequency and severity of ILD as provided in Table 5.5, we saw that the variables BL.RB and ET. EDPM are our most significant variables in terms of said metrics. If we cross-reference these variables' results with their concomitant graphical diagnostics, it is clear how significantly the parameter estimate deviates away from the OLS estimate in the higher quantile regions. We can therefore interpret this further that the parameter estimates would be greatly underestimated should normal OLS be applied.

The final model

We now review our final model to be used for determining component B as defined in Equation 6.2. Recall that we are using the extended DD-model we developed whilst slotting in the parameter estimates on quantiles where statistical significance was not present (see Section 3.4.1). We therefore provide the matrix of parameters estimates in tabulated form in Table 6.5.

6.2.5 Regression process - ELD

Introduction

This section follows analogously to that of the preceding regression for ILD, Section 6.2.4. We firstly perform the OLS regression on our ELD retaining the full setup as defined in the preceding section; i.e. with respect to cut-off levels, and the original and by definition, the extended DD-model. Subsequently we return to said extension of this model by using quantile regression; focusing on deciles and finally inspecting very high quantiles.

We discuss all results and ensuing reasoning here, but only present the results of the final quantile regression model. All other regression results are tabulated in Appendix A. However, we first reflect on the inclusion of the factor variable 'Region' which is now possible for the ELD.

Including the variable 'Region'

In this section we first perform OLS regression for the ELD and then the quantile regression. As before, we use level 'Payment and Settlements' under BLs, and 'Business Disruptions and System Failures' under ETs as our base cases as in Dahlen and Dionne (2008).

A significant difference between the ILD regression work and this ELD section lies with the inclusion of the 'Region' variable for the latter's regression. As mentioned

Table 6.5: Final model for ILD

$\theta =$	<i>Dependent variable: $lCurLoss$</i>									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
log(CurRev)	-1.704	-2.054	-2.229	-2.568	-2.951	-3.099	-3.357	-3.680	-3.548	
BL.CF	0.481	0.989	1.432	1.475	1.495	1.441	1.365	0.989	1.090	
BL.TS	0.181	0.181	0.181	0.181	0.181	0.256	0.181	0.419	0.684	
BL.RB	-0.080	-0.060	-0.090	-0.127	-0.147	-0.184	-0.350	-0.609	-0.789	
BL.CB	0.190	0.114	0.249	0.447	0.509	0.561	0.461	0.274	0.190	
BL.AS	-0.128	-0.312	-0.274	-0.312	-0.312	-0.312	-0.312	-0.312	-0.312	
BL.RBR	-0.322	-0.322	-0.322	-0.322	-0.322	-0.322	-0.322	-0.569	-0.998	
ET.CPBP	0.484	0.138	0.226	0.257	0.250	0.480	0.696	0.940	1.210	
ET.EPWS	0.103	0.329	0.601	0.638	0.658	0.753	0.764	0.879	0.795	
ET.EDPM	0.117	0.117	0.117	0.117	0.117	0.117	0.191	0.322	0.486	
ET.EF	0.062	0.055	-0.235	-0.235	-0.213	-0.241	-0.293	-0.332	-0.494	
ET.IF	0.170	0.467	0.852	0.863	0.893	1.043	1.248	1.514	1.767	
Constant	44.708	52.546	56.583	64.286	73.019	76.484	82.562	90.262	88.117	

in Section 6.2.4, the ‘Region’ variable by its very nature, is only applicable to ELD. We include this variable on the reasoning that we want the ELD regression model to capture the variability in operational losses associated with difference in geographical location of financial institutions’ operations.

Concerning the factor variable ‘Region’ we were not able follow the guidance in the Dahlen and Dionne (2008) work. In this study an earlier assessment of the factor level has been completed to determine which factor levels are statistically significant. Within the international financial context, the results obtained by Dahlen and Dionne (2008:17-19) follow our intuition; the main players as indicated by high statistical significance in the parameter estimation results are the United States of America (USA), Canada, and Europe. As is illustrated in this section, our regression results did not follow this same intuitive pattern.

This result may require some further reflection here. Recall that the Dahlen and Dionne (2010) study was centred on the scaling of ELD onto ILD for a financial institution which is based in the USA. Ours is based on a very different setting. The ILD and ELD are nowhere near comparison; i.e. the South African operational losses are not comparable to those of the northern hemisphere; not only in terms of pure magnitude, but scaling aspects of currency rates distorts the juxtaposition further. We can therefore consider our setting to be more *classic* with respect to the differences intuitively envisioned between ILD and ELD. However, given this scenario our intuitive expectations may be misaligned.

Another argument to consider is that the South African environment may indeed be better suited for comparison to other, so-called *smaller* or *less active* regions. As we illustrate in this chapter, we see that there is correspondence with the factor level ‘Other Americas’. For example, South Africa can be compared better on its economic profile to the likes of Argentina, Chile or Brazil than the USA or the United Kingdom (UK). Subsequently we follow strict statistical theory in evaluating our regression result where real world interpretation is lacking; we therefore use the parameter estimation for factor variable ‘Region’ as is.

The regression process

As with the ILD regression, we firstly inspect regression results for the individual model components; i.e. the size indicator, BLs, ETs and Regions respectively - each time retaining the size indicator. We discuss the ELD regression results whilst we also compare the aforementioned to the results obtained for ILD where appropriate. This overview considers the results provided in Tables A.1 through A.3.

For most of the different model setups defined, regression results between those of the ELD and ILD regressions are comparable. Firstly, with the size indicator, the adjusted R^2 values are comparable - 2.1% for ELD vs. 2.5% for ILD. ELD regression results for both BL and ET factor levels obtained an adjusted R^2 of 8.6%; whereas for ILD the BL and ET regression respectively obtained 9.0% and 10.7%.

Focusing our attention on the regression of the variable Region which was not included in the ILD regression; we see a very low adjusted R^2 value of 2.8%. However, viewed in the context of the low performance of the size indicator we further investigate the overall coefficient determination of a full model when using all factor levels of all variables.

Table 6.6: Reduced OLS regression for ELD

	<i>Dependent variable:</i>
	lCurLoss
log(CurRev)	0.068*** (0.012)
BL.RB	-0.567*** (0.074)
BL.RBR	-1.224*** (0.125)
ET.DPA	-0.921** (0.410)
ET.EPWS	-0.934*** (0.254)
ET.EDPM	-0.538*** (0.165)
ET.EF	-1.037*** (0.089)
ET.IF	-0.796*** (0.091)
Reg.OAmericas	0.985** (0.394)
Constant	15.138*** (0.286)
Observations	1,924
R ²	0.148
Adjusted R ²	0.144
Residual Std. Error	1.514 (df = 1914)
F Statistic	36.890*** (df = 9; 1914)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Adjusting for regression results

We continue our iterative process of the regression model for ELD. We see an adjusted R^2 14.4% in Table 6.6; the ILD regression obtained 14.0%. The ELD regression when reduced results in a much ‘smaller’ model, than that obtained for ILD. This is especially the case when considering the factor levels of the BLs; only two were included, namely those of BL.RB and BL.RBR. These specific terms may point to a concentration of operational losses within the retail environments of the ELD.

In terms of the ET factor levels presented in Table 6.6, there is the same strong representation of factor levels when comparing to the results of the ILD; Table 6.2. All factor levels also show high statistical significance when using a cut-off of a 95% confidence level for inclusion in the model.

We finally turn to the regression results of the variable Region. The only factor level included here is that of ‘Other Americas’. We already alluded to this result in our discussion on including the Region variable - see Section 6.2.5.

Quantile regression

As with the quantile regression for ILD the quantile regression estimates show the confidence level at which the coefficients pass for statistical significance. We retain the formal cut-off of a 95% confidence level. Furthermore, we also keep our high quantile approach by focusing on the tail of the loss distribution. Recall that our quantile regression is on every fifth percentile up to 95%.

Firstly, note that our ELD is now *further* extended on the extended DD-model of Equation 6.7 to accommodate the ‘Region’ term in our model

$$\ln(\hat{Y}_{\theta i}) = \alpha_{\theta i} \cdot \ln(Size_{\theta i}) + \sum_j \beta_{\theta ij} BL_{\theta ij} + \sum_k \delta_{\theta ik} ET_{\theta ik} + \sum_l \xi_{\theta il} REG_{\theta il}, \quad (6.8)$$

where $\theta \in \{0.05, \dots, 0.95\} \cup \{0.96, \dots, 0.99\}$; and $i \in \{1, \dots, n\}$ represents our dataset of losses. Recall that j and β sums over the factor levels of the Business Lines variables and it follows analogously for k and δ for the Event Type factor levels; and now we also see the same for l and ξ which sum over the factor levels of Region.

Since the regression process is analogous to that of the ILD, we do not provide the iterative quantile regression results here, nor discuss them. We refer the reader to Tables A.1, A.2, A.3, A.4, and A.5 for the stepwise decision-making. We consider the final model shortly.

Graphical diagnostics

We are already familiar with the setup of the graphical diagnostics for quantile regression and we therefore turn to the interpretation of the graphs obtained - see Figure 6.2.

We again see varied behaviours for the different variables. On average, the graphs indicated that the parameter estimates are more stable across quantiles when compared to those of the ILD - Figure 6.1. However, we note the significant standard error for variable ET.DPA. Despite the parameter estimates not appearing to be

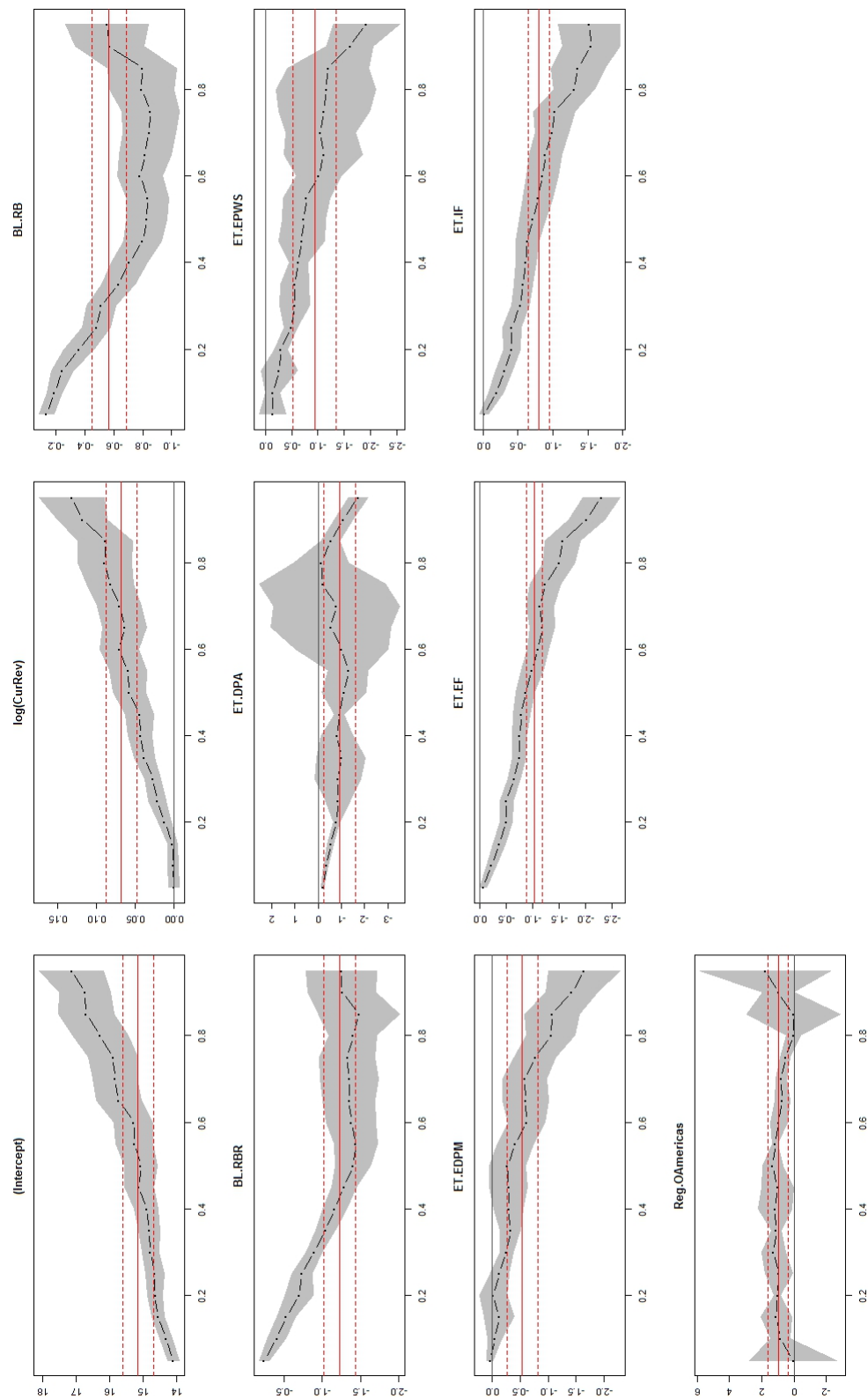


Figure 6.2: Quantile regression on the ELD model's terms

statistically different from the OLS estimates, the standard error around the estimates allow for significant deviation. We infer this from the small (14 observations) and widely distributed (mean = 20,660,000; standard deviation = 30,430,000) ET.DPA subset.

In terms of our variables size indicator - $\log(\text{CurRev})$ - and ET.EPWS, ET.EDPM, ET.EF, and ET.IF we again see that quantile regression appears to be highly statistically significant in the higher quantiles. An interesting observation is that of BL.RB. This variable is considered of medium importance in the context of loss frequency and severity for ELD - see Figures 5.6 and 5.6. Intuitively one would not expect large operational losses for a retail banking division, and we can see this delineated with the statistically significant quantile regression parameter estimates across the mid-level quantiles.

In general we consider the graphical diagnostics to be aligned to our statistical testing. However, with the greater stability seen for ELD quantile regression, the statistical testing is necessary for more detailed scrutiny on the parameter estimates at specific quantile levels. As mentioned, we observe more discrimination towards quantile regression when performing statistical testing; see Tables A.4, and A.5.

The final model

Our final model to be used for determining component A as defined in Equation 6.2, is provided in Table 6.7. As stated for our ILD model, we are using the extended DD-model we developed whilst slotting in the parameter estimates on quantiles where statistical significance was not present.

6.3 The scaling mechanism

6.3.1 Overview

Firstly recall that our proposed solution for scaling is focused mainly on the severity of the losses; i.e. we do not aim scale loss frequency based on external experience - we only wish to inform our ILD with ELD loss *magnitudes*. We therefore include the full scaled ELD dataset with the ILD dataset to form a pooled dataset; i.e. we do not consider a count distribution to reflect frequency from the ELD within the pooled dataset.

In this section we revisit the idea of Equation 6.2 whilst viewing it in the practical perspective of our study. We specifically focus on the conceptualization of components A and B , and do a graphical interpretation of the results originating from our preceding hybrid quantile regression models.

We then delve into so-called ‘*grouping*’ of our components. This is a special solution we developed for our model due to dramatic difference in number of observations between the respective ILD and ELD. We conclude this section by reviewing the final scaling mechanism by means of illustrating its theoretical form.

Table 6.7: Final model for ELD

$\theta =$	<i>Dependent variable:</i>									
	ICurLoss									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
log(CurRev)	0.002	0.013	0.028	0.044	0.058	0.071	0.071	0.090	0.119	
BL.RB	-0.187	-0.358	-0.511	-0.707	-0.825	-0.781	-0.843	-0.791	-0.571	
BL.RBR	-0.399	-0.694	-0.887	-1.160	-1.399	-1.381	-1.354	-1.401	-1.257	
ET.DPA	-0.346	-0.788	-0.840	-0.810	-1.101	-1.017	-0.791	-0.136	-1.089	
ET.EPWS	-0.139	-0.296	-0.555	-0.630	-0.725	-1.015	-1.050	-1.152	-1.603	
ET.EDPM	-0.044	-0.020	-0.256	-0.291	-0.273	-0.617	-0.581	-1.045	-1.421	
ET.EF	-0.221	-0.491	-0.648	-0.750	-0.859	-1.101	-1.135	-1.502	-2.022	
ET.IF	-0.192	-0.412	-0.532	-0.612	-0.714	-0.854	-0.986	-1.301	-1.545	
Reg.OAmericas	0.866	1.039	1.269	1.170	1.324	0.941	0.809	0.049	0.972	
Constant	14.312	14.633	14.768	14.893	15.061	15.270	15.835	16.282	16.736	

6.3.2 Components of the scaling mechanism

Conceptualization

Our focus on the components of the scaling mechanism, other than the actual operational losses being scaled and scaled to respectively, is firstly on the conceptual definition of said components. We subsequently review a graphical result of the components and discuss their visible properties. We offer solutions to basic problems concluded from this interpretation.

In order to gauge the context of the underlying components, we recall our scaling ratio, Equation 6.2. Once we apply our final models on our ILD and ELD, our results are the values $\hat{Y}_{B_i}$ and $\hat{Y}_{A_i}$ respectively; and i being the number of observations in the individual datasets. These are our predicted operational losses when using our *hybrid* quantile regression models.

Graphical review

For our graphical review of components A and B , we ordered our observations to be monotonously increasing in the respective datasets and then compared the true observed values against those we obtained from our scaling mechanism; i.e. a classic model evaluation of actuals vs. predicted.

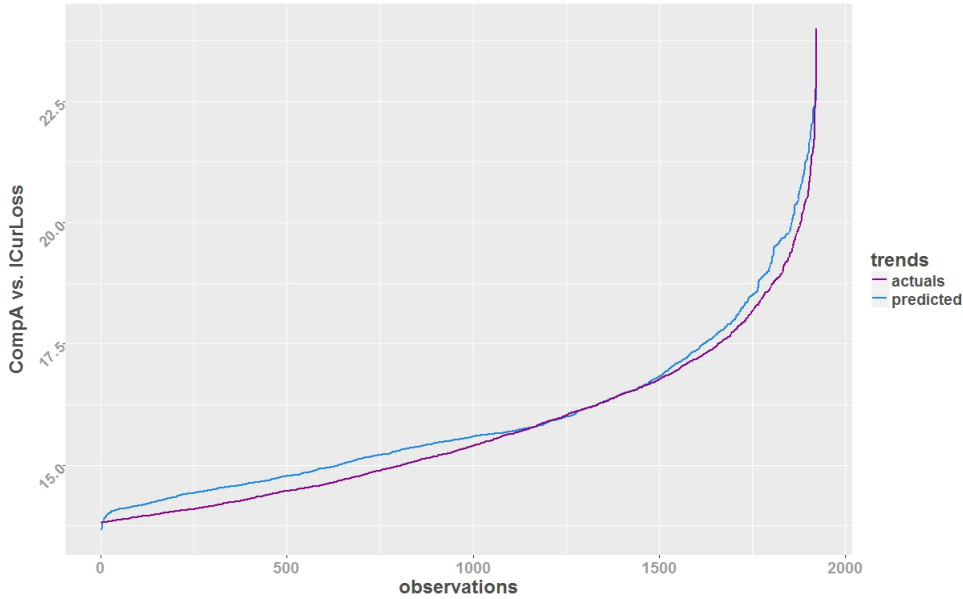


Figure 6.3: ELD - actuals vs. predicted

When considering Figure 6.3, we note positive results. To explain, firstly consider that our objective was to perform a regression with predicted values which closely follow the trend of the actual observed values. We can clearly see this is the case for our ELD regression. Secondly, and perhaps more relevant to the highly regulated financial world; we note the predicted values are mostly slightly higher than the actuals. Despite this not being an explicit aim of our study, it may certainly be included as motivation to a national regulator on the robustness of this procedure;

i.e. the procedure follows a conservative approach as illustrated by the slight overestimation across the trendline.

We note that at the very high end of the set of observations, the actuals are higher than the predicted values. Taking into account that the y -axis is on a logarithmic scale and that these values imply very significant values in pure monetary terms, this is a point which further consideration. We discuss a possible solution later in this section.

On the trendlines for component B in Figure 6.4, we see comparable results for most of the central and high parts of the actuals vs. predicted. However, in the lower of the trendline, we again notice actuals which are higher than predicted values. The difference here is larger than under component A , however the much lower scale may deflate the observable effect. Nevertheless, we consider a basic solution for these instances here.

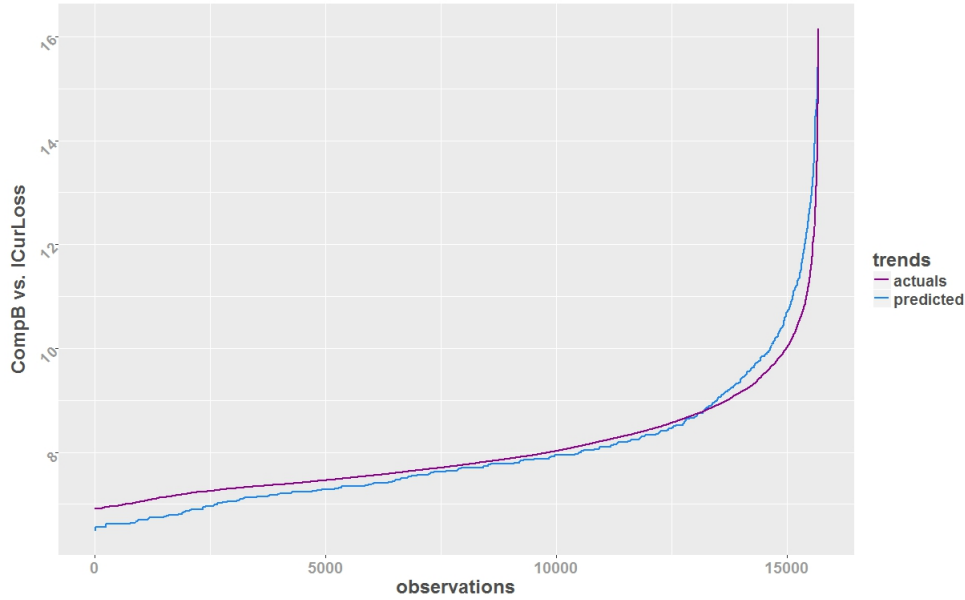


Figure 6.4: ILD - actuals vs. predicted

Suggested solution

Following an argument of conservatism only, we apply a discriminatory filter on the actuals vs. predicted; we thus compare the actuals vs. predicted for each observation unit and include the highest in our final dataset. Though this suggestion may seem crude, we subsequently see that this procedure adds greater stability to our scaling mechanism once we encounter a further problem on the grouping of the components for the final scaling.

6.3.3 Grouping

Theoretical reasoning

We refer to this step of the scaling mechanism as ‘*grouping*’. This procedure is considered a unique extension in our model since the original DD-model followed

a type of observation matching, Dahlen and Dionne (2008:23). Their approach was to match a *single* observed loss in the respective ILD and ELD from the same year with the loss attributes; the loss attributes being a loss's description descriptions in terms of its factor variables. We review our approach here.

Firstly, in our model we do not match by 'same year' at all. Recall from Section 5.4.1 that we applied cumulative inflation. We therefore transformed all data to be observed at a single point in time. We therefore do not consider it appropriate in our process to reference the year of an operational loss at this stage, since the transformed loss value we are working with as an actual, already contains that information through its inflation adjustment.

Secondly, we do retain the idea of matching according to loss attributes. We feel that this is an authentic approach when trying to link operational losses in greatly varying datasets. However, we did not undertake the further approach of matching single losses since we have a fairly large ILD set and want to fully utilize it. Given our earlier work in quantiles and its core place in this scaling mechanism, our approach was therefore to 'group' the loss ratio information by quantiles. Thus, for a single quantile as defined under the domain of Equation 6.8 and its analogous ILD form, we isolate all the respective ILD and ELD losses along with their respective component A and B . As mentioned, our ILD is large in comparison to the ELD (15,684 vs. 1,924 observations), and to fully apply this dataset, we opt for averaging across all the of individual components B for every defined quantile.

Graphical review

We see the boxplots of component B per defined quantile in Figure 6.5. Recall that the so-called hinges of the box itself correspond to the 25th and 75th percentiles respectively. Also, the concomitant upper and lower notches, or whiskers, extending away from the hinges correspond to a value of $1.5 \times \text{interquartilerange}$. In Figure 6.5 the green dots indicate outliers; defined as observations outside the interquartile range.

In our graph we note that in terms of the actual number of observations, most of the population lies within the quantile range $[0.05, 0.90]$. For the higher quantiles the location and scale of component B increases sharply. We take into account that the wide dispersion seen for these higher quantiles may also be somewhat distorted due to the smaller number of observations in these subpopulations. The same applies to the fact that there are no outliers classified within these higher quantiles.

6.3.4 The new dataset - a scaled and pooled view

The final step in solving the original Equation 6.2, is to substitute term $Loss_A$ with the actual losses from our ELD². We effectively then obtain our final prediction for a scaled value, $Loss_B$:

$$Loss_{B_{\theta i}} = Loss_{A_{\theta i}} \frac{\hat{Y}_{B_{\theta i}}}{\hat{Y}_{A_{\theta i}}} \quad (6.9)$$

²Recall that subscript A indicates the ELD and subscript B the ILD.

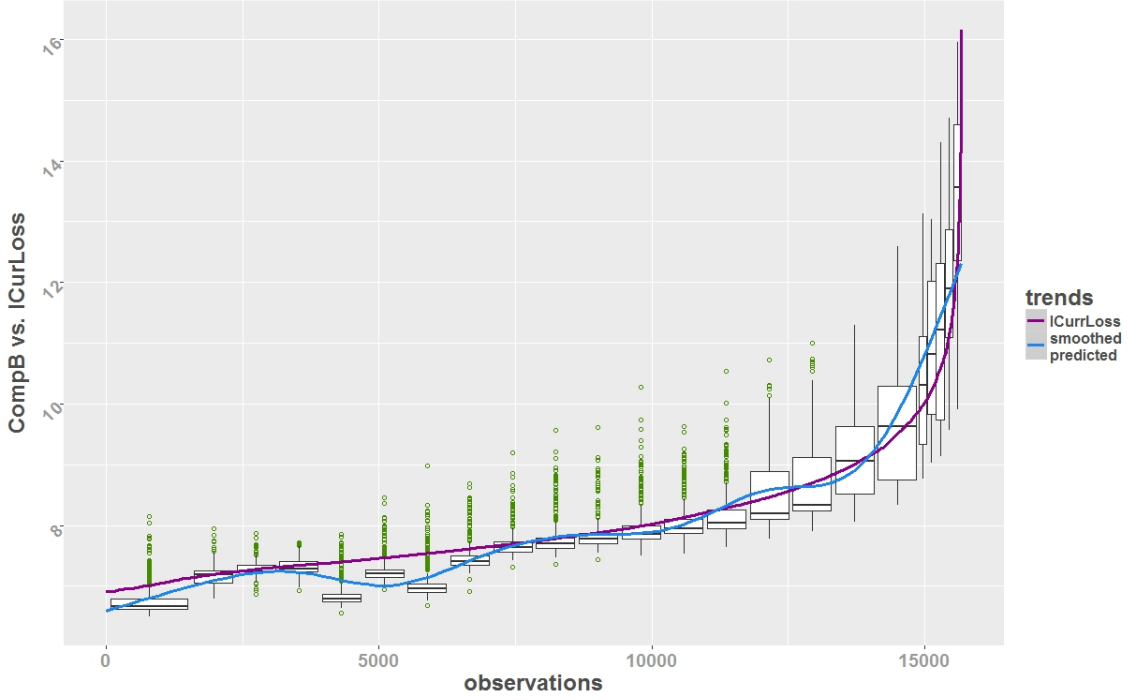


Figure 6.5: Boxplots per quantile of component B's ranges

where this resulting value is then intended to be used directly in the ILD set. Or further, using our respective quantile regression models, the theoretical expression becomes

$$Loss_{B_{\theta_i}} = Loss_{A_{\theta_i}} \times \frac{\hat{\alpha}_{\theta_i} \cdot \ln(Size_{\theta_i}) + \sum_j \hat{\beta}_{\theta_{ij}} BL_{\theta_{ij}} + \sum_k \hat{\delta}_{\theta_{ik}} ET_{\theta_{ik}}}{\hat{\alpha}_{\theta_i} \cdot \ln(Size_{\theta_i}) + \sum_j \hat{\beta}_{\theta_{ij}} BL_{\theta_{ij}} + \sum_k \hat{\delta}_{\theta_{ik}} ET_{\theta_{ik}} + \sum_l \hat{\xi}_{\theta_{il}} REG_{\theta_{il}}} \quad (6.10)$$

6.4 Impact analysis

6.4.1 Overview

In this section on the impact analysis, we inspect the differences between the original ILD and our updated pooled dataset; i.e. the original ILD combined directly with the scaled ELD data. Our inspection of the differences consists of both an empirical review along with a graphical review.

Our last step in inspecting the differences consists of calculating rudimentary undiversified capital charges for the respective datasets. In this way we can to what extent the capital charge increases when including scaled ELD observations.

Finally we conclude on the effectiveness of our approach. We base this opinion on our conclusions drawn from the dataset analysis and the capital charge evaluation. We also comment on how appropriate the overall effects of the scaling and direct combination are in solving our external data scaling problem.

6.4.2 Data analytics on the pooled set

Overview

In this section we use graphical and empirical analysis on our scaled data. We analyze the respective ILD, scaled and pooled data, and where applicable the ELD. By performing these analyses we can already confirm in a basic sense whether we have achieved our objective; i.e. scaling an external dataset to a lower and more comparable level so that we can use the scaled data directly in our ILD as a pooled dataset.

We firstly consider a graphical analysis where we consider the densities of the various datasets to form an overall perspective on how the external data has ‘*moved*’ to its scaled form. We also review how this scaled set influences the ILD when directly combined to form the pooled dataset.

Secondly we perform an empirical analysis to directly compare the densities’ descriptive statistics. We thus calculate and compare quantiles and other applicable statistical components to gauge whether our scaling results are appropriate and thus permissible to be used directly for a pooled dataset.

Graphical review

In this section we consider our scaled data results by means of a graphical analysis. In Figure 6.6 we see the densities of ELD, the scaled ELD data using our scaling mechanism, and ILD.

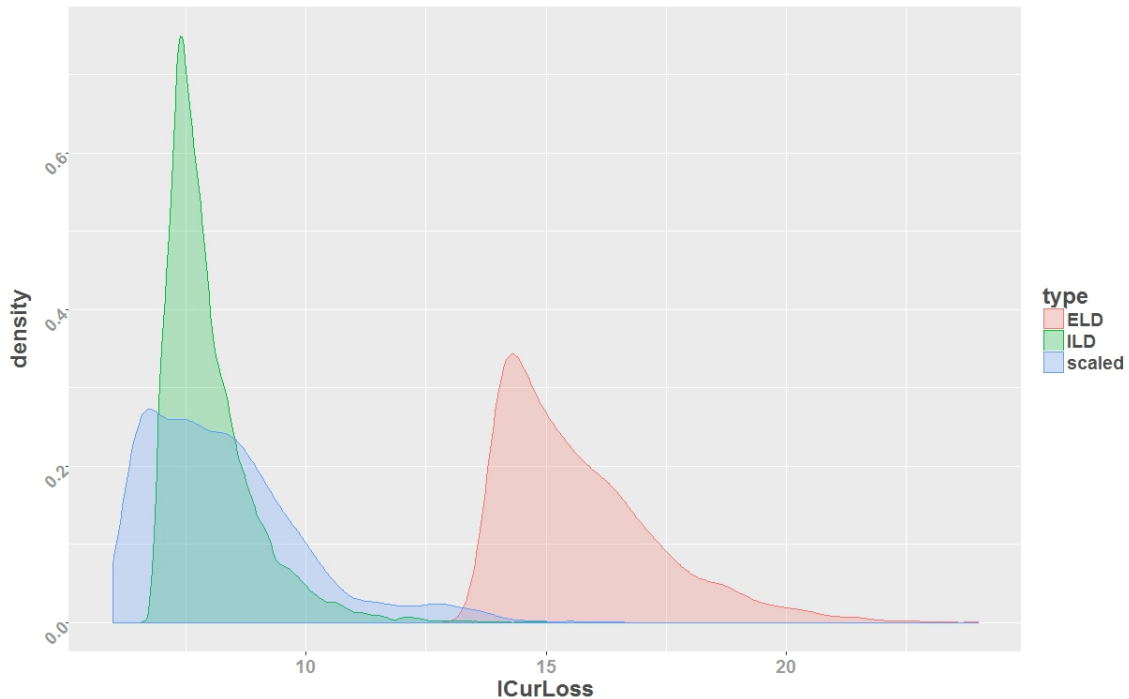


Figure 6.6: Densities of the scaled data vs. those of the ILD and ELD

When comparing the scaled and ELD densities, we can see that functions appear to be comparable in form. We see a mild peakedness and the clear positive skewness

with extended tails in both graphs. For the scaled data we specifically note that the aforementioned extended tail is also somewhat more leptokurtic than that of the ELD density. We consider this a desirable attribute since such data lead to more conservative - i.e. greater - capital estimates for the high loss events - see Figure 6.7. This is the very reason for the inclusion of ELD when the ILD is of a smaller scale - incorporating comparable but more and more severe operational densities allow for greater informed capital estimation.

When comparing the scaled and ILD, we can see alignment in terms of location and scale. The long tail of the scaled is clearly an extension from the observation in the ILD. As mentioned the leptokurtic tail of the scaled data is a considerable improvement on the very platykurtic tail we observed for the ILD. As mentioned, our aim was to extend this part of the function specifically to better our operational risk capital calculation.

We construct Figure 6.7 where we isolate our graph for the tails of the pooled and the original ILD. We clearly see that the pooled tail is significantly longer than that of the ILD - given that we are considering a logarithmic scale on the x -axis.

From our graphical analysis we conclude that the scaling of the ELD achieved an appropriate and comparable form when viewed against our core ILD. We conclude that the tail of a pooled dataset is longer and slightly heavier than the original ILD. Given our conclusion drawn from the graphical analysis on the scaled data, we now consider empirical analytics on the pooled dataset.

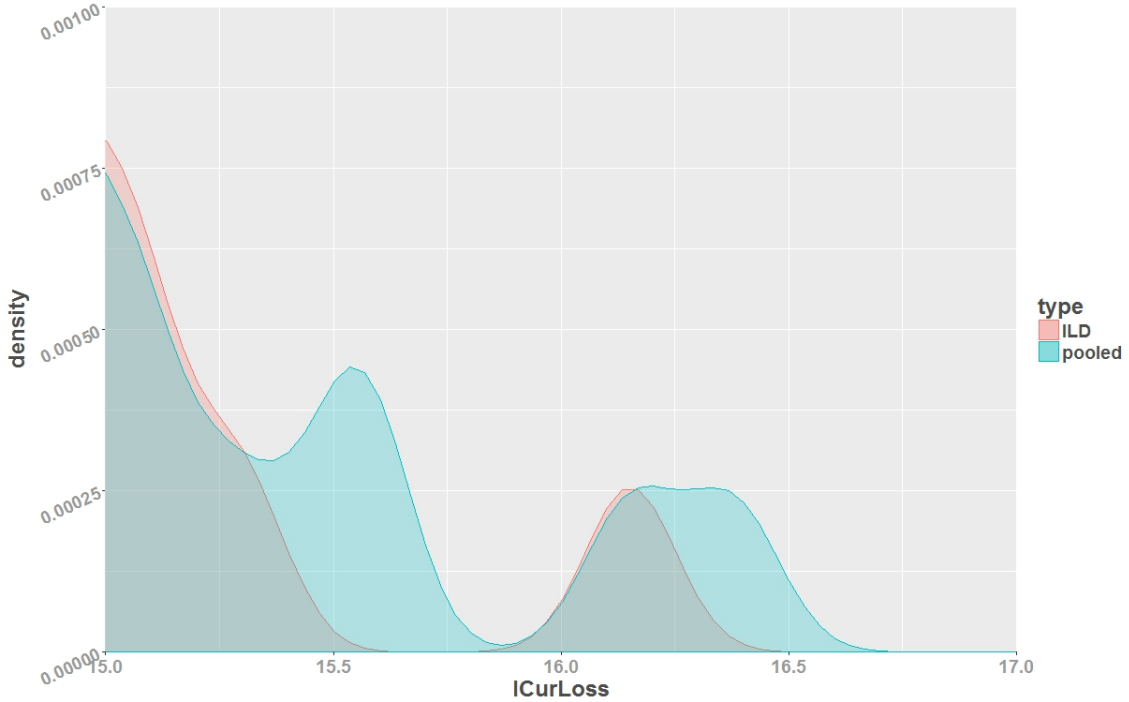


Figure 6.7: Tail densities of the ILD vs. full pooled dataset

Empirical results

We compare the statistical information of the pooled dataset with the original datasets. For this empirical overview we consider the quantiles of the different

datasets. We also provide a discussion on the statistical measures of skewness and kurtosis to gauge the changes in these distributional attributes which are especially informative for asymmetric distributions and their tails.

When considering the quantiles presented in Table 6.8, we firstly clearly note the pronounced difference between the ELD and all the other datasets. This table shows the scale and magnitude of the ELD which would make it impossible to do direct inclusion with the ILD. For the remainder of the quantiles, we again deem our results of firstly the scaled, and finally the pooled dataset as successful. We note that for most of the quantiles listed, the pooled datasets results are almost a type of *midpoint* between the quantile results for the ILD and scaled datasets respectively.

Table 6.8: Quantiles across the different datasets

data	25%	50%	75%	90%	95%	99%	100%
pooled	7.365	7.747	8.420	9.314	10.008	12.065	16.377
ILD	7.370	7.738	8.372	9.210	9.881	11.553	16.153
scaled	7.012	8.030	9.094	10.344	11.823	13.561	16.377
ELD	14.445	15.315	16.577	17.982	18.936	20.887	23.998

In Table 6.9 we look at the statistical measures which often provide a more definitive view on asymmetric distributions such as our loss distributions. In terms of skewness, which defines the symmetry, or in our case asymmetry, of a distribution. We firstly note that the scaled datasets was highly effective in retaining the asymmetry as contained within the ELD. Furthermore, once we combined the datasets to form the pooled dataset we then see that skewness shows greater alignment to that of the ILD.

With respect to (excess) kurtosis which provides an indication of the ‘tailedness’, we see a comparable result. The scaled datasets firstly retain the tail characteristic of the ELD, however we note that it is less platykurtic than the ELD tail. Mirrored results are seen for the ILD vs. pooled dataset; the pooled dataset is less leptokurtic than the ILD tail.

Table 6.9: Function attributes across different datasets

data	Skewness	Kurtosis
pooled	2.060	6.640
ILD	2.099	6.850
scaled	1.279	1.986
ELD	1.253	1.638

In conclusion, we deem the scaling and pooling of the data as successful judging from our empirical and graphical assessments. We saw that ELD attributes are well preserved after scaling, but also that the pooled dataset is very much representative of the original ILD.

Recall that this empirical analysis is based on the dataset as a whole. However, when modelling for capital estimates we focus on the individual ORCs. We therefore get an even better perspective on the comparability between the different datasets and the effective scaling of external data once we start calculating undiversified capital estimates per ORC.

6.4.3 Undiversified capital charges

Process description

We follow a rudimentary approach in calculating capital charges for our operational risk losses. We determine which of the ORCs have sufficient observations to fit loss distributions with the aim of performing Monte Carlo simulations to calculate VaR capital charges.

We align our classification of ‘sufficient’ observations to that of our host institution; i.e. a minimum of 25 observations. Out of a total of 56 possible ORCs (i.e. 8 business lines $\times$ 7 event types), we note that on average both our ILD and pooled datasets only have observations for 49 ORCs. Using 25 observations as a cut-off, we are able to calculate capital estimates for 22 ORCs - this amounts to approximately 45% of the population in terms number of observations. An interesting we noted when comparing the two datasets of terms of the observations’ magnitude is that the results are analogous to those of the ORCs’ frequencies; we therefore select the same 22 ORCs for capital estimation and deem them as representative of the ORCs’ frequency and severity.

For the fitting process we assume Weibull distribution throughout (Chernobai et al., 2007:117). The probability density function of the (two-parameter³) Weibull distribution is given by:

$$f(x; \eta, \beta) = \begin{cases} \frac{\beta}{\eta} \left(\frac{x}{\eta}\right)^{\beta-1} e^{-(x/\eta)^\beta} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \quad (6.11)$$

where β is the shape parameter and η signifies the scale parameter. In combination with the Lognormal distribution, the Weibull distribution is often the most popular for loss (severity) distribution fitting. Subsequent to recording our parameter estimations across the 22 ORCs we perform a Monte Carlo simulation for the respective datasets to obtain a VaR estimate - following our discussion in Section 2.4.2 on obtaining aggregated losses.

Lastly, for our loss frequencies, we assume a Poisson distribution with parameter λ for our convolution simulation exercise (Chernobai et al., 2007:88). Its probability mass function is given by:

$$f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} . \quad (6.12)$$

We approximate the λ parameter by using the frequencies obtained from our overall ILD as accumulated across the eight years. We perform 5 million iterations of Monte Carlo simulations for each of the selected ORCs.

³The most generalized form is the three parameter version which includes a third parameter γ , to allow for location in addition to shape and scale.

Table 6.10: Estimation of distribution parameters

ORC	<i>ILD</i>			<i>pooled</i>		
	$\hat{\lambda}_{ILD}$	$\hat{\beta}_{ILD}$	$\hat{\eta}_{ILD}$	$\hat{\lambda}_{pooled}$	$\hat{\beta}_{pooled}$	$\hat{\eta}_{pooled}$
3	7.38	7.55	8.73	7.50	7.60	8.73
6	28.00	5.41	8.50	29.13	5.46	8.50
11	3.38	4.64	9.81	6.75	4.38	10.11
14	20.63	5.96	9.06	45.63	5.39	9.47
15	11.38	8.30	9.24	26.00	6.60	8.79
16	3.75	3.60	10.30	12.38	3.67	10.15
22	6.38	9.23	8.33	6.88	8.82	8.27
30	14.88	7.38	9.04	15.88	7.49	9.03
32	5.63	7.53	10.14	6.00	7.40	10.06
33	5.63	5.90	8.91	5.88	5.92	8.90
34	8.00	7.29	8.88	8.88	6.42	9.06
35	49.38	6.14	9.37	51.00	6.19	9.37
36	3.75	7.17	10.35	4.00	7.21	10.39
37	23.00	4.66	9.45	24.25	4.72	9.46
38	181.50	6.61	8.68	184.38	6.62	8.67
39	10.13	8.77	8.43	12.38	8.10	8.41
40	25.00	4.82	9.59	26.63	4.80	9.56
43	78.50	6.91	8.98	107.38	6.96	8.89
45	18.25	7.55	8.11	19.00	7.60	8.11
46	1,377.63	8.23	8.17	1,413.88	8.22	8.16
48	9.50	8.63	7.92	9.88	8.44	7.98
54	33.63	6.66	9.57	63.88	5.93	9.31

Parameter estimation

We review the results of our distribution fitting in Table 6.10. Considering the ORC matrix of 8 BL rows and 7 ET columns we number the matrix cells *rows* $\times$ *columns* and report only the applicable ORCs.

Firstly we consider the $\hat{\lambda}$ -estimate of the two respective datasets; i.e. our indication of frequency of losses. We naturally expect the pooled dataset to have higher frequencies since more data have been added to the original ILD. Secondly we review the shape parameter estimates, $\hat{\beta}$, of the two datasets. Based on statistical theory we know that when a shape parameter estimate greater than 1 increases, the tail of the Weibull distribution becomes lighter. Recalling our implicit goal of obtaining a loss distribution with more severe losses - i.e. reflecting ELD experience - we see a desirable trend in the parameter estimates of the pooled datasets. On closer inspection we see that most of the parameter estimates (all greater than 1) have decreased; and therefore result in heavier tails which we want for our final loss dataset. The parameter estimates which did actually increase were all increases of 2% or less.

Finally, in terms of the scale parameter estimates, $\hat{\eta}$, we calculated average change as with the shape parameter estimates. Here we note a fairly stable trend; i.e. the scale of the distribution remains fairly stationary. We recall the high number of observations of the ILD vs. the much lower number of transformed (scaled) observations from the ELD. We identify these different concentrations as the reason for the distribution's scale stability.

Capital results

We see our simulated capital results in Table 6.11. For the selected ORCs we provide the respective ILD and pooled datasets' results on Expected Loss (EL) and Value-at-Risk (VaR). Our focus is on the VaR estimate which is our indicator of a capital charge, but we include the EL estimate - the mean of simulated loss distribution - to have an indication of the location of the loss distribution.

Table 6.11: Capital estimates for selected ORCs (in USD '000)

ORC	<i>ILD</i>		<i>pooled</i>		<i>% change</i>	
	$EL_{\bar{X}}$	$VaR_{99.9\%}$	$EL_{\bar{X}}$	$VaR_{99.9\%}$	$EL_{\bar{X}}$	$VaR_{99.9\%}$
3	60.45	141.34	61.53	143.11	1.78%	1.25%
6	219.50	362.15	228.56	374.01	4.13%	3.27%
11	30.28	95.51	62.18	152.11	105.37%	59.25%
14	173.32	306.09	398.45	598.12	129.89%	95.40%
15	99.16	203.15	213.11	356.34	114.93%	75.41%
16	34.80	106.88	113.32	231.84	225.69%	116.91%
22	50.36	123.22	53.78	128.31	6.78%	4.12%
30	126.08	240.54	134.60	252.40	6.76%	4.93%
32	53.54	137.28	56.63	141.88	5.77%	3.35%
33	46.43	119.69	48.46	123.06	4.38%	2.81%
34	66.58	151.69	74.89	165.71	12.48%	9.24%
35	429.89	635.42	444.17	652.93	3.32%	2.76%
36	36.33	108.16	38.95	113.04	7.19%	4.52%
37	198.82	343.93	209.82	358.29	5.54%	4.17%
38	1,468.64	1,823.49	1,491.27	1,848.86	1.54%	1.39%
39	80.72	170.88	98.03	196.23	21.44%	14.83%
40	219.68	372.48	233.20	389.67	6.15%	4.62%
43	658.82	904.59	892.53	1,175.26	35.47%	29.92%
45	138.95	251.57	144.76	259.66	4.18%	3.21%
46	10,613.29	11,518.08	10,879.72	11,795.82	2.51%	2.41%
48	71.14	153.41	74.39	158.58	4.56%	3.37%
54	300.34	476.07	551.15	781.30	83.51%	64.12%
	15,177.11	18,745.62	16,503.49	20,396.51	8.74%	8.81%

For every ORC listed, we see that the pooled dataset has both a greater EL and VaR estimate. In certain instances the increases are noteworthy - up to 225%.

When analysing for the large increases we are not concerned for so-called ‘*over-representation*’ of the cells in question: these ORCs overall contribute fairly little to the total capital charge. Each of the top three proportional changes contribute less than 2% in either the ILD or pooled datasets’ capital charges.

The ORCs which are the top three contributors to the total capital charges (ORCs 38, 43, 46) show mostly unremarkable proportional changes - less than 3%, save for ORC 43 where there is an increase of 36% (ILD), 30%(pooled). We therefore revert back to our ORC matrix encoding and see that this cell represents External Fraud losses within the Commercial Banking business line. Given the context of both the ET and BL type we do not consider the 36% as unreasonable.

Overall we consider the result very positive for our proposed solution to external scaling. The ultimate result of operational risk capital modelling is summarized in the single value of the final capital charge. For our impact study this value amounts to \$20,397,000 - or using approximated end-2016 exchange rates R285,551,000; i.e. our 99.9% VaR (capital charge) for the pooled dataset. This represents an increase in the capital charge from the ILD charge of 8.81%. This is certainly not an unreasonable increase given that we can now confirm through our study that our new pooled dataset reflects the external data experience.

6.5 Chapter conclusion

In this section we conclude on our proposed solution to the problem of external data scaling, the process followed, and the success of implementing the process for real-world data.

Firstly, we highlighted our theoretical starting point for the scaling mechanism as provided in Equation 6.2. We identified the key elements of the scaling mechanism which we need to calculate for our data; namely those of the regression on ELD, and the regression on ILD. We then see that we must scale the original ELD operational loss values by the ratio of the mentioned ELD-ILD regression to obtain a scaled value to be used for pooling into the original ILD. As part of this overview we briefly revisit the theoretical traction made in the study by Dahlen and Dionne (2010) and define our encoding of the binary variables for the factor levels associated with the factor variables BL, ET, and REG.

Our practical solution then starts off by performing various sets of OLS regressions on the ILD to gauge the various effects of our different models’ terms; i.e. size indicator vs. BL vs. ET. This provided us with an indication of which variables (e.g. factor levels where appropriate) ought to be included in our regression model for component B. Accounting for base cases of binary variables, our final model only excluded BL.AM and ET.DPA; where the included terms provided highly statistically significant p -values. Using the guidance from our OLS regression, we extended the DD-model by performing quantile regression for our model, given that operational risk is inherently focused on tail behaviour; i.e. the high quantiles of regression.

The quantile regression provided a type of ‘zoomed-in’ view on our regression; we were able to where OLS parameter estimates were significantly deviated from the obtained quantile regression estimates in the distributions’ tails. We also cross-referenced the results with graphical diagnostics which illustrated the need for

quantile regression in most of the variables. We specialized our quantile regression by performing even more detailed regression by regressing for single percentiles in the highest end of the distribution's tail; i.e. $\theta \in \{0.95, \dots, 0.99\}$. We subsequently finalized the model by using our hybrid approach of statistically significant quantile and OLS regression parameter estimates.

The regression, both OLS and quantile regression, followed analogously for the ELD. We only mention two points of variation here. Firstly, from the OLS *guidance* regression we noted that much fewer model terms were included in the final model when compared to ILD results. Secondly, we extended our improvement on the DD-model even further here, since we were able to include the REG variable to account for geographical information on where the operational losses took place. Again, our graphical diagnostics confirm clear alignment with the empirical testing.

Our next step was to perform the actual scaling using our regression results. We did not follow the direct 'loss-to-loss per annum' scaling as done in Dahlen and Dionne (2010) for two reasons; we had a much larger ILD which we wanted to utilize fully; and secondly, we had already accounted for annual inflation and therefore had cross-sectional data in effect as opposed to historical data.

We started off our scaling mechanism by reflecting on the efficacy of our regression results which forms the scaling ratio. For both ILD and ELD we noted that the predicted values reflected the actual values very well overall. Where there were issues - e.g. significant discrepancy in tail behaviour - we provided an overlay to enforce conservatism for the loss distribution. We then solved our issue of grouping due to the difference in dataset sizes by constricting averaging groups as proxies for our scaling intervals. Finally, we comment on the scaling mechanism and provide the formulaic expression of our extended model in Section 6.3.4.

In order to test the practical implications of our scaling mechanism, we inspect the pooled dataset whilst also comparing it to the respective ILD, ELD, and (isolated) scaled data. We consider both graphical and empirical interpretations of the scaling results and concluded that the scaling mechanism successfully captured the distributional aspects of the original ELD whilst still *transforming* said data to pool properly with our ILD.

We concluded our inspection by doing a mock undiversified capital charge calculation. We performed Monte Carlo simulations on our original ILD vs. pooled data and then calculate the expected loss and Value-at-Risk. We noted an increase in the capital charge of 9% which we consider a positive result, given that we have incorporated the information of highly divergent ELD which exaggerate capital, but still obtained a moderate final increase to appease opportunity costs associated with capital charges.

Chapter 7

Conclusion on external data scaling

7.1 Literature study

Our literature study opened with the review of power law transforms. It was explained that power-law relationships are often seen in physical and financial worlds. This technique also forms the core application of the Shih et al. (2000) study. However, the most important aspect which follows from Shih et al. (2000), is using a size indicator - i.e. a type of substitute or proxy - which captures the size of a financial institute.

We then provided a detailed account of the scaling mechanism as devised by Na et al. (2006), specifically taking note of the concept of ratios of common vs. idiosyncratic elements of operational losses. In this study we saw the operational risk characteristic of business lines introduced as a measure of obtaining an aggregate operational loss.

Dahen and Dionne (2010) built even further on these concepts by including the operational losses' event types, whilst also reworking the loss ratio to a new scaling ratio

$$L_{ILD} = L_{ELD} \frac{\hat{Y}_{ILD}}{\hat{Y}_{ELD}} . \quad (7.1)$$

These two aspects formed an important part of our proposed solution; firstly since we used said loss ratio as developed further by Dahen and Dionne (2010) from the preceding work of Na et al. (2006). Secondly, we continued with the idea of the DD-model, and extended the Dahen and Dionne (2010) scaling model to include the 'Region' variable where we were able to; i.e. in our ELD regression.

In order to propose a solution to Problem I for external data scaling, we combined the techniques learned from the core literature with a second angle on the regression of loss data; namely that of quantile regression.

Quantile regression is a regression concept where the general technique of OLS is extended; i.e. instead of regressing for a conditional mean value of the dependent variable; the regression dataset's explanatory variables are divided into the quantile group sought for regression. The regression is then towards a conditional quantile value of the response variable. Intuitively we already conclude that heteroskedasticity

is therefore accommodated since we are specifically regressing in smaller areas of the explanatory variables' variability. Said variability is thus captured in the regression results for the quantile at hand.

Our study of quantile regression centred on the theory as laid out by Koenker (2005) and the practical examples as provided in Davino et al. (2014). The statistical definition for obtaining a quantile regression is defined by Davino et al. (2014) as

$$\hat{\beta}(\theta) = \underset{\beta}{\operatorname{argmin}} E[\rho_{\theta}(Y - \mathbf{X}\beta)] , \quad (7.2)$$

where $(\theta) \in [0, 1]$ indicates the specific quantile for which we are performing the regression.

As seen in Na et al. (2006), there exists a possibility to also do regression and subsequent scaling on the frequency of the loss data. Our study explicitly only focused on the severity since we had decided firstly to incorporate the *full* ILD, and secondly, as *initial* study we definitely wanted overrepresentation of the ILD. However, we recall the results obtained by Na et al. (2006) - see Section 3.3.3 - that the frequency aspect in fact proved to have the most conclusive regression results in that study. There are various techniques available for doing discrete variable regression should we wish to do frequency regression; even within the field of quantile regression which we applied - see Koenker (2005:255) and Davino et al. (2014:3,15).

With respect to Dahlen and Dionne (2010), we saw an increased focus on in-depth analysis of operational risk loss data; i.e. Dahlen and Dionne (2010) used the natural ordering attributes of operational loss to a greater effect than was done in the Na et al. (2006) study. In the Dahlen and Dionne (2010) extended model (the so-called DD-model), 'BL' and 'ET' are used effectively to obtain regressions for the scaling ratio. Revisiting the definition of data types for operational risk as discussed in Section 2.3, we suggest further studies to consider how the results of scenario analysis and even BEICFs can be utilized in such scaling techniques. Both of these mostly qualitative data aspects can inform the selection of subsets for ILD and ELD prior to initiating a scaling mechanism.

There are extensive refinements we can consider in future studies when using quantile regression. Firstly, the censoring of observations as is clearly done for both ILD and ELD can be incorporated in quantile regression techniques. Koenker (2005) provides a theoretical overview on this topic. Another fitting technique we can consider is to test for quantiles not crossing; there are clear theoretical definitions for such approaches in the case of survival models. This technique is naturally closely linked to that of ensuring monotonicity in quantiles. Practical examples are, however, also aimed at survival models at present.

7.2 The data environment

In our analysis of the data environment we had a large ILD set and suitable ELD set. We analyzed the respective datasets in terms of basic statistical measures, as well as graphical diagnostics on both the dependent and independent variables.

A specific point we wish to mention here is the variable used as proxy for the financial institution's size - the size indicator. Most of our literature studies confirmed

the size indicator as a significant model term to be included during regression. Our own study also confirmed this. However, when the potential proxy variables were investigated, the statistical analyses resulted in problematic answers. For our ILD the proxy relationship seemed to be contradictory (confirmed during regression), whilst the relationship appeared to be very weak for ELD (opposite confirmed during regression).

We can argue from our basic intuition that such relationships are incorrect and the model should be reworked. However, directly the same intuition for institution across two highly different geographical areas are inappropriate. When we consider the pronounced effect that BEICFs may have on operational losses the indirect relationship seen for the size proxy in ILD could indeed be plausible.

With respect to the data environment, we would naturally advise to include as much ELD as possible. For this reason we propose including ORX data where possible. The existing ELD once filtered for the standard requirements already provides a suitable case study, however, widening the international aspect of ELD can only be beneficial for gauging overseas data environments.

7.3 Proposed solution

In our proposed solution we firstly followed the DD-model by performing OLS for our model's terms. We subsequently used these regression results to guide our variable selection for quantile regression. For our quantile regression we performed regression on every 5% quantile of the variable's distribution. We later extended this to zoom in on the very high quantile by regressing for every 1% quantile from 95% through 99%. We noted the tips and illustrations provided in Davino et al. (2014:48-60) on selecting the actual quantiles for which to regress. We followed the implication of using graphical diagnostics to get a rough estimate of the area where quantile regression is most appropriate. However, we feel confident in our selection since it adheres to operational loss focus of high-severity losses, and it links up with the graphical diagnostics as mentioned.

During our scaling mechanism build-up we performed a back-testing exercise where compared the scaling components A and B - see Figures 6.3 and 6.4. It was clear that the initial regression was successful and closely tracked the expected values. However, in the extreme tail sections, the tracking deviates; i.e. the tracking retains the natural regression trend whereas the tail deviates even further. Given that our regression did focus on extreme quantiles, we followed as conservative approach as possible by creating a maximum test. In this test we created a new dataset by testing whether the actual or predicted value was the highest and then substituting the highest into our working dataset. Linking back to the selection of quantiles, we advise future research to revisit this step in the scaling mechanism. We suggest that quantiles can be extended to regress for even '*deeper*' quantiles: e.g. $\theta \in [0.991, 0.992, \dots, 0.999]$. The quantile regression process is naturally able to distinguish between statistically significant parameters when these are present among the distinct quantiles. To illustrate, consider the graphical analysis of the model term ET.EF in Figure 6.1; here we see a very steep gradient in the parameter's graphical interpretation which may point to very high sensitivity in the extreme quantiles.

Despite the shortcomings we discussed in this section, we are still confident that our scaling mechanism provides an efficient, robust and conservative solution to the problem of external data scaling. We firstly consider our proposed solution to be efficient as the methods used can easily be executed in statistical computer languages whilst the modelling steps are also easy to follow for a quantitative analyst. Secondly, we feel that our combined steps to address the scaling mechanism are robust since we managed to find initial solutions to data and modelling complexities experienced during the modelling process. Last, we deem our approach as conservative since we always selected the most ‘restrictive’ option when faced with a model challenge; i.e. we applied our maximum test on the scaling components, for the grouping of scaling ratios we tested that variations away from the average ratio applied did not underestimate the scaling ratio.

Considering our result of a 9% increase in capital when using a pooled dataset, we feel that this scaling mechanism is appropriate for use in South Africa. As evident from the comparison to pure ELD statistics - refer to Tables 6.8 and 6.9 - a direct usage of ELD would be ruinous for a financial institution when calculating capital estimates. We therefore consider the capital increase to be within expectations of capital levels whilst we have still provided a solution to incorporating ELD *directly* - albeit *scaled* - in our loss dataset. We consider the 9% increase as a minimal penalty for the use of ELD during capital estimation. The actual increment value may be disputed in institutions; however, considering the significant capital decreases we see when applying dependence structures to obtain diversified capital, such an increase in the loss dataset may be deemed an appropriate reflection of an ELD-informed dataset.

In the next part of our study, and using our new pooled dataset resulting from our proposed solution of Problem I, we attempt to provide a solution optimizing capital diversification through creating dependence structures to accommodate interdependencies we are expecting in the loss dataset.

Part II

Problem II - Dependence optimization with factor copulas

Chapter 8

Review of academic literature

8.1 Introduction

In this literature study for preparing to propose a solution to Problem II on dependence structure optimization, we provide a detailed theoretical overview of copula theory.

We firstly consult the classic sources for copula theory and retrace the development of copulas (Nelsen, 2006). Although we do not consider the central proofs in our text, we clearly derive the theoretical definition of copulas and also provide graphical illustration of the theoretical concepts.

Our literature study then continues on to natural copula structures and their associated structural classifications. We specifically focus on the well-known elliptical and Archimedean copulas. In this overview we also devote special attention to the tail behaviour as linked to the respective copulas. We deem this an important component as our study is centred on operational losses which inherently become more pronounced, and therefore damaging to a financial institution, higher up in distributions' tails.

This initial overview then concludes with theoretical definitions of dependence measures with which we can use our copulas to provide more information on the dependence structure we are building into our capital models. We define and where appropriate derive the theoretical formulae to illustrate with what ease dependence structure can be extracted from copulas once constructed.

8.2 Background of dependence structures

We see the scene set for using dependence structures in Chernobai et al. (2007:81). The discussion follows the urging of the BCBS to investigate the grouping of operational loss data with respect to their 'different degrees and nature'. Should we directly calculate capital estimates per ORC, the natural summary technique would be to summate said capital estimates. This may result in very high total capital estimates, whilst also not acknowledging that operational losses may see dependent dispersion across the grouping we utilized on our operational loss data. In our study we only consider inter-cell (ORC) dependence; i.e. not intra-cell dependence. This distinction is explained in our discussion on data preparation (see Section 10.2).

As mentioned in Chernobai et al. (2007), linear dependence is the most common dependence measure, but we already know that operational loss data exhibit significantly nonlinear behaviour. When accounting for nonlinear dependence it is explained that literature studies often demonstrate reduction in capital estimates by up to 30-35%¹ (Chapelle et al., 2004).

Our study follows the class of aggregate loss dependence when analyzing our dataset. Chernobai et al. (2007) also note the possibility of analyzing for *frequency* and *severity* dependence respectively, but cites Frachot and Roncalli (2002) when mentioning that the BIS only points to dependence measurement once operational risk modelling had already reached the point of capital estimation, and we can therefore intuitively accept that our focus should be on aggregate loss dependence.

8.3 A core theory of copulas

8.3.1 Introduction

Trivedi and Zimmer (2005) provide a clear introduction to the field of copulas. We firstly see that when encountering dependence measurement problems, we often have more information about marginal distributions than joint distributions. Furthermore, copulas provide a method by which we can develop conceptualization beyond the known pairwise techniques we apply for correlation.

Perhaps the most useful aspect of copulas is the opportunity to perform in-depth analysis on the tails of our dependence; i.e. we can now analyze the tails for the presence and magnitude of concordance. As can be concluded from our overall discussions on operational risks, it is clear that such a method is beneficial for our study of operational risk as we are focusing on the tails of our loss distributions. By using copulas, we have a method by which we can determine how our operational loss data's underlying variables interact, and specifically in their tails where the most crucial loss events are centered.

In this section we consolidate our theoretical knowledge of copulas and their workings to be applied in our study. Moreover, we review Sklar's theorem which forms the core of theoretical definition of copulas. We then also consider a practical illustration of copulas, and subsequently close this section with an overview of special copulas.

8.3.2 Sklar's theorem

The first concrete link developed between multivariate distributions and their underlying marginal distributions is first seen in Sklar (1959) - as cited in Nelsen (2006:16-17). We follow the latter's exposition of Sklar's theorem where H is defined as a joint distribution function with underlying marginal function, F and G . A copula C is then defined for all stochastic values x, y within $\bar{\mathbf{R}}$ such that

$$H(x, y) = C(F(x), G(y)) , \quad (8.1)$$

¹It is mentioned in CHERNOBAI, A. S., RACHEV, S. T. & FABOZZI, F. J. 2007. Operational Risk: A Guide to Basel II Capital Requirements, Models and Analysis, Hoboken, New Jersey, John Wiley & Sons Inc., that this range can actually extend up to 40%.

where $\overline{\mathbf{R}}$ is defined as the extended real line $[-\infty; \infty]^2$.

The ensuing distribution function, copula C , is therefore the joint distribution with underlying marginal distributions F and G .

Using this result we continue directly onto the n -dimensional definition for copulas given in Cherubini et al. (2012) by means of applying the *Fisher (integral) transform*³ to our definition of a copula and its underlying marginals in a two-dimensional setting where

$$F(X, Y) = F(F_X^{-1}(U), F_Y^{-1}(V)) \equiv C(U, V) . \quad (8.2)$$

8.3.3 Illustration and definition of copulas

In order to illustrate the concepts we consider the reasoning provided in Cherubini et al. (2012:12) in combination with the illustration idea provided in Nelsen (2006:20). For a two-dimensional setting we consider the probability that a variable falls within a specific area where we define $x_1 > x_2$ and $y_1 > y_2$ - noting that probability is always greater than 0 - and subsequently illustrate this on the Cartesian plane in Figure 8.1. We then consider the mathematical argument with

$$F(x_1, y_1) - F(x_1, y_2) - F(x_2, y_1) + F(x_2, y_2) \geq 0 . \quad (8.3)$$

This result can be extended into further dimensions. Specifically, for three dimensions ($\overline{\mathbf{R}}^3$), the rectangle in the plane $\overline{\mathbf{R}}^2$ seen here becomes a volume space.

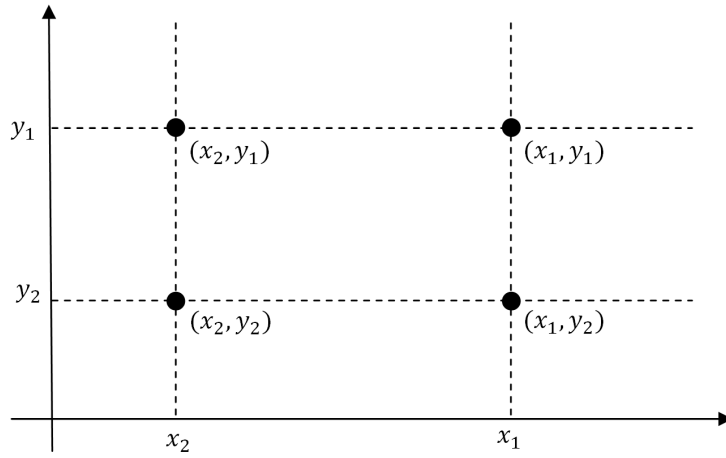


Figure 8.1: Defining probability on a plane

Cherubini et al. (2012) defines the n -dimensional form as a volume:

$$V_{[\mathbf{u}, \mathbf{v}]} = \sum_{V \in R} \text{sign}(V) \cdot C(V) , \quad (8.4)$$

² $\overline{\mathbf{R}}^2$ is therefore the extended real plane, and so forth. The inclusion brackets "[,]" are used since these reflect coordinate points, not actual values.

³This integral transformation technique is considered as upfront knowledge and is not discussed here. We refer the reader to FISHER, R. A. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10:16. We merely note that when considering any stochastic variable X with its cumulative distribution function $F_X(X)$ and then calculate $U \equiv F_X(X)$, the resultant variable is uniformly distributed and we therefore have $\mathbb{P}(U \leq u) = u$, $\forall u \in [0, 1]$. This applies for any uniform variable.

where $[\mathbf{u}, \mathbf{v}] = \prod_{i=1}^n [u_i, v_i]$. We then use this result to define a copula in a general sense; i.e. n -dimensions. The following is the definition summarized in Cherubini et al. (2012:14): "An n -copula is a function $C(u_1, \dots, u_n)$ from $[0, 1]^n \rightarrow [0, 1]$ satisfying the following requirements:

- *Grounded*:
 $C(u_1, \dots, u_{i-1}, 0, u_{i+1}, \dots, u_n) = 0$ for all i and all $u_1, \dots, u_n$;
- *Copula marginal*:
 $C(u_1, \dots, u_{i-1}, 0, u_{i+1}, \dots, u_n)$ is an $(n - 1)$ -copula for all i ;
- *n -Increasing*:
 $V_{[\mathbf{u}, \mathbf{v}]} \geq 0$ for all rectangles $[\mathbf{u}, \mathbf{v}]$."

As with any basic statistics understanding of dependence we naturally expect to see versions of complete independence as well as full dependence extremities. To review such relationships in the copula context, we first re-iterate the Cherubini et al. (2012) requirements for copulas, and accordingly rewrite these for the bivariate case where U and V are random uniform variables, and we therefore have:

- *Grounded*: $C(0, v) = C(u, 0) = 0$;
- *Uniform marginals*: $C(u, 1) = u$ and $C(1, v) = v$;
- *2-Increasing*: $C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_1) + C(u_2, v_2) \geq 0$ for $u_1 > u_2$ and $v_1 > v_2$.

Using the notation and conclusion for this bivariate example we provide the general relationship as indicated by Mazo (2014:8):

$$W(u, v) \leq C(u, v) \leq M(u, v) , \quad (8.5)$$

where $(u, v) \in [0, 1]^2$. In the given relationship $W(u, v) = \max(u + v - 1, 0)$ is the minimum copula, $C(u, v) = uv$ is the product copula, and $M(u, v) = \min(u, v)$ is the maximum copula. It is explained that the maximum and minimum copulas form the Fréchet-Hoeffding boundaries in a general sense in Nelsen (2006:11). For specific copulas such bounds change to reflect the specific copula's limits. We see a helpful illustration of the minimum, product and maximum copulas in Figure 8.2⁴.

8.4 Copula families and their attributes

8.4.1 Introduction

In Cherubini et al. (2012) there is an extensive discussion on the types of copulas. These include survival copulas, their underlying radial symmetry, the exchangeability of copulas, hierarchical and vine copulas. We are not featuring those discussion here, since our study is aimed at a single type of copula. We therefore point the reader to the discussions in Nelsen (2006), Cherubini et al. (2012), and Mazo (2014). Our proposed solution in fact focuses on another special type of copula, namely that of factor copulas, which we present in detail in Section 9.1. However, our progress onto factor copulas is initiated by considering the families of copulas.

⁴We amended the original graphs from Nelsen (2006) to align to the notation in Mazo (2014).

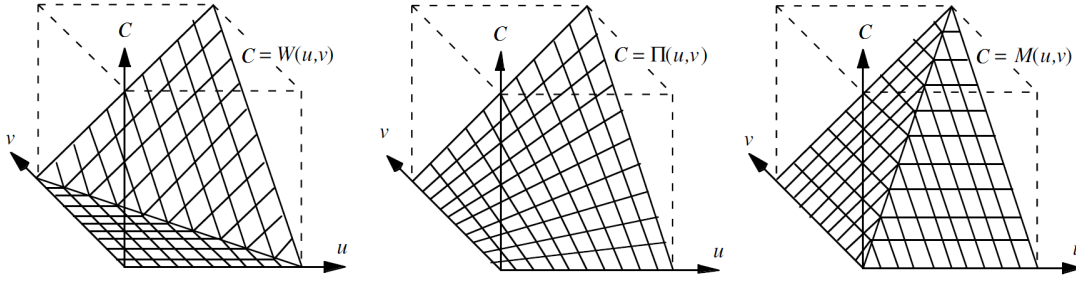


Figure 8.2: Illustration of the minimum (W), product (Π), and maximum (M) copulas on the unit square

SOURCE: Nelsen (2006:12); reformatted by author

8.4.2 Elliptical copulas

Overview

As clarified in Cherubini et al. (2012), the classification of copulas as *elliptical* is due to the underlying distribution being elliptical. Cherubini et al. (2012) goes on with an exposition of the derivation of elliptical distribution as affine transformation from spherical distributions. We do not consider this approach as central to our study and thus secondarily refer the reader to Fang et al. (1990) as referenced in Cherubini et al. (2012).

For our understanding of elliptical copulas we follow the overview as provided in Cherubini et al. (2012:29). The definition we see here is defined in terms of the Gaussian copula; therefore we know that all the contributing marginal distributions are standard normal distributions.

We note $\Phi(x)$ defined as a univariate standard normal distribution and $\mathbf{R}$ is an n -dimensional correlation matrix. Using these inputs, the joint normal distribution can be defined as

$$C(u_1, \dots, u_n) \equiv \Phi_n(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n); \mathbf{R}) . \quad (8.6)$$

This definition is easily extendible to the *Student's t*-distribution. This distribution is important for dependence study since its additional parameter, ν , the degrees of freedom allows for creating copulas with heavy tails such as is applicable in operational risk loss distributions. By redefining the joint distribution for a *Student's t*-distribution, where $\mathbf{T}_\nu(x)$ indicates a univariate *Student's t*-distribution. The *Student's t*-copula then follows as:

$$C(u_1, \dots, u_n) \equiv \mathbf{T}_n(\mathbf{T}_\nu^{-1}(u_1), \dots, \mathbf{T}_\nu^{-1}(u_n); \mathbf{R}, \nu) . \quad (8.7)$$

As noted thus far, elliptical copulas do not have a closed mathematical form, which is a clear difference from Archimedean copulas which we review in Section 8.4.3.

Tail behaviour

For our study of an operational risk problem, we specifically consider the attributes associated with tail behaviour of the respective copulas.

The first elliptical copula we defined is that of the Gaussian copula. This copula we note has no tail dependence; i.e. the copula structure does not accommodate any concordance / discordance of the varying input variables (Cherubini et al., 2012:30). Given this context, the *Student's t*-copula is considered to be an improvement on the Gaussian since it allows for association between the underlying variables. With the inclusion of the parameter for degrees of freedom in the *Student's t*-copula - parameter ν - the copula structure then generate heavy tails.

We mention here that elliptical copulas are often linked to distributions which are symmetric (Cherubini et al., 2012:31); i.e. this symmetry is then also translated to the tail behaviour of our selected copulas. In Section 8.6.4 we review the measurement of such tail behaviour with our discussion of tail dependence as a measure of association.

8.4.3 Archimedean copulas

Overview

For Archimedean copulas, as with elliptical copulas, we do not review the underlying mathematical derivation. We refer the reader to Nelsen (2006:109) for the details. For our understanding of the structure, and specifically the construction, of Archimedean copulas, we follow the explanation provided in Cherubini et al. (2012:32).

We see the argument that the easiest approach to constructing an Archimedean copula, is to obtain its generator function. Cherubini et al. (2012) defines a generator function as $\phi(x)$ along with its inverse, $\psi(x) = \phi^{-1}(x)$. The inverse generator function requires the following attributes to be able to construct an Archimedean copula:

- $\psi(x)$ must be infinitely differentiable⁵;
- the derivatives in their ascending order must be of alternating signs;
- $\psi(0) = 1$;
- $\lim_{x \rightarrow +\infty} \psi(x) = 0$.

As mentioned in Schmidt (2006:10), Archimedean copulas can be stated directly; i.e. they have closed-form formulae. We therefore consider specific well-known Archimedean copulas, considering both their generator functions as well as the actual copula formula.

The generator function of the *Clayton* copula is defined as:

$$\phi(x) = \frac{1}{\theta} \left(x^{-\theta} - 1 \right) , \quad (8.8)$$

whereas the generalized Clayton copula is then expressed by:

$$C_{\theta}^n(\mathbf{u}) = \left(\max[u_1^{-\theta} + u_2^{-\theta} + \dots + u_n^{-\theta} - 1, 0] \right)^{-\frac{1}{\theta}} . \quad (8.9)$$

⁵It is noted in CHERUBINI, U., GOBBI, F., MULINACCI, S. & ROMAGNOLI, S. 2012. Dynamic copula methods in finance, John Wiley & Sons Ltd., London, UK., that this attribute may be relaxed to a point where the inverse generator function is only differentiable $(n - 2)$ times.

with parameter $\theta \geq -1$.

Schmidt (2006:10-11) also considers the Gumbel(-Hougaard) copula with generator function:

$$\phi(x) \equiv (-\ln x)^\alpha, \quad (8.10)$$

and the actual Gumbel copula with parameter $\theta \geq 1$ is given as:

$$C_\theta^n(\mathbf{u}) = \exp\left(-\left[(-\ln u_1)^\theta + (-\ln u_2)^\theta + \dots + (-\ln u_n)^\theta\right]^{\frac{1}{\theta}}\right). \quad (8.11)$$

We mentioned these two copulas specifically due to their tail behaviour.

Tail behaviour

In Venter (2002:90) we see that the Clayton copula has its probability concentrated around the coordinate (0,0). It is explained that such an attribute is not proper for property-liability claims where the study originates. We agree that this is also the case for operational losses where we are expecting to see high severity losses which are associated in some dependence manner.

The second copula we discussed, the Gumbel copula, is of interest for our study since not only does it have strong association in its upper tail (as does the Frank copula; not previewed previously), but the Gumbel copula's association is also *dynamic* (Venter, 2002:72). This terminology used by Venter (2002) serves to denote that the association is asymmetrical; i.e. the association is more pronounced for one of the underlying variables.

As mentioned, we review the measurement of tail behaviour in Section 8.6.4.

8.4.4 Closing remarks

The reader may notice that we do not give any attention to extreme value theory. As pointed out in Joe (2015:34), using extreme value theory one can estimate extreme quantiles on the assumption that the underlying density is exponential or an inverse polynomial tail. We consider revisiting the issue of high quantiles an unnecessary repetition of Problem I. Furthermore, testing for the assumption of said exponential or inverse polynomial deviates from our envisioned simplification towards robust factor copula modelling.

We thus summarize this section as follow: we have obtained a basic overview of the copulas families which informs our further study on using copulas for dependence structure optimization. We firstly saw that elliptical copulas rarely have closed-form solutions or rather, representations. These copulas are often created of symmetric distributions of the underlying stochastic variables. For tail behaviour we saw that the *Student's t*-copula can accommodate heavy tails as we would like to study in our problem solution.

We then focused on Archimedean copula which do have closed-form expression, but are often rather expressed in terms of their generator functions. An interesting aspect of Archimedean copulas in terms of their tail behaviour is that they can accommodate asymmetric tail behaviour of the underlying stochastic variables.

Retaining our focus on copulas' tail behaviour, we now review how dependence as inherently established by a copula can be measured. We specifically then also focus on how the *tail dependence* is measured.

8.5 Goodness-of-fit techniques

8.5.1 Introduction

After we fit our copulas the next consideration is that of inherent model risk. In compiling a proposed solution we want assurance that the model is a proper reflection of the risk environment and that it is providing reasonable results in relation to its original purpose.

In order to get this assurance we assess our fitting according to goodness-of-fit. As distinguished in Chernobai et al. (2007:201), we consider the testing approaches for so-called 'in-sample' testing as data paucity remains a problem. In terms of graphical diagnostics we utilize the well-known technique of qq-plots and therefore do not discuss the details here. We specifically discuss the techniques for two empirical tests: the Cramér-von Mises and Anderson-Darling tests.

8.5.2 Cramér-von Mises test

Overview

The opening argument provided by Genest et al. (2009) is that so-called 'blanket tests' prove to be the best for assessing the goodness-of-fit when fitting copulas. These tests are highly robust since there is no need to account for parameters, weighting or kernel functions.

Genest et al. (2009) describe the test as a 'Monte Carlo experiment' where both the sample size and level of dependence during fitting is assessed. This assessment then concludes with respect to the robustness of blanket tests with the possibility of comparing between various copula.

Two approaches are mentioned by Genest et al. (2009), but we only consider the test based on non-parametric marginal distributions since we follow that approach in our proposed solution's assessment.

Test details

Firstly, Genest et al. (2009:200) define the null hypothesis test as:

$$H_0 : C \in \mathcal{C}_0 \quad , \quad (8.12)$$

which tests whether the dependence structure of our multivariate distribution is appropriately suited to a specified copula family, $\mathcal{C}_0$. Testing for the null hypothesis H_0 is therefore based on the inference of ranks of underlying variables. Genest et al. (2009) specifically indicates that these ranks can be considered as 'pseudo-observations' which arise from the rank transformation of variables; i.e. $\mathbf{U}_1 = (U_{11}, \dots, U_{1d}), \dots, \mathbf{U}_n = (U_{n1}, \dots, U_{nd})$.

However, Genest et al. (2009) caution that said pseudo-observations are by definition not mutually independent and that their underlying distributions only approximates a uniform distribution on the interval $[0, 1]$. It is these two aspects which Genest et al. (2009) include, and account for, in their test design.

Genest et al. (2009) explain that the information as represented by the pseudo-observations can be expressed by a concomitant empirical cumulative distribution function where:

$$C_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(U_{i1} \leq u_1, \dots, U_{id} \leq u_d) \quad ; \quad (8.13)$$

where $\mathbf{u} = (u_1, \dots, u_d) \in [0, 1]^d$.

Using this rank-based interpretation in the standard Cramér-von Mises test which we see defined as:

$$S_n = \int_{[0,1]^d} \mathbb{C}_n(\mathbf{u})^2 dC_n(\mathbf{u}) \quad . \quad (8.14)$$

Genest et al. (2009) postulate that a large test statistic, S_n leads to rejection of H_0 and add that p -values ought to be approximated from their limiting distributions.

Below we also provide the test statistic's expression as defined in Chernobai et al. (2007:210) since it facilitates understanding of the Anderson-Darling test which follows. The Cramér-von Mises test statistic is alternatively given by:

$$S_n = n \int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2 dF(x) \quad . \quad (8.15)$$

8.5.3 Anderson-Darling test

Overview

Chernobai et al. (2007:209) distinguish between a supremum and ‘quadratic class’ of the Anderson-Darling test. As is clear in our formulaic expression of the test statistic, we follow the quadratic form. It is explicated that, unlike the well-known Kolmogorov-Smirnov test, the Anderson-Darling test specifically allows for testing in an environment where a heavy tail is expected.

We see that Anderson-Darling follow the same basic null hypothesis definition as that of the Cramér-von Mises test (Marsaglia & Marsaglia, 2004). The most basic test of Kolmogorov-Smirnov has the approach of determining the maximum *distance* between the cumulative distribution function and the straight-line function, $y = x$, whereas the Cramér-von Mises test then squares this distance - thus obtaining the *area* of the difference between the cumulative distribution function and the straight line function. The Anderson-Darling provides a further improvement on Cramér-von Mises by incorporating a weighting function based on the standard deviations of the distance measurement points. It is this component which then creates greater sensitivity for tail behaviour.

Test detail

The Anderson-Darling test statistic is then expressed by:

$$A_n = n \int_0^1 \frac{[F_n(x) - x]^2}{x(1-x)} dx \quad (8.16)$$

$$= n \int_0^1 \frac{[F_n(x) - x]^2}{x} dx + n \int_0^1 \frac{[F_n(x) - x]^2}{1-x} dx \quad . \quad (8.17)$$

The approximation of this integral calculation is given as:

$$A_n = -n - \frac{1}{n} \left[\ln(x_1(1-x_n)) + 3 \ln(x_2(1-x_{n-1})) + 5 \ln(x_3(1-x_{n-2})) + \dots + (2n-1) \ln(x_n(1-x_1)) \right] \quad . \quad (8.18)$$

Marsaglia and Marsaglia (2004) report that Anderson and Darling (1954) applied numerical integration and obtained the values for the 90th, 95th and 99th quantiles which are 1.933, 2.492 and 3.857 respectively.

8.5.4 Closing remarks

In this section we reviewed the underlying theory of two important goodness-of-fit tests for in-sample cases. We noted the possibility of graphical diagnostics as mentioned in Chernobai et al. (2007) which we include in our testing on the copula fitting.

We most importantly note the relationship between the Cramér-von Mises and Anderson-Darling tests. Whilst the former can provide an immediate indication of the suitability of a fitting, further testing is required by means of the Anderson-Darling test to account for the possible presence of tail dependence. This being the very objective of our proposed solution, we initiate our testing with Cramér-von Mises, but require a final decision via Anderson-Darling.

Once the copula has an appropriate fitting we wish to investigate the fitting further by examining the dependence structures which the copula fitting has put in place. We review such techniques in the following section.

8.6 Dependence: definition and measurement

8.6.1 Overview

In this section we now return to the core function of copulas. The final use of the copula is to measure between variables across their joint distributions. We see this point approached in Nelsen (2006:157) in the context of *concordance*. As known from general statistical theory, the shortcomings of the linear dependence of correlation coefficients need to be overcome by more robust yet flexible ‘measures of association’.

8.6.2 Kendall’s τ

Relating the common measure of association, Kendall’s (τ), to the copula environment Nelsen (2006) states and proves that a so-called *concordance function* of the differences

in probabilities of concordance and discordance between vectors of two continuous stochastic variables, only depends on their marginal distributions by way of their combining copula. We consider this result here.

We observe the definition of two independent vectors (X_1, Y_1) and (X_2, Y_2) of continuous stochastic variables where their joint distributions are given as H_1 and H_2 . Following the definition seen earlier in Equation 8.1, we have $H_1(x, y) = C_1(F(x), G(y))$ and $H_2(x, y) = C_2(F(x), G(y))$ with C_1 and C_2 indicating the copulas of the respective independent vectors. Nelsen (2006:159) then defines the concordance function, Q , as

$$Q = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0] , \quad (8.19)$$

and proves⁶ the result that

$$Q = Q(C_1, C_2) \quad (8.20)$$

$$= 4 \iint_{\mathbf{I}_2} C_2(u, v) dC_1(u, v) - 1 . \quad (8.21)$$

Nelsen (2006:162) then continues by explaining Kendall's τ as the 'first concordance set' and redefines the obtained integration result as the expected value on the function $C(u, v)$ of the uniform variables U and V with joint distribution C

$$\tau_C = 4 E[C(U, V)] - 1 . \quad (8.22)$$

These concise results are our tool for determining the measure of association between our operational risk loss variables as combined in our copula function.

However, in the case of Archimedean copulas, the process of determining Kendall's τ can be simplified even further

$$\tau_C = 1 + 4 \int_0^1 \frac{\phi(t)}{\phi'(t)} dt . \quad (8.23)$$

8.6.3 Spearman's ρ

A further measure of association which we consider for the copula environment is that of Spearman's ρ . Nelsen (2006:167) provides a derivation argument analogous to that of Kendall's τ . We provide the details here.

We see three independent stochastic vectors defined (X_1, Y_1) , (X_2, Y_2) , and (X_3, Y_3) . The joint distribution function is again H with marginals F and G . Spearman's ρ is then defined as a ratio of the probability of *concordance* less the probability of *discordance* of two hypothetical vectors with the same margins on the condition that one vector follows the distribution H . Thus, we have (X_1, Y_1) and (X_2, Y_3) , and we note

$$\rho_{X,Y} = 3P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0] . \quad (8.24)$$

⁶We include the full extended proof hereof in Appendix B

In reference to concordance axes, and Kendall's τ being the first concordance set, Spearman's ρ is then defined as the 'second concordance set'. As with the Kendall's τ measure, it follows that⁷

$$Q = 3 Q(C, \Pi) \quad (8.25)$$

$$= 12 \iint_{\mathbf{I}_2} uv \, dC(u, v) - 3 \quad (8.26)$$

$$= 12 \iint_{\mathbf{I}_2} C(u, v) \, dudv - 3 . \quad (8.27)$$

Note that the second of the two copulas is the product copulas because X_2 and Y_3 are independent - the defined discordance.

In conclusion, we again see the integration result redefined for expected values on the copula function:

$$\rho_C = 12 E[C(U, V)] - 3 . \quad (8.28)$$

8.6.4 Tail dependence

A concomitant measure of association is 'tail dependence' or 'tail index'. As illustrated measures of association delineates the extent to which one stochastic variable's large / small values vary in relation to another's large / small values. We see Nelsen (2006:214) explain tail index as a comparable measure since it determines the association of the variables in the upper right and lower left quadrant of the unit square $\mathbf{I}^2$.

We consider the Nelsen (2006:162) definition of two continuous stochastic variables X and Y with F and G as their distribution functions respectively. The *upper* tail index is then given as

$$\lambda_U = \lim_{t \rightarrow 1^-} P[Y > G^{(-1)}(t) | X > F^{(-1)}(t)] , \quad (8.29)$$

whilst the *lower* tail index follows analogously

$$\lambda_L = \lim_{t \rightarrow 0^+} P[Y \leq G^{(-1)}(t) | X \leq F^{(-1)}(t)] . \quad (8.30)$$

The most important result which Nelsen (2006) proves is that these indices are '*nonparametric and depend only on the copula of X and Y* '. We provide the proof of this results below.

The interpretation of these results is then that a copula has upper tail dependence if $\lambda_U \in (0, 1]$. If $\lambda_U = 0$ the copula has no upper tail dependence. The interpretation follows analogously for λ_L . We only note the case for the upper tail index here as detailed in Nelsen (2006).

⁷A point we mention in passing is the so-called *normalization constant* '3' which we consider beyond the scope of our study and refer the reader to Nelsen (2006).

If the upper tail index exists, then we calculate its value as:

$$\lambda_U = \lim_{t \rightarrow 1^-} P[Y > G^{(-1)}(t) | X > F^{(-1)}(t)] \quad (8.31)$$

$$= \lim_{t \rightarrow 1^-} P[G(Y) > t | F(X) > t] \quad (8.32)$$

$$= \lim_{t \rightarrow 1^-} \frac{\bar{C}(t, t)}{1 - t} \quad (8.33)$$

$$= \lim_{t \rightarrow 1^-} \frac{1 - 2t + C(t, t)}{1 - t} \quad (8.34)$$

$$= 2 - \lim_{t \rightarrow 1^-} \frac{1 - C(t, t)}{1 - t} \quad (8.35)$$

$$= 2 - \delta'_C(1^-) , \quad (8.36)$$

where δ_C is the diagonal section of the copula.

In the case of Archimedean copulas, these tail indices can be evaluated directly using their generator functions and inverses.

8.7 Chapter conclusion

In our review of the core academic literature topics on copulas, we were able to acquire an understanding of the goal which copulas aim to fulfil; i.e. to address our concerns for expressing dependence structures when modelling, especially in environments where we have limited information on joint distributions, but significant background on variables' marginal distributions. This goal is directly linked to our study of Problem II where we are attempting to model interaction / dependence elements of our operational loss information as captured across business lines and event types across the period under study. As confirmed in Chapelle et al. (2004), incorporating dependence optimization in capital estimation may lead to a decrease in such capital estimates of at least 30%.

In tracking the historical development of copulas, we followed Nelsen (2006) when starting with Sklar's theorem on how the copula is defined as the '*manner*' in which two marginal distributions can be combined to obtain a joint distribution which we firstly, did not possess, and secondly, would also provide us with an indication of the underlying variables' co-movements. Continuing on from this classic exposition, we consulted the Cherubini et al. (2012) technical definition, specifically that of defining the copula's probability on a plane. We subsequently use these definitions to inspect key copulas; the minimum, product and maximum copulas.

Our study then deepened into understanding the copula families. We considered the theoretical derivation and / or definitions of both elliptical and Archimedean copulas. We provided examples of the most common copulas known and used in dependence measurement. We review tail behaviours of said copula in particular since we are expecting to apply such techniques in our study of operational losses where the tailedness of distributions is a key input to operational risk capital estimation.

A further step in using copulas was to evaluate the means by which we can determine whether a selected and fitted copula is appropriate given our data. For this purpose we consulted the overview provided in Cherubini et al. (2012) where the

dual approach of graphic and empirical diagnostics are explained. We focused on the theoretical definition of the empirical tests noting that generalized test for goodness-of-fit ought to be supplemented by tests which take into account the possibility of heavy tails in our data.

We concluded our review of basic copula literature topics by examining the techniques with which one can measure dependence as present in copulas. We discussed the basic derivation of the two common measure of Spearman's ρ and Kendall's τ and how these can be extracted from a copula. An interesting point to note was that an Archimedean copula's generator function can be used to obtain the Kendall's τ measure. There is a similar case for calculating tail dependence using generator functions, which is where one is interested in measuring dependence in the upper right or lower left quadrants of our copula probability plane on $\mathbf{I}^2$.

In the next chapter we focus on a specific type of copula, namely that of factor copulas. Factor copulas arise from applying conditioning when defining a copula. We examine this type of copula in detail along with a case study since it guides us in our proposed solution to Problem II of dependence structure optimization.

Chapter 9

Advanced literature topics

9.1 Introduction

In our problem statement for Problem II, we already mentioned that dimensionality becomes a problem in the case of operational risk loss data. This statement may not be so clear at face value. However, since we unlikely know the joint distribution of many units of measures (UoM), examining joint behaviour and underlying dependencies becomes a laborious task. This follows since we then have to evaluate for marginal distributions of the UoMs, determine the dependence structures and only then compile a joint distribution from which we estimate capital for operational risk.

The issue of dimensionality consists of the separate dimensions of the UoMs and how these then combine into a joint distribution. When considering the full set of ORCs from a basic operational risk matrix of measurement, this leads to 56-dimensional joint distribution. A powerful approach would be to ‘summarize’ these dimensions and then focus on modelling the dependence structure for the so-called ‘summary’. Copulas already provide a useful manner to obtain the joint distribution and thus estimate Value-at-Risk, but we still need to compress the dimensions involved in analyzing the dependence relationships.

As stated in (Zhang & Jiao, 2012:4), when applying a ‘factor approach’ we are able to significantly reduce the dimensionality of our data. If we thus use any of the common factor analysis techniques available to us - e.g. principal component analysis - we can thus isolate common factors and therefore rather focus on analyzing those common factors’ underlying dependence structures. This is achieved by using factor copulas.

In order to define factor copulas, we first review how factors are expressed in copula format; conditional probability expressed in copula notation. We then recall the popular one-factor model as designed by Vašíček (2002) and what its copula interpretation. Finally we see our core case study for two factors and consider the details of the technique. Using the definitions acquired throughout this literature we then construct our own two-factor t -copula which we propose as a solution for the problem of optimizing dependence structures for capital diversification benefits in Chapter 10.

9.2 Conditional probability

9.2.1 Introduction

With the composition, performance and measurement techniques of copulas already reviewed in Chapter 8, we first define the concept of ‘conditional probability’ in terms of copulas. We do this review since the conditionality concept strongly underlies the further exposition of factor copulas.

9.2.2 Theoretical overview

We revert to the derivation as provided in Cherubini et al. (2012:40) following the Bayesian principle so that we have

$$\mathbb{P}(X \leq x \mid Y = y) = \lim_{\Delta y \rightarrow 0} \mathbb{P}(X \leq x \mid y \leq Y \leq y + \Delta y) \quad (9.1)$$

$$= \lim_{\Delta y \rightarrow 0} \frac{F(x, y + \Delta y) - F(x, y)}{F_Y(y + \Delta y) - F_Y(y)} \quad (9.2)$$

$$= \lim_{\Delta y \rightarrow 0} \frac{C(F_X(x), F_Y(y + \Delta y)) - C(F_X(x), F_Y(y))}{F_Y(y + \Delta y) - F_Y(y)} \quad (9.3)$$

$$= \left. \frac{\partial C(F_X(x), v)}{\partial v} \right|_{v=F_Y(y)} . \quad (9.4)$$

It is explained in (Cherubini et al., 2012) that this proof requires further background that $\frac{\partial}{\partial u}C(u, v)$ and $\frac{\partial}{\partial v}C(u, v)$ exist for almost all u and v . There is also a proof in (Nelsen, 2006) that for u and v it holds true that $0 \leq \frac{\partial}{\partial u}C(u, v) \leq 1$ and $0 \leq \frac{\partial}{\partial v}C(u, v) \leq 1$ - see Appendix B.

This derivation provides a very powerful result; namely that a conditional probability is the partial derivative of the copula, with the derivation in the direction of the variable for which we are conditioning. (Cherubini et al., 2012) expresses this as

$$\mathbb{P}(X \leq x, Y \leq y) = \int_0^{F_Y(y)} \frac{\partial C(F_X(x), w)}{\partial w} dw . \quad (9.5)$$

We also consider the Palaro and Hotta (2006) definition of the log-likelihood for conditional copulas here (with amended notation to align to Cherubini et al. (2012)), as this likelihood function is our estimation method of choice through maximum likelihood estimation in our proposed solution.

By rewriting Equation 8.1 for conditionality we see the result:

$$H(x, y) = C(F(x; \vartheta_x), G(y; \vartheta_y)) , \quad (9.6)$$

recalling that $H(x, y)$ is the joint distribution of functions F and G , whilst $H(x, y; \vartheta)$ denotes the conditional joint distribution. By defining the conditional variables $X, Y \mid W$ with associated conditional (marginal) density functions $f(x \mid \mathbf{w})$ and $g(y \mid \mathbf{w})$ and applying the Cherubini et al. (2012) result from Equation 9.5 the copula density is expressed by:

$$c(u, v \mid \mathbf{w}) = \frac{\partial^2 C(u, v \mid \mathbf{w})}{\partial u \partial v} , \quad (9.7)$$

whilst the log-likelihood function for a sample size of T is given as:

$$l(\vartheta) = \sum_{t=1}^T \ln c(F(x; \vartheta_x | \mathbf{w}_t, G(y; \vartheta_y | \mathbf{w}_t))) + \sum_{t=1}^T (\ln f(x_t; \vartheta_x | \mathbf{w}_t) + \ln g(y_t; \vartheta_y | \mathbf{w}_t)) \quad . \quad (9.8)$$

9.3 The one-factor Gaussian copula

9.3.1 Introduction

In this section we present the one-factor Gaussian model as composed by Vašíček (2002) in order to have an anchor point for the extension to a two-factor Gaussian model (Zhang & Jiao, 2012), and finally our own extension to a two-factor *Student's t*-model. The literature review in Zhang and Jiao (2012:2) is quite sparse, but provides some background information to the conceptualization of the Vašíček-model as utilized in the secondary references of; Andersen and Sidenius (2004) and Patton (2006). Our own approach tracks the Vašíček-model's concept usage in portfolio theory, and then gives the theoretical derivation to better link to the theoretical reasoning in Zhang and Jiao (2012) and our own study.

Continuing from this basis, in order to support any portfolio with capital, the creator of said portfolio will need information on the probability distribution of the portfolio's losses. The capital allocated to this portfolio support must then be equal to the desired percentile of the portfolio's loss distribution.

This is the introduction to the Vašíček (2002) one-factor (Gaussian) model which we present in more detail below. He also mentions that the ability to calculate VaR from this method is useful for regulatory reporting; this indeed forms an integral part of our proposed solution.

9.3.2 Theoretical overview

The discussion is opened in Vašíček (2002) where we consider a risk event that a credit extension will default should the value of the borrower's assets at some time T when the extension matures, fall below some contractual value B .

Vašíček (2002:160) defines for a portfolio of assets that A_i is the asset value of the i -th borrower. This process is then expressed as:

$$dA_i = \mu_i A_i dt + \sigma_i A_i dx_i \quad . \quad (9.9)$$

where μ_i refers to the assets mean (expected value) and σ_i refers to its standard deviation (variability). At time T this asset value is rewritten as follow:

$$\log A_i(T) = \log A_i + \mu_i T - \frac{1}{2} \sigma_i^2 T + \sigma_i \sqrt{T} X_i \quad , \quad (9.10)$$

where X_i follow the standard normal distribution. Vašíček (2002) then continues by developing the problem for n credit loan extensions of equal value. We see the

assumption that the borrowers' asset values are correlated with a ratio of ρ when considered for any two companies.

Since the X_i in Equation 9.10 are *jointly* distributed as standard normal variables with equal ρ for pairwise correlations, Vašíček (2002) states the following expression to hold true:

$$X_i = \sqrt{\rho}Y + \sqrt{1-\rho}Z_i \quad , \quad (9.11)$$

where $Y, Z_1, Z_2, \dots, Z_n$ are all standard normal variables as well as independent of each other. The pairwise correlation between two borrowers I and J would thus be expressed as $\text{corr}(x_i, x_j) = \sqrt{\rho_i \rho_j}$.

Vašíček (2002) specifically points out that Y is taken as the 'portfolio common factor'. Equation 9.11 can then be interpreted as consisting of element $\sqrt{\rho}Y$ which represents a borrower's exposure to the *common* risk factor and element $\sqrt{1-\rho}Z_i$ in turn represents the same borrower's *specific* risk exposure. The closing remark of this exposition is that the loss probability is calculated as *conditional* given Y . It is specifically this conditioning which we are utilizing in our proposed factor copula methodology as derived from the conditional probability for copulas - Section 9.2.

With this one-factor base description now established, we remain with the Gaussian-type model, but consider an extension to two factors in the following section.

9.4 A two-factor Gaussian copula model

9.4.1 Introduction

We now consider factor copula models as developed by Zhang and Jiao (2012). In their study on exchange rate returns they construct dependence structures through various copulas by means of conditioning distributions on common factors. Given the core function of factor analysis to reduce dimensionality, this method facilitates the construction of dependence structures considerably.

The efficacy of the factor copula method is immediately apparent when Zhang and Jiao (2012:7) explain that instead of directly using the marginal distributions of each individual variable (exchange returns in their study), the method applies the marginal distributions which are conditioned for the common factors. This implies that we only need to extract a few common factors from the large set of variables.

9.4.2 Theoretical overview

Building on the one-factor copula as presented in Section 9.3, Zhang and Jiao (2012) compiles a two-factor copula model; i.e. using only two factors resulting from factor analysis on the original set of variables.

We use slightly adapted notation to have greater consistency across the various theoretical views.

Zhang and Jiao (2012:7) start off their theoretical formulation by reverting to Hull (2009). They transform their set of variables into new variables by means of the transformation function $x_i = \Phi^{-1}[Q_i(r_i)]$ with $i = 1, 2, \dots, n$. In this function Q_i is

the cumulative distribution function of the individual variables, and Φ^{-1} the inverse of the cumulative normal distribution. In this transformation the resulting variables x_i have a standard normal distribution; i.e. $x_i \sim \mathcal{N}(0, 1)$. We point out that this step already adds significant robustness to this method; with this transformation the marginal distributions and their fitting, become unnecessary. Zhang and Jiao (2012) continue on to explain the percentile-to-percentile transformation allows for the correlation structure of the original variables to be evaluated by the correlation structure of the new transformed variables, x_i .

Using the basic structure of the one-factor Gaussian copula as presented in Section 9.3, we see an extension to a two-factor Gaussian copula model:

$$X_i = \alpha_i F_1 + \beta_i F_2 + \sqrt{1 - \alpha_i^2 - \beta_i^2} Z_i \quad , \quad (9.12)$$

where Zhang and Jiao (2012) define F_1 and F_2 as the two common factors extracted by means of factor analysis from the correlation structure as obtained from the transformed variables, x_i . These factors are defined as factors commonly affecting all the variables as derived from a common financial market environment. Further, Z_i functions as a random error term and we see $Z_i \sim \mathcal{N}(0, 1)$.

The coefficients of said factors, α_i and β_i are constant parameters such that $\alpha_i, \beta_i \in [-1, 1]$. The correlation structure of the transformed variables is then expressed as $\text{corr}(x_i, x_j) = \alpha_i \alpha_j + \beta_i \beta_j$.

Equation 9.12 can then be rewritten as:

$$Z_i = \frac{X_i - \alpha_i F_1 - \beta_i F_2}{\sqrt{1 - \alpha_i^2 - \beta_i^2}} \quad ; \quad (9.13)$$

continuing from this step on, we can estimate the parameters using maximum likelihood estimation on Equation 9.13. Using the results a new correlation structure is compiled which can be applied in copula fitting and simulation.

It is with this last technique that we clearly see the reduction in dimensionality. A new correlation structure with the same dimensions as the full set of variables - whether transformed or original - can be calculated using only the two vectors α_i and β_i .

Zhang and Jiao (2012:8) also provide methods for Archimedean copulas where the variables are not transformed, but instead the common factors are extracted directly from the original variables. These common factors are not normally distributed, but rather Gamma. We do not consider further details of these processes as our proposed solution centres on elliptical copulas; we rather consider the (Student's) t -copula in the following section; and extend for two factors.

9.5 The two-factor t -copula

9.5.1 Introduction

Another elliptical copula with factor loading application is that of the t -copula. As opening argument Bluhm et al. (2003) illustrate in Figure 9.1 that moving from a Gaussian copula to a t -copula firstly allows for tailedness; i.e. high dependence is seen for high concordant values of the underlying variables. Secondly, in terms of spread

we notice that when moving from normal to (Student's) t -marginal distributions of the underlying variables adds to a wider spread of the dependence structure; i.e. retaining normal marginal distributions leads to copulas densely structured around a central so-called ‘cloud’.

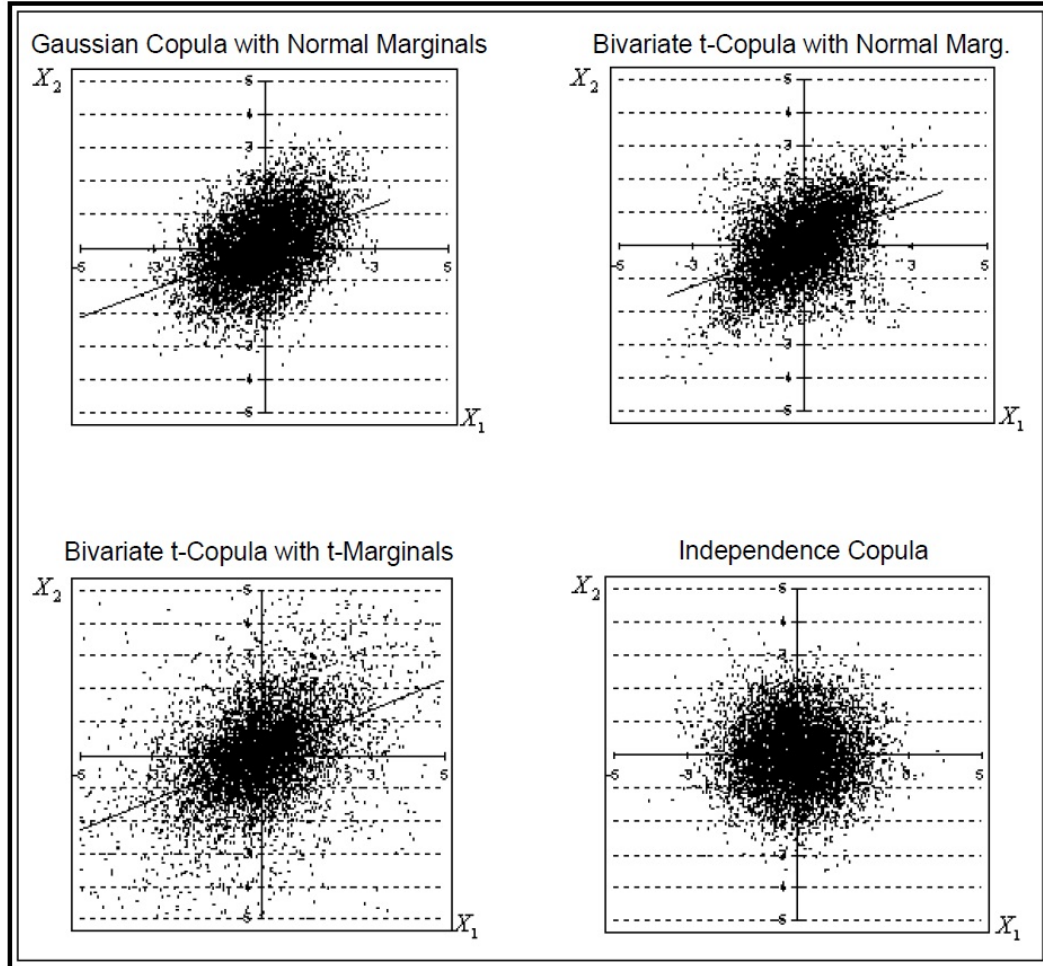


Figure 9.1: Illustration of elliptical copulas on spread and tailedness: Gaussian vs. t -copulas

SOURCE: Bluhm et al. (2003:105)

9.5.2 Theoretical overview

For this section, we again adapt the notation for consistency with the preceding expressions.

In their theoretical formulation, Bluhm et al. (2003:107) also initiate the definition by recalling the one-factor Gaussian copula (Section 9.3). When the base one-factor Gaussian copula model is scaled with a ratio of $\sqrt{\nu/W}$, where $W \sim \chi^2(\nu)$; the Gaussian copula is changed into a t -copula with underlying t -distributed marginals. We thus have the underlying marginals as $\tilde{x}_i \sim t(\nu, \Gamma_\rho)$.

Taking W as independent of Y and Z_i , the one-factor Gaussian copula is thus

extended as follow:

$$\tilde{X}_i = \sqrt{\frac{\nu}{W}} X_i \quad (9.14)$$

$$= \sqrt{\frac{\nu}{W}} \left(\sqrt{\rho_i} Y + \sqrt{1 - \rho_i} Z_i \right) , \quad (9.15)$$

and is distributed as $t(\nu)$ with $i = 1, \dots, m$ and ν indicating the degrees of freedom associated with the t -distribution.

We wish to point the reader to the degrees of freedom, ν . Note that unlike α_i and β_i , which have vector-type structures (alternatively single-column matrices), ν is of a scalar-type format; it is a single point to be estimated.

Using this result from Bluhm et al. (2003), we rewrite this expression analogously to Equation 9.13 to incorporate the two-factor element, and obtain the following:

$$Z_i = \frac{\sqrt{\frac{W}{\nu}} \tilde{X}_i - \alpha_i F_1 - \beta_i F_2}{\sqrt{1 - \alpha_i - \beta_i}} . \quad (9.16)$$

Note that it is a ν -scaled variable used in the aforementioned equation.

The parameters are also scaled with the degrees of freedom. Consulting Wang et al. (2009:280), we confirm this scaling ratio as $\frac{\nu}{\nu-2}$ - also refer to Neugebauer (2007:12). For our two-factor t -copula, we thus extend this scaling ratio and define the correlation structure generating function as $\text{corr}(\tilde{x}_i, \tilde{x}_j) = \frac{\nu}{\nu-2}(\alpha_i \alpha_j + \beta_i \beta_j)$.

9.6 Chapter conclusion

In this chapter we confronted the issue of dimensionality in dependence structures for operational risk more directly. We investigated a specialized technique for using copulas to evaluate dependence structure. This technique, factor copulas, is proposed since it heavily reduces the high dimensionality commonly associated with modelling dependencies for operational risk.

Before developing the theory for factor copula, we considered the conceptualization of factors in copula theory. We saw that factor copulas are based on *conditional probability* as defined for copulas. The underlying theory confirmed a noteworthy result; the conditional version of a copula is obtained by calculating the partial derivative of the copula, deriving in the direction of the conditioning variable.

We then reviewed Vašíček's one-factor model and how it relates to modelling with copulas. This result was then extended to two factors for a normal distribution (or Gaussian) copula on the basis that two factors often explain a sufficient proportion of variance of the original variables; thereby accommodating significant dimension reduction.

Finally we applied the preceding information on factor copulas and the different distributional interpretations, and we then constructed our own two-factor model. Our proposed two-factor copula model makes use of the Student's t -distribution which allows us to model dependence with tail behaviour. We express this two-factor

t -copula by:

$$Z_i = \frac{\sqrt{\frac{w}{\nu}} \tilde{X}_i - \alpha_i F_1 - \beta_i F_2}{\sqrt{1 - \alpha_i - \beta_i}} \quad . \quad (9.17)$$

We now propose our solution to Problem II in the following chapter; incorporating the existing Gaussian two-factor copula from Zhang and Jiao (2012) as well as our definition of a two-factor t_ν -copula.

Chapter 10

Proposed solution

10.1 Introduction

In this chapter we provide our proposed solution for Problem II on optimizing dependence structures to facilitate capital estimation with diversification benefits.

We discuss our data environment as arising from the output Problem I; i.e. the pooled scaled dataset. We detail our handling and preparation of the data to ‘create’ a larger dataset for more reasonable modelling results. A short analysis provides an indication of the underlying dependence structure.

We discuss the modelling steps as derived from Zhang and Jiao (2012) and include and our own extension of a two-factor t -copula. For said modelling process we report on how we use the factor analysis and dependence generators for dimension reduction in copula modelling. We end the modelling discussion by providing result on copula fitting and which copulas show the desired behaviour of heavier upper tails.

We conclude our proposed solution by conducting an impact analysis. This analysis consists of capital estimation for operational risk; we reflect on the changes between undiversified capital estimation and the capital charges obtained when using the various copulas.

10.2 The data environment

10.2.1 Introduction

In this section we document our data environment as used for proposing a solution to Problem II. We report the details on our data preparation where several data issues commanded solutions before allowing further modelling; e.g. choice of periodicity.

Once the dataset has suitable attributes for modelling we do very basic data analysis to obtain an overall impression of the dependence structures inherent to our raw dataset.

10.2.2 Data preparation

Choice of dependence unit of measure

Our base raw dataset is that which we obtained as result from our proposed solution to Problem I - see Section 6.4.2. This is the starting set for our further analysis for Problem II; all analytics results from said section are thus applicable in principle for this *base* dataset.

Our approach to the optimizing the dependence structure is built on intercell dependencies; i.e. the dependencies as expected between the aggregate losses - ($Loss_{aggregate} = Severity \circ Frequency$), see Section 2.4.2 - of the 56 ORCs. Other dependence units of measure (UoM) are available for modelling - dependence can also be tracked across BLs and ETs. We suspect a risk manager functioning at a portfolio-level within a financial institution would best be able to decide between such options; or in fact decide to investigate all possible options.

Our choice of ORC intercell dependence UoM, however, is not restricted in application. We first constructed our ‘beta’ version of the proposed solution on both ETs and BLs before settling and reporting on ORCs as dependence UoM.

We note that at this stage, each of the underlying components of the ORC UoM contains data for some of the row observations; i.e. the time component of how losses are captured. We start addressing this aspect in the subsequent section.

Choice of accumulation methods

A frequent practice to initiate dependence analysis for operational risk is to sum losses per UoM across each year of observation. For our pooled dataset this would provide us only eight observations (i.e. eight years) at most.

Also, due to our choice of ORCs as UoM - the most granular UoM - data availability becomes a problem. As explained by Brunel (2014:179) this is a common problem for operational risk dependence modelling. Using the ORCs to analyze dependence at its most granular creates a dependence structure with very high dimensionality - 56 as the number of ORCs. Preparing a dataset with a suitable number of observations for dependence analysis we conduct an iterative process between the number of observations and the inclusion of all the UoMs.

In order to ‘generate’ more data we divide the annual summation for semesters, trimesters and quarters. Following this approach we are able to obtain the expected 16, 24 and 32 observations for the respective accumulations methods. Focusing on the UoMs we are compelled to reconsider the inclusion of UoMs. Inspecting for UoMs across the different accumulation methods we remove ORCs which we do not more than 25% of observations missing. When filtering on this rule we have a final set of 15 ORCs.

Based on these various filters we see that the semester set is complete, the trimester set is 88% complete, and the quarter set is 56%. We therefore select the trimester accumulation as the most ideal for our analysis, allowing us 21 complete observations.

10.2.3 Data analysis

Initial correlation structure

Based on our dataset for ORC intercell dependencies - 21 observations across 15 variables - we compile a simple correlation structure to gauge the dependence structure. We obtain the correlation diagram as presented in Figure 10.1.

When considering only in terms of the matrix's triangle - whether upper or lower - it is clear from the diagram that more than half of the ORCs show positive correlations, approximately 60%. On average the positive correlations are stronger than negative associations.

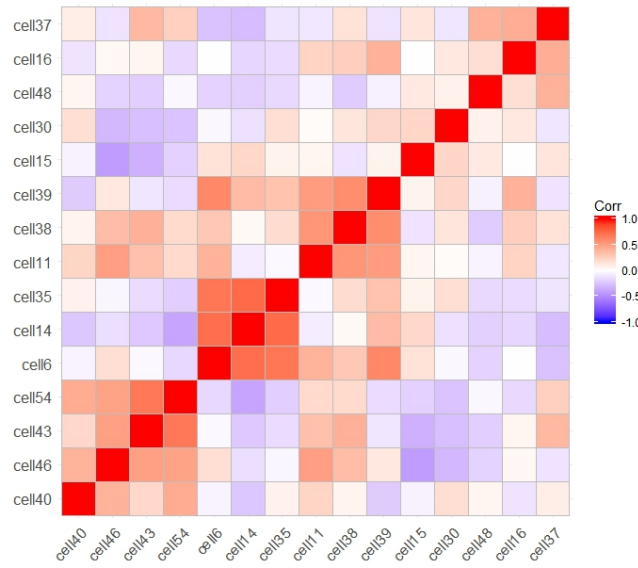


Figure 10.1: Correlation structure of the pooled dataset

Adapted correlation structure

As explained in Groenewald (2014:354), negative correlations cannot be justified when applied for capital estimation. We therefore substitute all negative values with a small value to avoid this justification in capital estimation. However, as we discuss in Section 10.3, we first perform the factor analysis and conditioning to obtain the dependence generating structures. Once these final structures are obtained we adapt the correlation structures for positive-only values prior to initiating the copula simulations.

10.3 Two-factor copulas

10.3.1 Introduction

Zhang and Jiao (2012:2) explain their method as a rework of the well-known one-factor copulas. An extended theoretical discussion on the well-known one-factor (Gaussian) models is presented in Andersen and Sidenius (2004). The application of this technique is the first idea behind the Zhang and Jiao (2012) method. A second

concept is then included as arising from bivariate copulas; i.e. a method where the factor copula model is extended to consist of two factors.

In our study we use the methodology as compiled by Zhang and Jiao (2012:7), specifically also modelling for the Gaussian copula conditioning on two factors. However, we do not follow the further investigation on Archimedean copula applications, but instead create our own two-factor t -copula to allow for heavy tailedness. We amend the Gaussian copula's likelihood function for estimation along with the dependence generating parameter estimates.

With respect to the Zhang and Jiao (2012) method we transform the ORC variables to standard normal marginals for our Gaussian copula, as well as to Student's t -marginal distributions with selected degrees of freedom for various (Student's) t -copulas.

10.3.2 Variable transformation

Following the Zhang and Jiao (2012) method we transform the raw ORCs variables in two ways using the transformation of $x_i = \Phi^{-1}[Q_i(r_i)]$ for standard normal marginals and $x_i = t_\nu^{-1}[Q_i(r_i)]$ for t -marginals - see Section 9.4.2. Recall that the transformation step of $Q_i(r_i)$ is the cumulative distribution function and for our method we use the empirical cumulative distributions for each of the respective ORCs variables.

10.3.3 t -copula - degrees of freedom

We acknowledge the fact that a Student's t -distribution converges to a standard normal distribution once the degrees of freedom become large. We specifically want to avoid this as a Gaussian copula based on the normal distribution does not have adequately heavy tails for modelling extreme risk event scenarios which we want to incorporate in our proposed solution. Also, Afambo (2006:82) explains that copula theory predicts higher capital charges associated with low degrees of freedom since the implication is that numerous extreme events occur 'simultaneously'. We therefore aim to model various t -copulas to continue searching for a *conservative* solution regarding diversification benefits whilst still aiming for low degrees of freedom to create heavy tails.

As confirmed in Hull and White (2004:23), the most ideal results - in terms of tailedness - were obtained with degrees of freedom ranging from three through five. Their results actually point to $\nu = 4$ which proves to be the best. A further argument is made in Afambo (2006) that estimation of ν for various sets of variables is often infeasible when considering loss data due to their paucity. We also expect that convergence between our ORCs variables when estimating degrees of freedom is not a given, moreover for large sets of variables.

Based on these various views we attempt to construct a fusion of alternatives by choosing to model three t -copulas based on three fixed degrees of freedom; i.e. $\nu \in (3, 4, 5)$ within the so-called ideal range for loss estimation, and specifically operational losses.

10.3.4 Factor analysis

Overview

As with the Zhang and Jiao (2012) methodology we perform factor analysis on our transformed variables; i.e. for the four respective sets of the normal marginal transformation and the three t -marginal transformations - $\nu \in (3, 4, 5)$. We discuss the results of the factor analysis in the next section.

Results

We conduct a principal component analysis to extract two common factors from our four datasets to use in our two-factor copula models. We first consider the results for our normal marginals transformation, reporting only on the first eight factors.

In Table 10.1 we illustrate the results of factor analysis, listing the eigenvalues and proportion of variance explained by the factors. By inspecting the eigenvalues we note that four factors are positive for a value above 1.00 as cut-off, although the fourth factor's eigenvalue can be considered a boundary value. This result is also clear from the scree plot in Figure 10.2.

Table 10.1: Proportion of variance - normal marginals transformation

Factors	Eigenvalue	Difference	Proportion	Cumulative
1	5.83810	2.73500	0.389	0.389
2	3.10310	0.59323	0.207	0.596
3	2.50986	1.47981	0.167	0.763
4	1.03004	0.05166	0.069	0.832
5	0.97838	0.44087	0.065	0.897
6	0.53750	0.17703	0.036	0.933
7	0.36047	0.12049	0.024	0.957
8	0.23997	0.02101	0.016	0.973

In terms of the proportion of variance explained by the factors extracted, we see in Table 10.1 and Figure 10.2 that a proportion of 90% is only reached by around the fifth factor. Returning to Zhang and Jiao (2012) the proportion of variance explained is only reported for the first two factors; their general proportion settles around 75%. In our analysis this level is only achieved for three factors. However, we explicitly point out here that Zhang and Jiao (2012:4) already started with fairly low dimensionality - only 4 original variables were to be accounted for through factor analysis.

We note the results of the scree plot and supporting graph for 'variance explained', and highlight the differences between factors' proportions of variance. The increase in proportion of variance between factors 2 and 3 appears to be fairly low before a large increase between factors 3 and 4 is seen.

Besides our method being constructed for two factors the increase in proportion of variance explained may not be sufficiently large to motivate for the inclusion of

a third factor. As seen in Table 10.1 our first two factors allow for a proportion of 60% of variance to be explained.

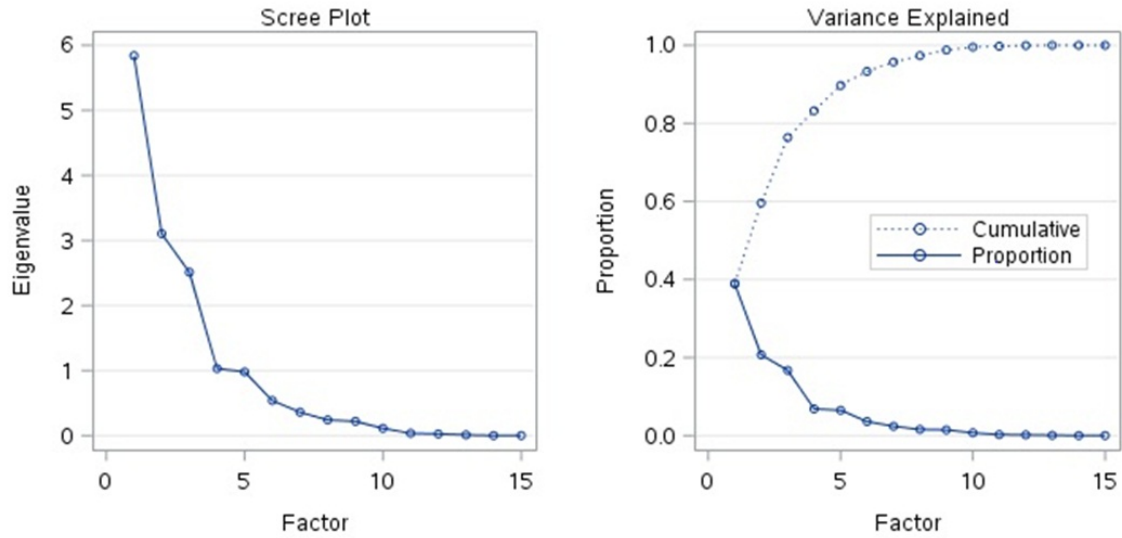


Figure 10.2: Scree plot and proportion of variance for factor analysis

We do not report the results for the t -marginals' factor analysis; these are listed in Appendix C, see Section C.1 for the t_3, t_4, t_5 marginals' factor analysis. The results obtained for the t -marginals are highly comparable to those of the normal marginal. We specifically see similar results for proportion of variance as well - approximately 60% of variance explained by the first two factors. We found closely comparable results for all the t -marginal transformed variables.

Having obtained our two factors for the copula model and using our given data of transformed variables, we now focus on calculating the reduced-dimension dependence generating structures in the next section.

10.3.5 Dependence generators

Overview

Using two factors as extracted from the respective transformed variable sets we now calculate the dependence generators - parameters α_i and β_i as defined in the Gaussian (Equation 9.13) and t -copula (Equation 9.16) two-factor models. We calculate the dependence generators for our four copulas of investigation - Gaussian, t_3, t_4 and t_5 . - using maximum likelihood estimation.

Results

We report the dependence generators in Table 10.2. For each of the four prospective copulas the two parameters are listed. Each α_i and β_i has 15 row-wise elements which correspond to the 15 ORCs variable for which our copulas are constructed.

Even before applying the generators functions to create factor correlation structures it becomes that the parameter estimates are closely related. When doing the pairwise comparison we see that there is an inflationary effect moving from Gaussian

Table 10.2: Dependence generator parameter estimates

i, j	<i>Gaussian copula:</i>		t_3 copula:		t_4 copula:		t_5 copula:	
	α	β	α	β	α	β	α	β
1	-0.130382	0.350327	-0.183169	0.370548	-0.188542	0.388855	-0.185721	0.393700
2	0.0732615	0.220401	0.143053	0.210530	0.126372	0.228930	0.117288	0.234779
3	-0.381259	0.136545	-0.462259	0.246802	-0.441701	0.225686	-0.430275	0.212499
4	-0.424576	-0.042467	-0.465098	-0.045940	-0.451780	-0.037491	-0.445211	-0.033597
5	-0.186353	0.051609	-0.138759	-0.040266	-0.169519	-0.003389	-0.186631	0.015336
6	-0.373445	0.046727	-0.530777	0.119349	-0.518757	0.096065	-0.495815	0.080810
7	-0.224949	0.263095	-0.242264	0.261670	-0.239678	0.270845	-0.234777	0.274676
8	-0.091148	-0.191991	-0.179240	-0.145247	-0.178116	-0.166519	-0.167375	-0.177792
9	-0.054621	0.226921	-0.037243	0.217609	-0.041139	0.224895	-0.044321	0.230694
10	-0.307104	0.242769	-0.509653	0.124051	-0.468147	0.175965	-0.442828	0.203088
11	0.050289	0.138715	0.060903	0.0243294	0.069930	0.042787	0.073032	0.056804
12	0.324125	-0.091745	0.346411	-0.264545	0.372495	-0.239261	0.380566	-0.217436
13	0.345670	0.012051	0.476619	0.023399	0.452489	0.026253	0.433496	0.027099
14	-0.156423	-0.227689	-0.089668	-0.270444	-0.102805	-0.266270	-0.110795	-0.262193
15	0.291360	-0.001183	0.227938	-0.057736	0.268943	-0.073452	0.293451	-0.076463

to t -copulas dependence generators. Comparing within the t -copula estimates only, we mostly see a stepwise inflationary effect; however, the direction - increasing or decreasing - varies across the tabulated vector elements (i, j) .

It is with this result that the power of factor copula technique becomes more apparent. When using the correlation structure formulae - i.e. $(\alpha_i\alpha_j + \beta_i\beta_j)$ for Gaussian copulas and $\frac{\nu}{\nu-2}(\alpha_i\alpha_j + \beta_i\beta_j)$ for t -copulas - in conjunction with the paired (15×1) vectors to generate much more complicated correlation structures with the same high dimensions of the base problem; (15×15) . We illustrate the full correlation structure for the Gaussian copula in Table 10.3.

With the paired sets of dependence generators, we now fit our selected copulas and perform further analytics to gauge the efficacy of the fitting.

10.3.6 Copula fitting

Overview

To assess the goodness-of-fit of our selected copulas we use the empirical tests of Cramér-von Mises along with Anderson-Darling as presented theoretically in Section 8.5. We draw all quantile-quantile plots (qq-plots) for the four copulas we fitted and interpret the relationships observed.

Goodness-of-fit testing - graphical diagnostics

When investigating the qq-plots of the four copulas we fitted in Figure 10.3 we see that there are proper relationships in the fitting for all copulas. We specifically note that the Gaussian and t_3 -copulas falls within the strictest (widest in the figure) confidence interval of $\alpha = 0.01$. In the case of the t_4 -copula we note that the lower tail deviates and falls outside of the shaded ranges. The t_5 -copula shows comparable deviation but to a lesser extent.

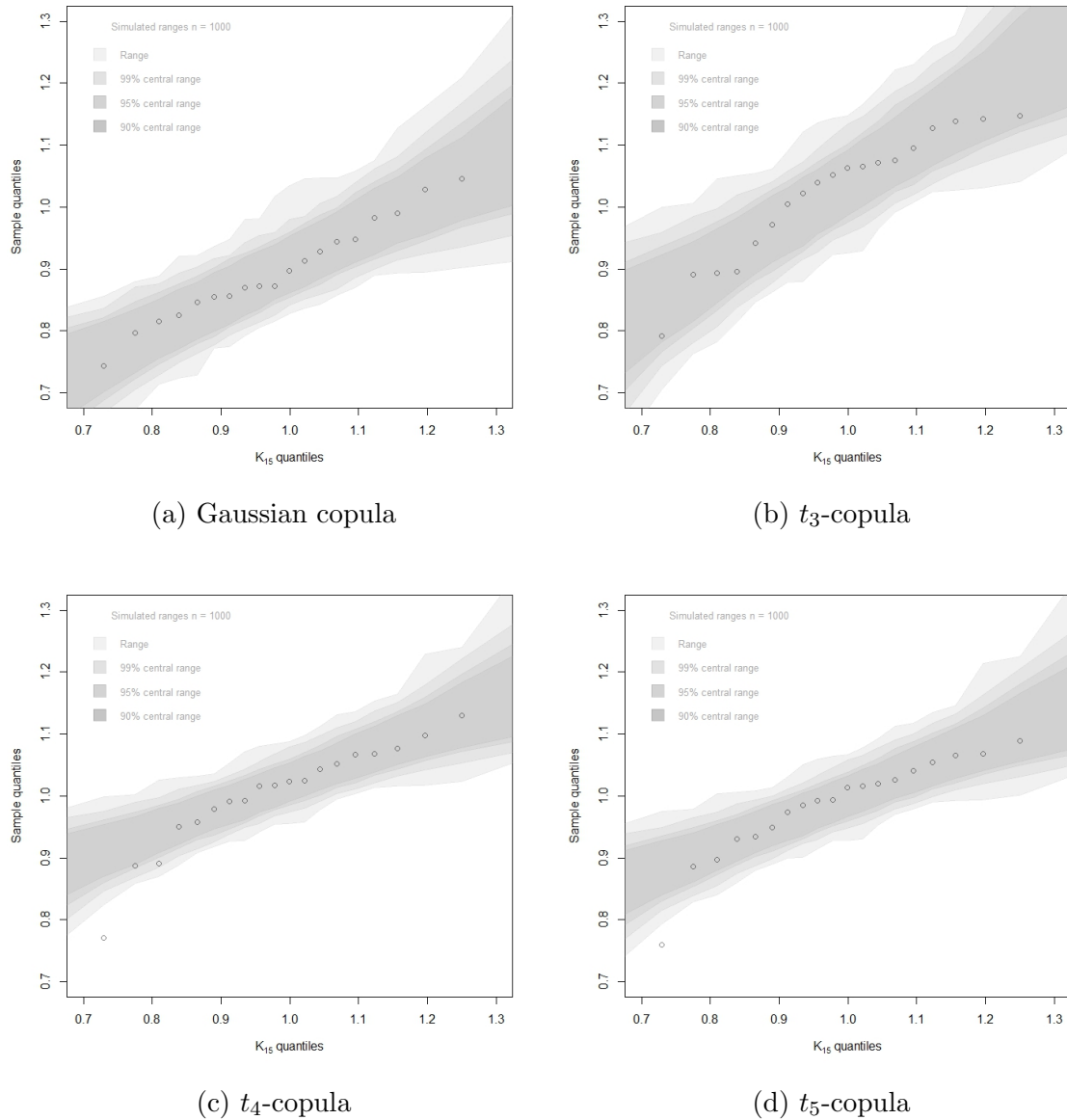
We equalized the graphing scale across all qq-plots to further assess how the fitting between copulas compare. It appears that the Gaussian copula follows the best most direct relationship in its fitting. However, since we know it does not account for heavy tails, we would deem the t_3 -copula to show the best fitting.

A further point of motivation for choosing the t_3 -copula follows from the observation given in Chernobai et al. (2007:202). It is explained that a heavy right tail can easily be gauged from a qq-plot when the plotted points start to dip below an imaginary 45° -line in the upper-right corner of the plot. We can clearly observe this from Figure 10.3b.

We now inspect the empirical tests of the goodness-of-fit assessments.

Goodness-of-fit testing - empirical results

Recalling that the null hypothesis tests whether the fitting follows the specified distribution (or copula in our case) - see Section 8.5 - we conduct the Cramér-von Mises and Anderson-Darling tests on our copula fittings and reject the hypothesis of a good fit when the p -value is smaller than $\alpha = 0.05$. As Chernobai et al. (2007:219) states, a high p -value proposes a good fitting.

Figure 10.3: qq-plots of the fitting of the Gaussian and three t_ν -copulas

Noting that our study is focused on obtaining appropriate heavy tails in our proposed solution we consider the simulated Cramér-von Mises test as an initial indication of the appropriateness of our fitting, but we use the Anderson-Darling test as the final confirmation of the fitting, whilst still giving preferences to the fittings with the lowest test statistic (highest p -value). These results are tabulated in Table 10.4.

We note that the test statistics for the Cramér-von Mises tests are very close. This is expected as Tong et al. (2010:278) mention: the underlying distance calculations in the test formula is much smaller than that of the Anderson-Darling test. Therefore we report S_n to high decimals to illustrate this aspect.

When we revisit the critical values obtained for the Anderson-Darling test (see Section 8.5.3) we confirm the rejection of the Gaussian copula's fit as based on the

Table 10.4: Goodness-of-fit test results

<i>copula</i>	<i>Cramér-von Mises:</i>			<i>Anderson-Darling:</i>		
	S_n	p -value	results	A_n	p -value	results
Gaussian	0.0475802	0.738	do not reject	4.192	0.007	reject
t_3	0.0475965	0.698	do not reject	1.572	0.160	do not reject
t_4	0.0475936	0.629	do not reject	1.803	0.119	do not reject
t_5	0.0475795	0.767	do not reject	1.853	0.111	do not reject

Anderson-Darling test where the test statistic A_n is higher than the critical value. In all other incidences and tests the null hypothesis is not rejected. We therefore to the guidance that low test statistics coupled with high p -values indicate a good fit. This leads us to conclude that the t_3 -copula would be the best fit.

Measures of association

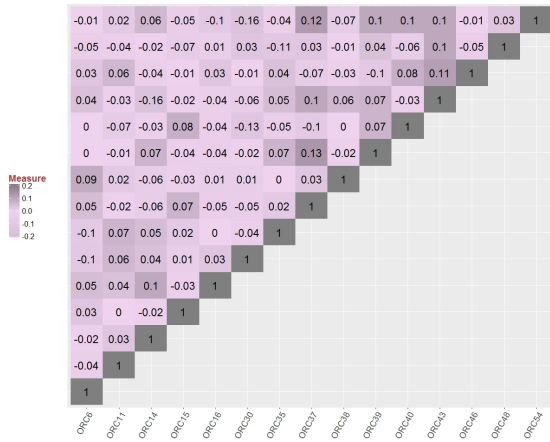
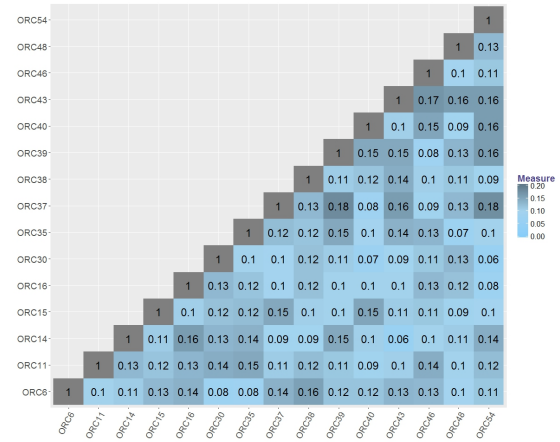
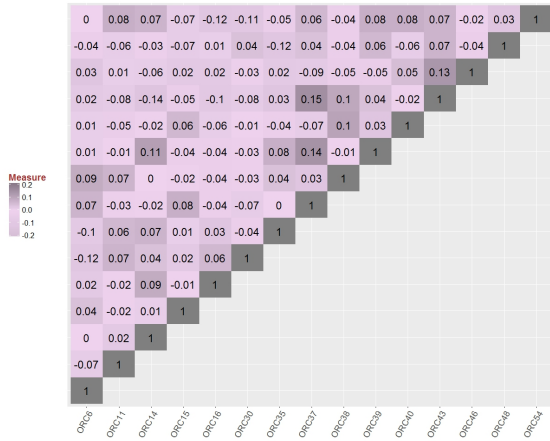
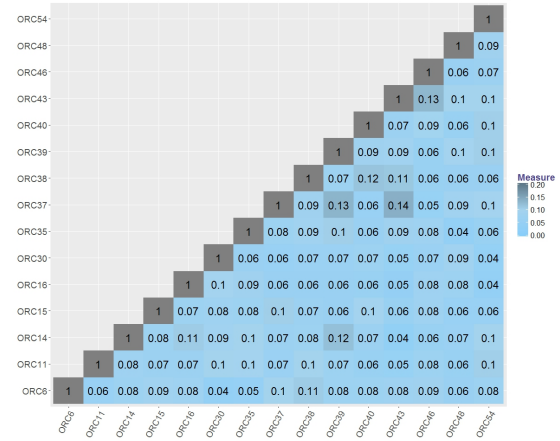
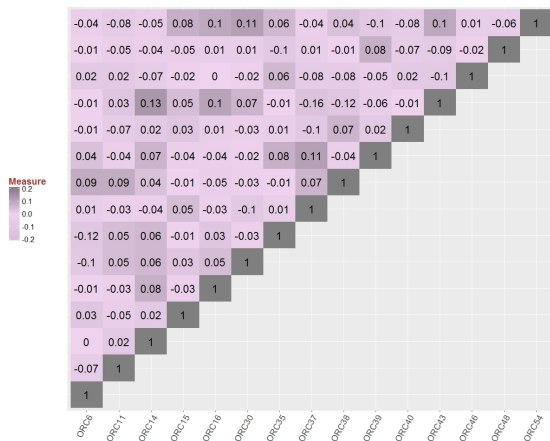
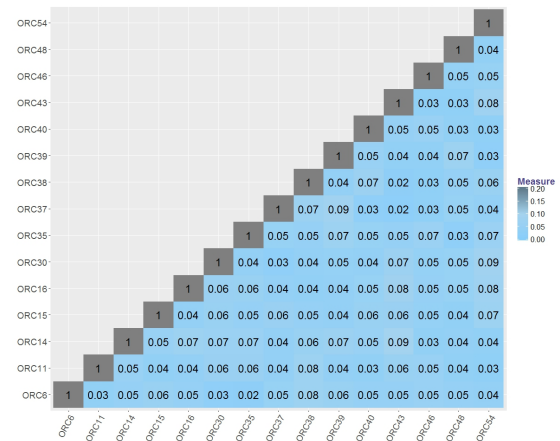
For measures of association we calculated both Kendall's τ as well as the tail indices of our three copulas. We report both measures in Figure 10.4; i.e. Kendall's τ in purple and tail indices in blue. We only illustrate the upper and lower triangles for the respective measures.

The Kendall's τ measures reported in Figures 10.4a, 10.4c, and 10.4e. On average we note that most pairwise measure confirm that the dependence is inflated as degrees of freedom decrease. However, there are some pairwise comparisons where the relationships are not intuitive. We turn to the tail indices for a clearer interpretation of dependence behaviour, specifically that of the tails.

Inherently a Gaussian copula should have full **0** tail index by definition and we therefore only report on tail indices for the t_ν -copulas. This underscores our choice for aiming our proposed solution towards t_ν -copulas which account for tail dependence. Further, note in Figures 10.4b, 10.4d, and 10.4f that we used a cut-off value of a maximum tail index of 0.20 when shading our graphics to better gauge the granular differences visually and retain comparability between the copulas.

In the tail index matrices directly underneath each other, we clearly see the 'inflationary' effect across the t_ν -copulas as the degrees of freedom decrease. Using the consistent colour shading scale, we can see the tail indices decrease as the degrees of freedom increase. The darker blue shading of the t_3 -copula (Figure 10.4b) contrasts sharply with that of the t_5 -copula (Figure 10.4f).

When confirming the actual change in the tail index values between copulas, we can also see the relationship noted by Demarta and McNeil (2005:115). The relationship states that correlations (ρ) and tail indices (ti) are directly related whilst both are indirectly related to the degrees of freedom (ν). This also holds true for our example: $\frac{1}{\nu} \propto \rho \propto ti$. We present a specific example in more detail in Table

(a) t_3 -copula - Kendall's τ (b) t_3 -copula - tail index(c) t_4 -copula - Kendall's τ (d) t_4 -copula - tail index(e) t_5 -copula - Kendall's τ (f) t_5 -copula - tail indexFigure 10.4: Measures of association for the t_ν -copulas

10.5 deriving from the tail indices from Figure 10.4 for ORC37 and ORC39.

Table 10.5: Tail index vs. degrees of freedom

ν	ρ	ti
3	0.0334	0.18
4	0.0244	0.13
5	0.0176	0.09

Based on this result we naturally want to maximize the tail index to obtain conservative estimates for extreme events which show dependence, and therefore would choose the option of the t_3 -copula.

10.4 Graphical analysis of a bivariate case

10.4.1 Overview

In order to construct the bivariate case for graphical diagnostics, and since our correlations on average are not above 0.5, we select the paired cell combination which has the highest correlation for the cell combination across our selected copulas.

This combination is ORC43 $\times$ ORC46. This in turn is qualified as the BL / ET combination of EF.CB and EF.RB. It is unsurprising that external fraud should feature here; it is one of the worst event types the hosting institution has to account for - see Section 5.4.2. Also, intuitively we would expect commercial and retail banking to see some correlation since they are related by definition through both being direct client-facing engagement business lines.

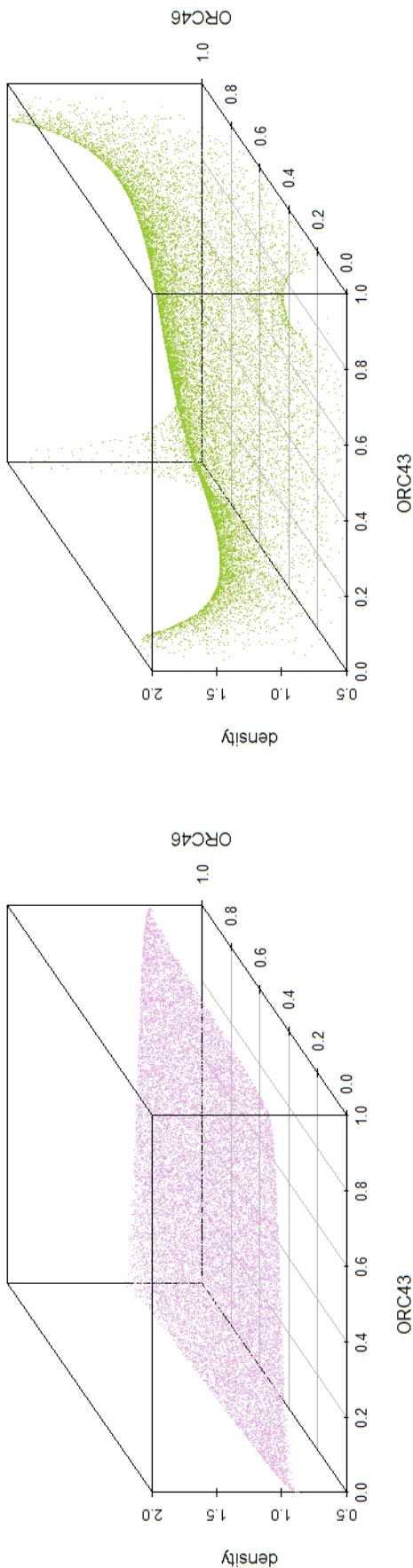
In analyzing our bivariate case graphically, we remain cognisant of the fact that this pairwise selection is merely a subset of the overall datasets. However, we did full and separate 2-dimensional copula fittings for the four different copulas.

We calculated and plotted the copulas' density and distribution functions and plotted these by way of scatter and contour plots. The results for the cumulative function proved to be very granular to notice illustrated differences, and we therefore only report and discuss the copulas' densities.

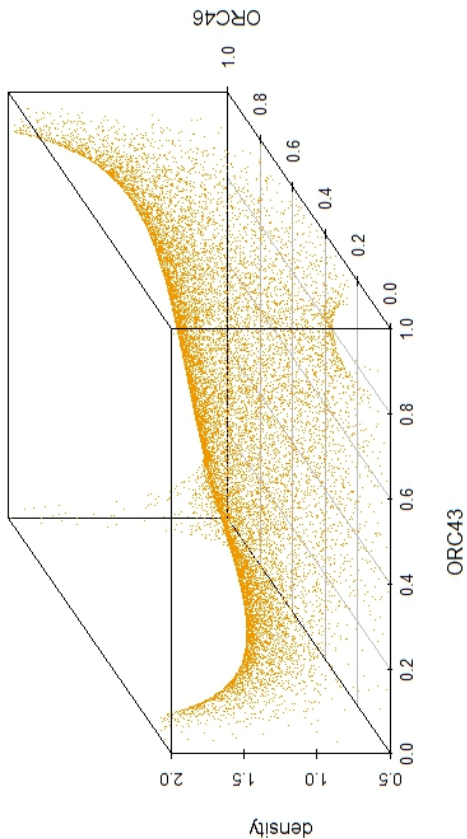
Note that the density graphs are 3-dimensional; i.e. ORC43 $\times$ ORC46 $\times$ *copula density*. Also, we applied a consistent cut-off range on the copula density such that $c(u, v) \in [0.5, 2.0]$ to illustrate distributional differences to a maximum whilst retaining consistency among various copulas.

10.4.2 Graphical diagnostics

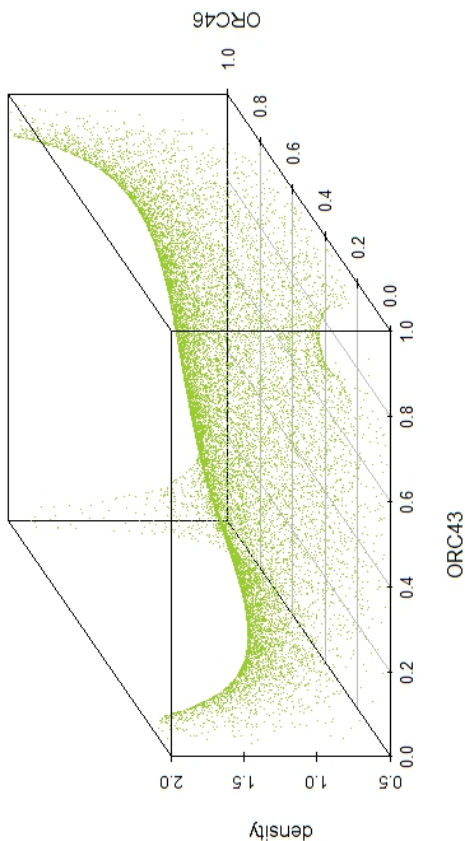
Firstly we wish to point out that the density graphs appear to have 'arms' stretching out in the upper-left and lower-right corners of the bivariate unit square. This is due to the overall low correlations in our dataset. Practical illustrations often use high correlations to illustrate the tail effects only and thus such unexpected *uncorrelated* are not delineated.



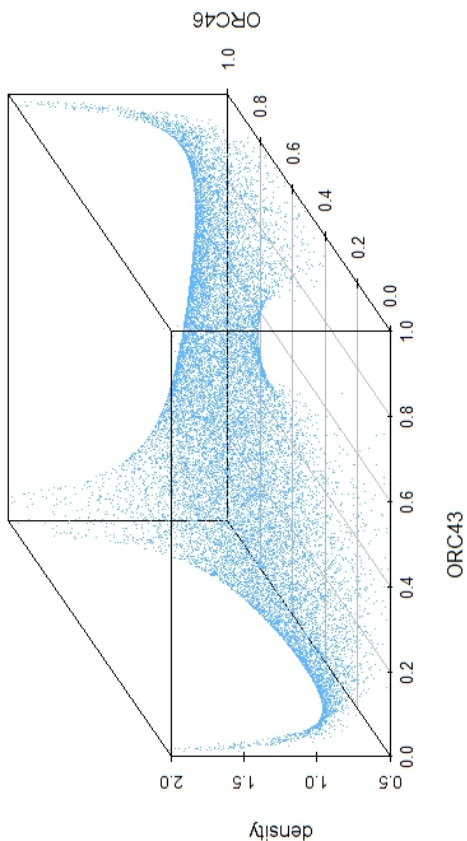
(a) Gaussian copula



(c) t_4 -copula



(b) t_3 -copula



(d) t_5 -copula

Figure 10.5: 3d-scatterplots of the copula densities

When evaluating the graphical results in Figure 10.5 we first consider the Gaussian copula. The graph illustrates that the effect of a varying copula density is extremely subtle. A first drawing initially revealed a sharply defined saddle point rotated at 45° . However, once we incorporated a consistent scale for all copula densities, this subtlety of the Gaussian copula is clearly illustrated. This graph also confirms to us how ineffective a Gaussian copula would be when estimating for tail events.

Moving on to the t_3 - and t_4 -copulas we observe that the difference between them is not so obvious. The density in the central unit square of the bivariate appears to be somewhat steeper. Both show a clear right tail in the upper right-hand corner of their respective graphs. Linking back to our cut-off range on copula density, we can only see the dramatic difference between the two aforementioned copulas when the upper ranges on copula density are allowed to be unrestricted. The maximum density point for the t_3 -copula is around 1,200, whilst the t_4 -copula only reaches around 1,000. Thus confirming a much heavier upper tail for the t_3 -copula.

Another point to consider is the underlying correlation structure of the respective copula fittings. Reporting the correlations as $\rho_{t_3} = 0.416$, $\rho_{t_4} = 0.376$ and $\rho_{t_5} = 0.312$ we already note that said correlations also influence the copula formations. The t_3 - and t_4 -copulas have correlations closer than those of the t_4 - and t_5 -copulas. However, inspecting the t_5 -copula the scale again plays a role.

A maximum copula density of approximately 300 can be seen for unrestricted graphing confirming a much lighter tail than those of the t_3 - or t_4 -copulas. Interestingly, this copula now seems to be inflating the so-called ‘arms of uncorrelation’ as mentioned earlier. Due to the scale of the graph can confirm that it is a highly trivial inflation, but would nevertheless require further motivation should this copula be selected.

When analyzing the contour plots of the bivariate copula densities, most of the information from the density scatter plots is confirmed; however, some additional points are contained in Figure 10.6. Noticeably the futility of seeking tail behaviour information from the Gaussian copula is illustrated. Very granular inspection of the unit square’s four corners reveal the sharp inclines / declines at the respective points.

Again we observe the strong contour inclines for the lower and upper tails, and more so upper, for the t_3 - or t_4 -copulas. Also, the slight elevation in central contour - i.e. density - we suspected from the density scatter plots of the t_3 -copula is confirmed by the contour plot.

From the graph for the t_5 -copula the high density for points noncorrelation. The contouring illustrated in this case appears to be much steeper than those seen for the lower and upper tail. However, upon further analysis with unrestricted ranges on the copula density, we confirm the expected t -copula behaviour of heavier lower and upper tail of dependence.

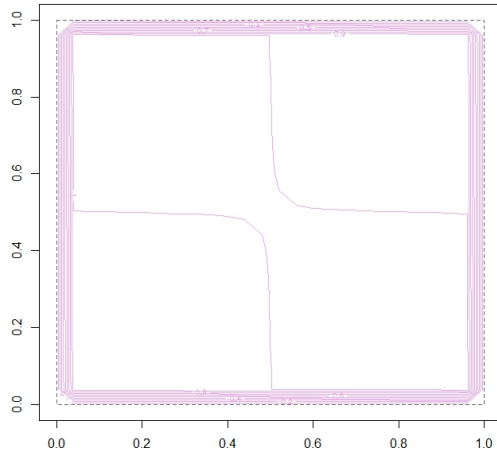
10.5 Impact analysis

10.5.1 Introduction

A part of starting the copula fitting, is to prepare the correlation structures we generated for use in our copulas. We recall the question of negative correlations

as raised in Section 10.2.3. As discussed we substitute all such values with a small value of 0.000001 to counter the deleterious effect of negative correlations. We fit four 15-dimensional copulas - Gaussian and t_3, t_4, t_5 .

Subsequently we calculate the 50th and 99.9th quantiles which represent the expected loss (EL) or provisioning, and unexpected loss (VaR) as derived from capital management. The normal level for regulatory capital is set at the 99.9th whilst the economic capital is preferred at a lower level 99.7th.



(a) Gaussian copula

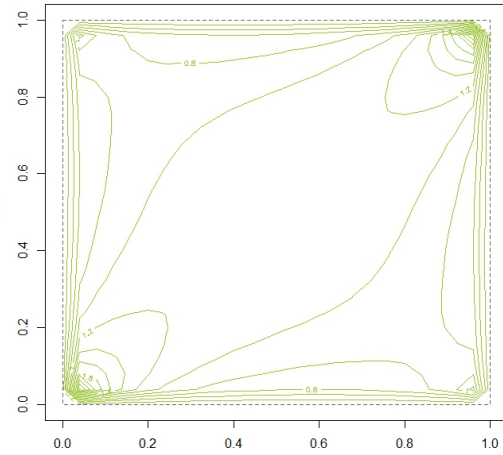
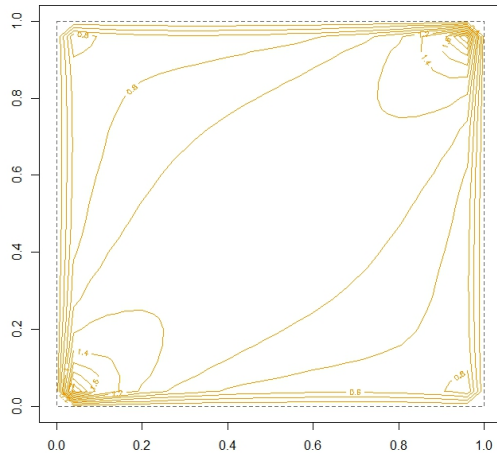
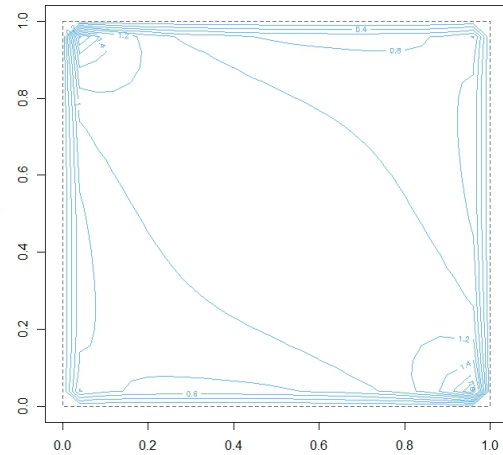
(b) t_3 -copula(c) t_4 -copula(d) t_5 -copula

Figure 10.6: Contour plots of the copula densities

We firstly do the EL and VaR estimation for the undiversified capital charge; i.e. when there is no allowance for dependence between ORCs. We then calculate the EL and VaR estimates for the diversified capital for each of our four copulas. We tabulate the results in Tables 10.6 and 10.7 respectively. For the copula capital

estimates we also calculate the change in provisions as well as the diversification benefit in regulatory capital charge when applying the various copulas.

10.5.2 Capital estimation

Overview

Firstly, the results of undiversified capital charges presented here are not directly comparable with those obtained in Problem I (Table 6.11). This is due to the fact that we used slightly fewer ORCs (i.e. columns) in Problem II due to data scarcity. Secondly, within the cells themselves we only used the periodical observations for which we were able to obtain full observation sets (i.e. rows). Thirdly, for Problem I we assumed parametric loss distribution since alternative options were not relevant to solving the problem of scaling. For Problem II we used empirical distributions since it was very relevant to solving the problem - thus ELs may differ substantially. Nonetheless, when comparing for matched ORCs - especially on the VaR estimates - it is clear that there is still correspondence between the two data sets within the applicable ranges of each ORC.

Table 10.6: Capital and provisioning estimation - undiversified ('000)

ORC	<i>simulated quantiles</i>		
	$EL_{q=.500}$	$VaR_{q=.999}$	Capital
6	21.83	341.28	319.46
11	19.33	1,185.33	1,165.99
14	141.49	1,744.04	1,602.55
15	31.96	166.49	134.53
16	306.59	2,454.87	2,148.27
30	30.73	112.20	81.46
35	143.44	1,132.15	988.71
37	68.17	1,641.37	1,573.20
38	253.63	845.11	591.48
39	11.37	32.48	21.12
40	70.07	1,541.76	1,471.69
43	112.26	1,223.88	1,111.62
46	622.17	6,396.59	5,774.42
48	3.79	35.32	31.53
54	165.63	591.29	425.66
Total	2,002.47	19,444.15	17,441.68

Estimation results

Clearly in terms of EL there seems to be a significant underestimation of provisioning levels. In terms of capital levels the diversification benefit is clear. In all incidences

of the copulas the EL increased by more than 60%. This result also ties in with the graphical diagnostics of Figure 10.5 where we noted the slightly raised central section of the copulas - even increasing when the degrees of freedom do. However, when comparing holistically for EL and VaR between the undiversified results and those of the copula it is already clear that more weight has moved ‘further back’ into the distribution. However, with respect to our explanation on the choice between parametric and empirical distribution; said choice influences EL estimation to a high degree.

Given the EL situation, our proposed solution is concerned with the regulatory capital charge - the VaR estimate. For the copula VaR estimates the values are discernibly lower than those reported in Table 10.6.

Table 10.7: Capital and provisioning estimation - copula diversification (’000)

copula	<i>simulated quantiles</i>		Capital	<i>% change from undiversified</i>	
	$EL_{q=.500}$	$VaR_{q=.999}$		change in provisioning	diversification benefit
Gaussian	3,484.64	10,666.75	7,182.11	74%	45%
t_3	3,263.00	14,526.04	11,263.04	63%	25%
t_4	3,307.65	13,787.10	10,479.45	65%	29%
t_5	3,345.47	13,209.65	9,874.19	67%	32%

When reverting to our goodness-of-fit and measures of association testing which showed that our best choice for a heavy tail copula would be the t_3 -copula, we see that this copula provides the most conservative result for VaR (and the least conservative for EL). Capital estimation from said copula allows for a fairly low diversification benefit of only 25%. When relating our results to those discussed in Chernobai et al. (2007:276), our results fall outside the expected range of diversification benefits of between 30% and 50%. However, our other copulas do approximate this range.

Again when comparing along with the EL estimates such a final decision of a choice of modelling techniques should be made by a group-level operational risk manager. The reasoning at this stage can then centre on the fact that for our copulas’ results the diversification benefits increase when the provisioning estimates increase. A further interesting point we mention here is that the rate of increase of EL appears to be slightly slower than that of the VaR, when moving across t_ν -copulas with increasing degrees of freedom.

However, weighing the two capital management components against each other falls outside the scope of our study.

10.6 Chapter conclusion

In this chapter we reported our data preparation for solving Problem II on dependence structure optimization. We reaffirmed our choice of analyzing *inter-cell* dependence as UoM in an attempt to realize diversification benefit for capital estimation. We explained the omnipresent issue of data scarcity and how we effectively created more data by means of changing our periodicity from years to trimesters. A rough analysis of the data indicated that negative dependencies are present and for a conservative capital estimation approach we decided to substitute said values with small positive dependencies.

We subsequently followed the technique of factor copula modelling according to the guidance from Zhang and Jiao (2012). For our Gaussian copula we perform the same modelling steps strictly, whereas for our t_ν -copulas, where $\nu \in (3, 4, 5)$, we follow our approach as formulated in Section 9.5.2. The two copula types follow the respective variable transformations of standard normal and Student's t; the latter possessing the fixed degrees of freedom as for the chosen copulas.

We performed factor analysis on the transformed variables and observed that for all four transformation we only manage to obtain a level of around 60% for proportion of variance explained by two factors. Even though the level is considered less than ideal, but in comparison to Zhang and Jiao (2012) it represents an appropriate improvement.

Using the extracted factors, we executed the optimization procedure as defined for our factor copulas, maximizing for the parameters of α_i and β_i . From these two vectors we were able to generate our factor correlation structures using the expressions of $(\alpha_i\alpha_j + \beta_i\beta_j)$ and $\frac{\nu}{\nu-2}(\alpha_i\alpha_j + \beta_i\beta_j)$.

These two steps of factor analysis and maximum likelihood estimation form the core of the dimension reduction we used for Problem II. Given that the data is 15-dimensional, through factor analysis and factor copula modelling we were able to construct a 15-dimensional dependence structure for capital estimation using only two factors via the two vector dependence generators.

Having fit our four copulas we reported on their goodness-of-fit for both the Cramér-von Mises and Anderson-Darling tests, as well as measures of association, specifically the copulas' tail indices. Based on the aforementioned testing we concluded that the t_3 -copula produced the best fit for our data.

In order to provide a more visual interpretation of our copula fitting we added a special bivariate case from our dataset and compiled various graphical diagnostics to better illustrate the differences between the copulas. The resulting copula densities assisted in better understanding of said differences whilst still affirming the result of the overhead 15-dimensional copulas.

We closed the chapter by estimating capital charges using our selected copulas. We found that diversification benefits followed expected results from operational risk literature; i.e. within the range between 30% and 50%. Only for the t_3 -copula we saw a slightly lower result of 25%. However, this lower result was accompanied with a significantly increased provisioning level. A final alternative can be sought by opting for the t_4 -copula in capital estimation. We leave a final decision in terms of capital management to an operational risk manager.

Chapter 11

Conclusion on dependence optimization

11.1 Literature study

Given that our proposed solution to Problem II centred on copula techniques, we naturally followed the development of copula theory in our literature study. As explained in Trivedi and Zimmer (2005), we often have easier access to marginal distribution information, and also *more* of said information. Copulas also allow us to perform deeper analysis into specific areas of dependence behaviour; e.g. central tendency, tailedness, and areas of noncorrelation.

We therefore opened our literature study on copula theory with a short review of original starting point - Sklar's theorem. With this theorem we saw how marginal distributions $(F(x), G(y))$ which are often easily obtainable, can be dynamically combined by means of a copula $C(F(x), G(y))$ to construct a joint distribution $H(x, y)$ from which we can then interpolate and inference to analyze joint dependence behaviour of the underlying marginal variable such that $H(x, y) = C(F(x), G(y))$.

Using the *Fisher integral transformation* we noted the technique of transforming our raw variables to uniform-parametric variables which then accommodates operations on the unit square when considering a Cartesian plane¹. Continuing on from this unit square interpretation we provided an illustration of how the probability associated with a copula structure is interpreted; first for a 2-dimensional flat service, then for a 3-dimensional volume. We concluded this overview by providing the theoretical expression for an n -dimensional copula.

With the basics of copula theory explicated, we considered special cases of copulas. These are numerous and still popularly used - we also used the well-known elliptical copulas for our proposed solution.

When thinking analogously of the limits and central point of independence of the linear correlation, ρ , we see similar possible structures for copulas. Working from Nelsen (2006) we illustrated the minimum, product and maximum copulas - the product copula being implying full independence. Other special copulas refer to the two main families of copulas; namely those of the elliptical and Archimedean

¹This transformation technique is also a core step in our proposed solution where transformation of our raw variables implicitly undergo transformation to uniform distribution before the second distribution-selected transformation takes place; i.e. standard normal, standard t .

families. We investigated the definition of said copulas and what possibilities they provide in terms of dependence structure analysis.

With a basic knowledge of how and why to use copulas - as well as which copula would be most suitable - we needed to determine how we can confirm a proper fitting of a selected copula to our data profile. In order to gauge the appropriateness of such a fitting we summarized goodness-of-fit techniques. We consulted the theory of the Cramér-von Mises and Anderson-Darling tests. We observed an important difference; Anderson-Darling in its formulaic construction allows for tail behaviour explicitly. Given that our proposed solution aimed to cross over from a Gaussian copula environment to a more heavy-tailed environment, the latter test is considered most important when determining a final copula selection.

We then reviewed the measures of association; which allowed us to interpret copula output and thus also aided in selecting the most desired dependence behaviour. Specifically we reviewed Kendall's τ and Spearman's ρ , as well as *tail indices*. Kendall's τ and Spearman's ρ provided us with a means to analyze dependence structure arising from copula fittings, as opposed to reverting to linear correlation. Using the aforementioned measure of association we were able to compare differences in the efficacy of dependence behaviour modelling between our various selected copulas. A similar comparison was made possible by calculating tail indices for the various copulas. When determining the tail indices we compared how heavy the tails of the respective copulas are; with the Gaussian copula naturally providing a tail index of $\mathbf{0}$.

Our literature review then focused on an isolated topic; the topic for our proposed solution - factor copulas, where the factors are extracted from an original dataset by applying factor analysis. We first set the scene for investigating factor copulas, by first reviewing the theoretical origin of factor copulas as derived from conditional probability. Through this derivation we learned that the conditioning aspect - as is done with factors - can be achieved from copulas by calculating the partial derivative of the copula in the direction of the conditioning variable. Thus if we are to construct a factor copula we thus need to condition for the factors ($\mathbf{w}$) to be used in the copula $c(u, v|\mathbf{w})$; and in turn obtain the partial derivatives for said factors:

$$c(u, v|\mathbf{w}) = \frac{\partial^2 C(u, v|\mathbf{w})}{\partial u \partial v} . \quad (11.1)$$

Subsequently we considered the one-factor model as designed by Vašíček (2002). Here we saw the principle of a model in a standard normal environment accounting for a *common* risk element and a *specific* risk element; i.e. risks which are common and specific to our observations. We noted the following two alternatives to this model below.

Firstly, building further on this result, Zhang and Jiao (2012) constructed a two-factor model where the common risk element is split into two factors, F_1 and F_2 . Working on the assumptions from Vašíček (2002) that the common and specific risk elements are independent of each other and standard normally distributed, they define a two-factor model to allow for greater dependence complexity to be expressed in the two factors as opposed to a single factor. This model is expressed as:

$$X_i = \alpha_i F_1 + \beta_i F_2 + \sqrt{1 - \alpha_i - \beta_i} Z_i \quad , \quad (11.2)$$

with the correlation structure generator defined as $\text{corr}(x_i, x_j) = (\alpha_i \alpha_j + \beta_i \beta_j)$.

Secondly, Bluhm et al. (2003) provided a definition for such a model in an environment where the underlying marginals follow Student's t -distributions and not standard normal. This solution is obtained easily enough by scaling the standard normal results from Vašíček (2002) with the ratio $\sqrt{\frac{\nu}{W}}$.

11.2 Proposed solution

When setting out our proposed solution to Problem II, we first reflect on the data environment and what data workings we performed prior to copula modelling. As much of the data work had been carried out under Problem I and we use the pooled (scaled) datasets as resulting from our solution for Problem I, we only provided a short overview of the data situation as at the start of solving Problem II.

We firstly confirm that we worked with intercell dependencies; i.e. the dependencies arising after aggregate loss (simulation) modelling has already taken place. This therefore amounts to 56 unique cell definitions corresponding to the various business lines and even type combinations - our ORCs. However, not all ORCs contained sufficient data across the years of investigation.

In deciding between the choice of unit of measures - our ORCs in this case - and the observation definition (summate losses per annum) we had to derive a type of balance, or compromise, in filtering our data. Given our explicit desire of work with the ORCs as unit of measure, we opted to make the periodicity of the loss summation more granular. Upon investigation of alternatives we found that trimester periodicity provided the best (i.e. most) results. When thus settled on a dataset using 15 ORCs' and 21 trimesters' data.

Analysis of the dataset indicated that there are negative correlations present. Given the desired conservatism to be applied when modelling for capital estimation as with our problem, we resolved to substitute such values with very small positive values when performing the final impact study of capital estimation.

We then started with our modelling process. Following from the literature study, we firstly decided to use the two-factor copula, a Gaussian copula, as defined in Zhang and Jiao (2012). This serves as our benchmark model, since we specifically wanted to see what changes occur when we attempted to model for tail behaviour. In order to model said tail behaviour, we turned to t -copulas. However, we needed to construct our own two-factor t -copula model.

We created this model by following the scaling stipulation with the ratio $\sqrt{\frac{\nu}{W}}$. We thus constructed the following model:

$$\tilde{X}_i = \sqrt{\frac{\nu}{W}} \left(\alpha_i F_1 + \beta_i F_2 + \sqrt{1 - \alpha_i - \beta_i} Z_i \right) \quad , \quad (11.3)$$

with all terms defined as in a Gaussian copula, save for W which is denoted as $W \sim \mathcal{X}^2(\nu)$. We also noted that the dependence generator is now given as $\text{corr}(\tilde{x}_i, \tilde{x}_j) = \frac{\nu}{\nu-2}(\alpha_i \alpha_j + \beta_i \beta_j)$. As per reported results for operational risk losses, we restricted our degrees of freedom for the t_ν -copula to be fixed and in the range $\nu \in (3, 4, 5)$ as these provide appropriately heavy tails - and noted that high degrees

of freedom lead to normal approximation which we explicitly wanted to avoid for its lack of tail behaviour.

In initiating our model process, we transformed our variables: for the Gaussian copula we performed from the empirical distribution functions of the marginals (in effect uniform transforms already). For our t -copula we transformed according to a t_ν -distribution adapting for the three chosen degrees of freedom. Subsequently we extract the two factors as defined in our model.

The reason for the choice of *two* factors appeared to be the high proportion of variance explained in Zhang and Jiao (2012). Also, we naturally assumed it as the most obvious extension of the Vašíček (2002) model. In our case, the proportion of variation explained was somewhat lower - 10% lower at around 60%. However, we draw confidence from the fact that our original dataset had significantly higher dimensionality than that of the Zhang and Jiao (2012) study - 15-dimensional in our data and 4-dimensional in Zhang and Jiao (2012). Furthermore, by including additional variables, the proportion of variance explained increased slower than with the first two factors. Especially the increase seen between two and three factors led us to abandon this extra inclusion.

At this point we already had sufficient information to fit copulas. By using the factor scores as calculated from the transformed variables after the two-factor extraction has taken place we fit the four copulas. We then performed goodness-of-fit tests and inspected the results of the measures of association, as well illustrated graphical diagnostics. Overall, the t_3 -copulas provided the best fit results, and as expected also the most ‘inflated’ dependence results.

With the resulting information of the extracted factors and transformed variables, we calculated the dependence generator vectors α_i and β_i by the optimization technique of maximum likelihood estimation, using our factor model expression rewritten in terms of the specific risk element - see equations 9.13 and 9.16. Then, applying our dependence generator function for the Gaussian and t_ν -copulas respectively, we are able to construct correlation-type structures which we incorporate in our impact study for capital estimation through copula simulations.

Our impact study provided us with the best practical interpretation on, firstly, how well the factor copula performed, and secondly, which copula is best suited to our dataset. For this impact study we calculated expected and unexpected loss for the ORCs using full independence - ORCs evaluated in isolation - and for the four copulas. We found that there is clear diversification benefit available when reverting to these factor copulas - up to 45% for the Gaussian. As per our statistical analysis our model of choice was the t_3 -copula where diversification benefit was only 25%. Given results from the industry, a lower limit is normally around 30%, which we closely approximate with our t_4 -copula. We also posed the juxtaposing argument that the copulas led to substantial increase in expected loss - provisioning. And that the rate of increase in expected loss is somewhat slower than the increase in diversification benefit when we increase degrees of freedom.

Such a balance act we recommend for the judgement call of an operational risk manager.

Part III

Conclusion

Chapter 12

Overall summary

12.1 Quantitative response to two operational risk problems

In closing this study we revisit our two problem definitions from Chapter 1 and reflect on whether our proposed solutions adequately solved said problems and contributed to our field of study.

We recall Problem I defined as:

- *How can we scale external loss data in such a way as to allow for subsequent direct combination / pooling of internal and external loss data to model capital estimates which reflect a financial institute's operational risk profile which is also informed by external operational loss tendencies?*

Our proposed solution to this question is contained in our formulaic expression of our quantile regression scaling mechanism. As defined below, we suggest that a modeller can scale down large external losses by using a ratio multiplier. This ratio multiplier consists of two *quantile* regression models where the regression models aim to predict loss values of across diverse quantiles for the ILD and ELD respectively:

$$Loss_{B_{\theta_i}} = Loss_{A_{\theta_i}} \times \frac{\hat{\alpha}_{\theta_i} \cdot \ln(Size_{\theta_i}) + \sum_j \hat{\beta}_{\theta_{ij}} BL_{\theta_{ij}} + \sum_k \hat{\delta}_{\theta_{ik}} ET_{\theta_{ik}}}{\hat{\alpha}_{\theta_i} \cdot \ln(Size_{\theta_i}) + \sum_j \hat{\beta}_{\theta_{ij}} BL_{\theta_{ij}} + \sum_k \hat{\delta}_{\theta_{ik}} ET_{\theta_{ik}} + \sum_l \hat{\xi}_{\theta_{il}} REG_{\theta_{il}}} \quad (12.1)$$

Building on the Regression scaling model by Dahlen and Dionne (2010) (DD-model) where a regression model is constructed for the core operational risk measures of business lines and event types, we firstly added the ‘region’ component as proposed in an earlier version by Dahlen and Dionne (2008). Naturally this step is only possible for the ELD since all our ILD are Africa-based.

We retained the concept of using a size indicator proxy in this extended DD-model as initiated by Shih et al. (2000). Statistical analysis of such a relationship proved dubious for the case of ILD, yet later modelling results were meaningful, and the final scaling result had the desired effect.

Our study then incorporated a different angle by drawing on quantile regression as popularised by Koenker (2005). During graphical diagnostics on our datasets it became clear that most parameter estimates away from the central quantile do not lie close to a conditional model as is implied with OLS. Empirical testing confirmed that lower and especially upper quantiles indeed provide significantly divergent quantile parameter estimates. When such parameter estimation did not prove to be statistically significant, we followed the hybrid approach suggested by Davino et al. (2014) to revert to OLS estimation for those points since a larger underlying dataset is used in such estimation.

Using an iterative regression process we obtained a final results with the hybrid OLS aspect where necessary. Upon analysis of the results produced by said model we found the model-predicted values closely follow those of actual values - both ELD and ILD. Where such trends diverged we suggested an approximating yet conservative solution by testing for maximum values between actual and predicted values when compiling a final scaled dataset.

Our scaled ELD was then pooled together with our existing ILD dataset allowing us to use a dataset informed by international larger-scale operational loss experience, yet firmly applicable to the South African industry. We tested this with an impact study where we found that *undiversified* capital estimation would increase by only 9% when pooled together scaled ELD using our scaling mechanism.

With this quantile regression scaling mechanism we postulate that we have contributed to greater understanding in quantitative risk analysis of data scaling for operational risk. We illustrated that extending the tractable DD-model to include the ‘Region’ aspect of operational loss is especially worthwhile from a South African viewpoint where scaling cannot be derived directly. Also, in using quantile regression we have shown that for the South African environment special care is needed when estimating capital at high levels so as not to underestimate the impact of unexpected loss events.

Using our resulting pooled dataset we turn to our study of Problem II, which we defined as:

- *How can we create an optimal dependence structure which facilitates the high-dimensional dependence analysis associated with the taxonomy of operational risk loss data, whilst still allowing for undemanding diversified capital estimation using our scaled and pooled dataset?*

As with the preceding problem, our proposed solution is easily summarized in the formulaic expression of our two-factor t_ν -copula. We indicate in the expression below how only two factors extracted through factor analysis from a correlation matrix can allow us to obtain *dependence generators* which are able to generate high-dimensional dependence structure for inclusion in copula modelling:

$$Z_i = \frac{\sqrt{\frac{W}{\nu}} \tilde{X}_i - \alpha_i F_1 - \beta_i F_2}{\sqrt{1 - \alpha_i - \beta_i}} \quad . \quad (12.2)$$

Aiming for results from elliptical copulas only, we focused on the method developed by Zhang and Jiao (2012) for working with two-factor copulas. The concept was

already usable from the Vašíček (2002) one-factor model, but the two-factor extension allowed for more dependence information to be extracted from datasets in a more ‘sparing’ manner using only two factors.

We then combined this two-factor Gaussian copula suggested by Zhang and Jiao (2012) and reworked it as a t_ν -copula version deriving from the one-factor version in Bluhm et al. (2003). When appropriately adapting the dependence generators for Student’s t -distribution scaling, we were able to obtain copula dependence structures which could be used for capital estimation.

In testing our fitting for a Gaussian and three t_ν -copulas, we observed graphically and empirically that not only is diversification easily obtainable with this method, but extending for t_ν -copulas also allow us to model dependence structure with varying tail behaviour, as is expected for extreme operational loss events. Goodness-of-fit testing pointed to a t_3 -copula as the best-suited for our data. However, capital estimation had a more complex insight on the matter.

Our impact study took on the the form of such capital estimation, and we contrasted capital estimates for undiversified estimation and diversification through copula simulation. Using a two-factor Gaussian copulas as a benchmark, we additionally modelled t_3 -, t_4 -, and t_5 -copulas. We restricted the degrees of freedom to maximize tail dependence for conservative capital estimation, but also concluding from literature that said range is best-suited for operational risk losses (Hull & White, 2004).

However, overly conservative capital estimation is not necessarily the best option for active risk management. Using a t_3 -copula confirms a conservative approach only leading to a 25% reduction in VaR; yet it comes at the price of a high expected loss - 75% more than for undiversified results. We therefore recommend rather to use a t_4 -copula when modelling for dependence structures.

With this two-factor t_ν -copula we postulate that we contributed to greater understanding in quantitative analysis of optimizing dependence structures for capital diversification in operational risk. Using the simplifying and dimension-reducing technique of factor copulas, we illustrated how to extend such a result to include tail behaviour modelling for dependent extreme loss events. We also contrasted copula behaviour and discussed how optimization of dependence structures also includes reflection on capital estimates and conservative modelling.

12.2 Future perspectives

12.2.1 Problem I

Extensions to quantile regression

When considering extensions to quantile regressions as a standalone concept, we turn to our core theoretical source, Koenker (2005). Several alternative and additional interpretation to using quantile regression are cited; including inter alia quantile crossing, censoring, extremal quantile regression, nonlinear quantile regression, and discrete quantile regression.

We firstly reflect on adjusting for quantile crossing. As Koenker (2005:56) explains, quantiles crossing transgresses the principle of distribution functions and their inverses

being monotone increasing, and adds that this usually only occurs in the extremities of the quantile ranges. This consideration links to two other features of quantile regression; extremal and nonlinear quantile regression. Firstly, Koenker (2005:130) defines the parameter estimation $\hat{\beta}_n(\tau_n)$ as extremal when the quantile approximates the range limits; i.e. $\tau_n \rightarrow 0$ or $\tau_n \rightarrow 1$. Again data paucity poses a problem since such isolated regressions require ever smaller subsets of an already data-challenged environment. Secondly, for nonlinear regression quantile crossing will be removed by inherent definition of the nonlinearity. Koenker (2005:211) indicates that nonlinearity is adapted from the analogous theory for nonlinear least squares.

Concerning censoring in quantile regression the application is obvious. We note this includes the two linked situations of true censoring - i.e. observations were never observed, but are intuitively expected to exist such as with ELD where operational losses are obtainable above a certain threshold - and the case of truncation - i.e. where observations are captured but removed through filtering for whatever reason such as with ILD where we discard modelling of small losses. Censoring quantile regression can aid in model fitting to reflect these so-called ‘missing’ values. We note that this aspect mostly affects the frequency modelling of operational losses.

Finally, we also recommend investigating quantile regression modelling for discrete models (Koenker, 2005:259). Our study only considered the severity aspects of ELD scaling since the ELD was already much smaller than our ILD and a reduction in data points is undesired. However, when focusing on frequency scaling by applying discrete quantile regression in a ratio type format such as our severity-driven model, could assist in resampling with replacement and thereby actually increasing our dataset to an extent. However, we observe the warning of Koenker (2005) on the intrinsic shortcoming of overdispersion in Poisson distribution modelling which may become more pronounced in quantile regression contexts.

Variations to our model

We consider possible variations in our proposed model methodology in this section.

In our analysis of the scaling model we included the ‘Region’ dimension when regressing for ELD. As mentioned Dahlen and Dionne (2010) did not include this in their final model. Our own results proved unexpected as well (see Section 5.4.8); only including the factor level ‘Other Americas’. However, this region is intuitively expected to be the most comparable to South African context and the inclusion of it is therefore sensible. For the Dahlen and Dionne (2010) study their focus was centred on scaling for a bank operating in North America and including ‘Region’ may thus be unnecessary. We recommend monitoring the development of this variable in the future as data availability increases.

Splitting distributions is closely aligned to the concept of operational risk’s most challenging area; low frequency high severity losses. The most common approach follows extreme value theory Chernobai et al. (2007:163). We recommend that such extensions be considered in detail since we expect that including extreme value theory will result in a further increase in capital estimates which will in turn require immediate attention for risk and capital management contexts. Building on this perturbing risk management possibility, we again note a possible deficiency in using extreme value theory. Chernobai et al. (2007:178) points out that there

is noteworthy bias in parameters estimation when samples are small - which we expect for operational losses - and for certain extreme value procedures there are no analytic approaches, but only graphical diagnostics for determining threshold of extreme values to subsequently model. A last concern is that shape parameters stemming from empirical studies are often larger than one, leading to estimates for capital charges which are unreasonably high.

12.2.2 Problem II

Extensions to copula modelling

An aspect we would definitely recommend for future investigation is the usage Archimedean copulas. When we consider members of the Archimedean copula family such as the Frank and Clayton copulas, we see that more complex dependence relationships can be captured in such instances. With the Clayton copula we could account for dispersion in the tail or upper right corner of the unit square; i.e. in a bivariate case one variable increases faster than the other. For a Gumbel copula we could accommodate greater dispersions in the middle to lower region of the unit square.

A further feasible alternative is to examine instances of negative association; i.e. where underlying variable show opposite dependence behaviour. We suggest such possibilities as the skewed t -copula. For such a copula this association can even be restricted to a single quadrant in the unit square; i.e. either the upper left or lower right corner. However, recalling our analyses flagging for such negative associations due to low correlations, it is advisable to thoroughly confirm underlying data to be sufficient and truly negatively associated.

A different approach to address the issue of high dimensionality is that of hierarchical copulas (Cherubini et al., 2012:36) as derived from Archimedean copulas. The dense structure associated with high dimensionality can be repeatedly split and refined in terms of low dimensionality copula structures by means of hierarchical copulas.

A typical structure would be defined as $C(u_1, u_2, u_3, u_4) \equiv \phi_3^{-1}(\phi_3 \circ \phi_1^{-1}(\phi_1(u_1) + \phi_1(u_2)) + \phi_3 \circ \phi_2^{-1}(\phi_2(u_3) + \phi_2(u_4)))$, where the two generator functions ϕ_1 and ϕ_2 represent dependence structure for the paired variables of u_1, u_2 and u_3, u_4 respectively, whilst a further complex generator $\phi_3 \circ \phi_i^{-1}$ is then used to represent dependence structures between the pairs u_1, u_3 and u_2, u_4 .

Certain restrictions are necessary for such models: firstly, the dependence parameters should decrease as the hierarchical levels increase; and secondly, as with cluster analysis we should ideally work towards high *within-groups* dependence and lower *between-groups* dependence. Highly effective solutions for hierarchical copula modelling is given by the Marshall-Olkin copulas.

Variations to our model

A first variation point to our model concerns the variables transformation before factor analysis. In our study transformations to standard normal and Student's t were easily obtained. However, as explained in Zhang and Jiao (2012) other transformation ought to be applied when transforming especially for use in Archimedean copulas.

We also suggest that additional factors be added to the factor model. Given that the ORC dimensionality requires significant reduction and two factors may be considered insufficient when the full set of ORCs are included, such an extension may become unavoidable. A further expansion is of course linked to multivariate analysis on factors; e.g. the different methods available for factor analysis. Note should be taken of the usage of either principal component or exploratory factor analysis. Furthermore, rotations of extracted factors can be amended by applying orthogonal or oblique rotations where appropriate.

An increase in degrees of freedom can naturally be considered for t -copulas. However, we caution against large increases in degrees of freedom; i.e. significantly higher than the so-called ‘ideal’ range for operational risk: $\nu \in (3, 4, 5)$. High degrees of freedom may diminish the desired tail behaviour to too great an extent.

Part IV

Appendices

Appendix A

A.1 Results of ELD regression modelling process

Results of the regression process for the ELD are presented in Tables A.1, A.2, A.3, A.4, and A.5.

Table A.1: Exploratory OLS regression for ELD - part i

	<i>Dependent variable:</i>	
	lCurLoss	
	(1)	(2)
log(CurRev)	0.083*** (0.013)	0.080*** (0.013)
BL.CF		−0.084 (0.289)
BL.TS		−0.198 (0.262)
BL.RB		−1.001*** (0.243)
BL.CB		−0.532** (0.248)
BL.AS		0.254 (0.603)
BL.AM		0.035 (0.277)
BL.RBR		−1.308*** (0.263)
Constant	13.889*** (0.285)	14.638*** (0.372)
Observations	1,924	1,924
R ²	0.021	0.086
Adjusted R ²	0.021	0.082
Residual Std. Error	1.619 (df = 1922)	1.568 (df = 1915)
F Statistic	42.131*** (df = 1; 1922)	22.514*** (df = 8; 1915)

Note:

*p<0.1; **p<0.05; ***p<0.01

Table A.2: Exploratory OLS regression for ELD - part ii

	<i>Dependent variable:</i>	
	lCurLoss	
	(3)	(4)
log(CurRev)	0.057*** (0.013)	0.080*** (0.013)
ET.CPBP	-0.650 (0.396)	
ET.DPA	-1.396** (0.573)	
ET.EPWS	-1.719*** (0.466)	
ET.EDPM	-1.139*** (0.422)	
ET.EF	-1.624*** (0.397)	
ET.IF	-1.403*** (0.397)	
Reg.Africa		0.566* (0.328)
Reg.Asia		0.298 (0.240)
Reg.Europe		0.516** (0.234)
Reg.NAmerica		0.298 (0.228)
Reg.OAmericas		1.595*** (0.473)
Constant	15.644*** (0.485)	13.599*** (0.375)
Observations	1,924	1,924
R ²	0.089	0.031
Adjusted R ²	0.086	0.028
Residual Std. Error	1.565 (df = 1916)	1.614 (df = 1917)
F Statistic	26.856*** (df = 7; 1916)	10.135*** (df = 6; 1917)

Note:

*p<0.1; **p<0.05; ***p<0.01

Table A.3: Full OLS regression for ELD

	<i>Dependent variable:</i>	
	Coefficient	SE
log(CurRev)	0.073***	(0.013)
BL.CF	−0.227	(0.283)
BL.TS	−0.315	(0.254)
BL.RB	−0.696***	(0.239)
BL.CB	−0.079	(0.246)
BL.AS	0.396	(0.585)
BL.RBR	−1.372***	(0.256)
BL.AM	−0.174	(0.269)
ET.CPBP	−0.646*	(0.387)
ET.DPA	−1.597***	(0.557)
ET.EPWS	−1.566***	(0.458)
ET.EDPM	−1.196***	(0.412)
ET.EF	−1.726***	(0.388)
ET.IF	−1.477***	(0.387)
Reg.Africa	0.584*	(0.311)
Reg.Asia	0.342	(0.230)
Reg.Europe	0.380*	(0.223)
Reg.NAmerica	0.323	(0.217)
Reg.OAmericas	1.342***	(0.446)
Constant	15.500***	(0.566)
Observations	1,924	
R ²	0.153	
Adjusted R ²	0.144	
Residual Std. Error	1.514	(df = 1904)
F Statistic	18.065***	(df = 19; 1904)

Note:

*p<0.1; **p<0.05; ***p<0.01

Table A.4: Quantile regression across deciles for ELD

$\theta =$	<i>Dependent variable:</i>									
	lCurLoss									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	OLS
log(CurRev)	0.002 (0.004)	0.013** (0.005)	0.028*** (0.006)	0.044*** (0.009)	0.058*** (0.013)	0.071*** (0.015)	0.071*** (0.017)	0.090*** (0.021)	0.119*** (0.020)	0.068*** (0.012)
BL.RB	-0.187*** (0.032)	-0.358*** (0.063)	-0.511*** (0.061)	-0.707*** (0.075)	-0.825*** (0.085)	-0.781*** (0.096)	-0.843*** (0.111)	-0.791*** (0.135)	-0.571*** (0.144)	-0.567*** (0.074)
BL.RBR	-0.399*** (0.074)	-0.694*** (0.110)	-0.887*** (0.064)	-1.160*** (0.093)	-1.399*** (0.139)	-1.381*** (0.197)	-1.354*** (0.232)	-1.401*** (0.192)	-1.257*** (0.270)	-1.224*** (0.125)
ET.DPA	-0.346*** (0.059)	-0.788*** (0.078)	-0.840 (0.593)	-0.810* (0.422)	-1.101* (0.576)	-1.017 (1.216)	-0.791 (1.645)	-0.136 (0.716)	-1.089*** (0.238)	-0.921** (0.410)
ET.EPWS	-0.139* (0.080)	-0.296*** (0.080)	-0.555*** (0.172)	-0.630*** (0.103)	-0.725*** (0.250)	-1.015*** (0.256)	-1.050*** (0.398)	-1.152** (0.578)	-1.603*** (0.270)	-0.934*** (0.254)
ET.EDPM	-0.044 (0.075)	-0.020 (0.146)	-0.256*** (0.075)	-0.291* (0.155)	-0.273 (0.193)	-0.617*** (0.202)	-0.581** (0.239)	-1.045*** (0.267)	-1.421*** (0.283)	-0.538*** (0.165)
ET.EF	-0.221*** (0.060)	-0.491*** (0.070)	-0.648*** (0.073)	-0.750*** (0.085)	-0.859*** (0.110)	-1.101*** (0.126)	-1.135*** (0.157)	-1.502*** (0.181)	-2.022*** (0.210)	-1.037*** (0.089)
ET.IF	-0.192*** (0.064)	-0.412*** (0.073)	-0.532*** (0.077)	-0.612*** (0.088)	-0.714*** (0.117)	-0.854*** (0.137)	-0.986*** (0.147)	-1.301*** (0.178)	-1.545*** (0.250)	-0.796*** (0.091)
Reg.OAmericas	0.866** (0.366)	1.039*** (0.128)	1.269*** (0.443)	1.170* (0.631)	1.324*** (0.379)	0.941*** (0.305)	0.809*** (0.192)	0.049 (0.273)	0.972* (0.577)	0.985** (0.394)
Constant	14.312*** (0.110)	14.633*** (0.141)	14.768*** (0.160)	14.893*** (0.219)	15.061*** (0.300)	15.270*** (0.353)	15.835*** (0.407)	16.282*** (0.506)	16.736*** (0.467)	15.138*** (0.286)

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.5: Quantile regression across high quantiles for ELD

$\theta =$	<i>Dependent variable:</i>									
	lCurLoss									
	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
log(CurRev)	0.119*** (0.020)	0.122*** (0.022)	0.137*** (0.021)	0.142*** (0.021)	0.142*** (0.019)	0.132*** (0.026)	0.190*** (0.035)	0.151*** (0.040)	0.187*** (0.027)	0.331*** (0.054)
BL.RB	-0.571*** (0.144)	-0.580*** (0.151)	-0.602*** (0.131)	-0.541*** (0.117)	-0.565*** (0.150)	-0.552*** (0.174)	-0.456** (0.227)	-0.613** (0.289)	-0.654*** (0.223)	-0.758* (0.393)
BL.RBR	-1.257*** (0.270)	-1.350*** (0.239)	-1.332*** (0.261)	-1.283*** (0.294)	-1.338*** (0.299)	-1.249*** (0.282)	-1.237*** (0.181)	-1.642*** (0.400)	-1.629*** (0.418)	-1.734*** (0.446)
ET.DPA	-1.089*** (0.238)	-1.211*** (0.212)	-1.338*** (0.202)	-1.460*** (0.184)	-1.582*** (0.346)	-1.716*** (0.248)	-1.691*** (0.325)	-2.492*** (0.358)	-2.913** (1.334)	-3.103** (1.549)
ET.EPWS	-1.603*** (0.270)	-1.644*** (0.310)	-1.843*** (0.198)	-1.963*** (0.369)	-2.086*** (0.318)	-1.919*** (0.382)	-2.267* (1.193)	-2.423** (1.039)	-2.462*** (0.488)	-3.428*** (0.436)
ET.EDPM	-1.421*** (0.283)	-1.461*** (0.282)	-1.483*** (0.231)	-1.711*** (0.290)	-1.778*** (0.268)	-1.639*** (0.389)	-1.756** (0.823)	-1.401** (0.689)	-1.932*** (0.294)	-2.410*** (0.535)
ET.EF	-2.022*** (0.210)	-2.074*** (0.203)	-2.101*** (0.195)	-2.277*** (0.183)	-2.273*** (0.187)	-2.300*** (0.215)	-2.283*** (0.260)	-2.682*** (0.355)	-2.892*** (0.301)	-2.510*** (0.520)
ET.IF	-1.545*** (0.250)	-1.517*** (0.245)	-1.455*** (0.239)	-1.604*** (0.231)	-1.611*** (0.244)	-1.517*** (0.266)	-1.371*** (0.378)	-1.542*** (0.385)	-1.690*** (0.315)	-1.415*** (0.488)
Reg.OAmericas	0.972* (0.577)	0.818 (2.039)	0.613 (2.170)	0.549 (2.219)	1.896 (2.414)	1.772 (2.453)	1.571*** (0.312)	1.014*** (0.306)	0.553 (1.123)	0.144 (0.526)
Constant	16.736*** (0.467)	16.787*** (0.520)	16.593*** (0.491)	16.688*** (0.499)	16.808*** (0.430)	17.138*** (0.588)	16.014*** (0.830)	17.543*** (0.970)	17.285*** (0.632)	14.738*** (1.124)

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix B

B.1 Proof of deriving Kendall's τ from copula function

The proof is directly aligned to Nelsen (2006:159).

The stochastic variables were defined as being continuous, $P[(X_1 - X_2)(Y_1 - Y_2) < 0] = 1 - P[(X_1 - X_2)(Y_1 - Y_2) > 0]$; it therefore follows that

$$P[X_1 > X_2, Y_1 > Y_2] = P[X_2 < X_1, Y_2 < Y_1] \quad (\text{B.1})$$

$$= \iint_{\mathbf{R}^2} P[X_2 \leq x, Y_2 \leq y] dC_1(F(x), G(y)) \quad (\text{B.2})$$

$$= \iint_{\mathbf{R}^2} C_2(F(x), G(y)) dC_1(F(x), G(y)) . \quad (\text{B.3})$$

We then use the probability transforms and substitute $u = F(x)$ and $v = G(y)$ into our equation; thus obtaining

$$P[X_1 > X_2, Y_1 > Y_2] = \iint_{\mathbf{I}^2} C_2(u, v) dC_1(u, v) . \quad (\text{B.4})$$

We then consider the analogous example

$$P[X_1 < X_2, Y_1 < Y_2] \quad (\text{B.5})$$

$$= \iint_{\mathbf{R}^2} P[X_2 > X_1, Y_2 > Y_1] dC_1(F(x), G(y)) \quad (\text{B.6})$$

$$= \iint_{\mathbf{R}^2} [1 - F(x) - G(y) + C_2(F(x), G(y))] dC_1(F(x), G(y)) \quad (\text{B.7})$$

$$= \iint_{\mathbf{I}^2} [1 - u - v + C_2(F(x), G(y))] dC_1(F(x), G(y)) . \quad (\text{B.8})$$

However, given that C_1 is the joint distribution function of our stochastic univariate variable pair (U, V) on $[0, 1]$, we thus expect $E(U) = E(V) = 1/2$. It then follows

that

$$P[X_2 < X_1, Y_2 < Y_1] \quad (\text{B.9})$$

$$= 1 - \frac{1}{2} - \frac{1}{2} + \iint_{\mathbf{I}^2} C_2(u, v) dC_1(u, v) \quad (\text{B.10})$$

$$= \iint_{\mathbf{I}^2} C_2(u, v) dC_1(u, v) . \quad (\text{B.11})$$

We can then conclude that

$$P[(X_1 - X_2)(Y_1 - Y_2) > 0] = 2 \iint_{\mathbf{I}^2} C_2(u, v) dC_1(u, v) . \quad (\text{B.12})$$

B.2 Proof that tail indices are nonparametric and only depend on the underlying variables' copula(s)

We refer the reader back to Section 8.6.2 for the theoretical introduction to this proof:

$$\lambda_U = \lim_{t \rightarrow 1^-} P[Y > G^{(-1)}(t) | X > F^{(-1)}(t)] \quad (\text{B.13})$$

$$= \lim_{t \rightarrow 1^-} P[G(Y) > t | F(X) > t] \quad (\text{B.14})$$

$$= \lim_{t \rightarrow 1^-} \frac{\overline{C}(t, t)}{1 - t} \quad (\text{B.15})$$

$$= \lim_{t \rightarrow 1^-} \frac{1 - 2t + C(t, t)}{1 - t} \quad (\text{B.16})$$

$$= 2 - \lim_{t \rightarrow 1^-} \frac{1 - C(t, t)}{1 - t} \quad (\text{B.17})$$

$$= 2 - \delta'_C(1^-) . \quad (\text{B.18})$$

B.3 Proof that $\frac{\partial}{\partial u}C(u, v)$ and $\frac{\partial}{\partial v}C(u, v)$ exist for almost all u and v

We follow the proof directly as provided in Darsow et al. (1992:610) which is provided in terms of Markov processes. Note the analogy in the notation where we can substitute for our instance that $u \equiv x$ and $C(u, v) \equiv A_{,1}(t, y)B_{,2}(x, t)$.

The lemma states: let A and B be copulas. For almost all x it then follows that:

$$\frac{\partial}{\partial x} \int_0^1 A_{,1}(t, y) B_{,2}(x, t) dt = \int_0^1 A_{,1}(t, y) B_{,2}(x, dt) . \quad (\text{B.19})$$

The proof states: let ϕ be defined as any copula function C^∞ which terminates at

points 0 and 1. We then see that:

$$\int_0^1 \phi(x) \left(\int_0^1 A_{,1}(t, y) B_{,1}(x, dt) \right) dx = - \int_0^1 \phi'(x) \left(\int_0^x \int_0^1 A_{,1}(t, y) B_{,1}(\xi, dt) d\xi \right) dx \quad (\text{B.20})$$

$$= - \int_0^1 \phi'(x) \left(\int_0^1 A_{,1}(t, y) B(x, dt) \right) dx \quad (\text{B.21})$$

$$= - \int_0^1 \phi'(x) \left(\int_0^1 A_{,1}(t, y) B_{,2}(x, t) dt \right) dx \quad (\text{B.22})$$

$$= \int_0^1 \phi(x) \frac{\partial}{\partial x} \left(\int_0^1 A_{,1}(t, y) B_{,2}(x, t) dt \right) dx , \quad (\text{B.23})$$

which proves the lemma.

Darsow et al. (1992) indicates that the second step requires further justification. We observe the mapping expression as:

$$f \rightarrow \int_0^x \left(\int_0^1 f(t) B_{,1}(u, dt) \right) du \quad \text{and} \quad f \rightarrow \int_0^1 f(t) B(x, dt) , \quad (\text{B.24})$$

with both described as positive linear functionals which are bounded. These mappings are defined on all right-continuous step functions on the interval $[0, 1]$. Said functions are stated as equal for all f . Given that right-continuous step functions are dense on (the Lebesgue measure) $L^1([0, 1])$, it follows that the linear functionals *extend uniquely* to linear functionals that are defined as equal for all functions where $f \in L^1([0, 1])$.

Appendix C

C.1 Results of factor analysis

Results of the factor analysis process are presented in Tables C.1, C.2, and C.3.

Table C.1: Proportion of variance - t_3 -marginals transformation

Factors	Eigenvalue	Difference	Proportion	Cumulative
1	5.83423	2.71165	0.389	0.389
2	3.12257	0.53338	0.208	0.597
3	2.58919	1.57281	0.173	0.770
4	1.01637	0.07859	0.068	0.838
5	0.93777	0.41655	0.063	0.900
6	0.52122	0.18152	0.035	0.935
7	0.33970	0.10182	0.023	0.957
8	0.23787	0.02230	0.016	0.973

Table C.2: Proportion of variance - t_4 -marginals transformation

Factors	Eigenvalue	Difference	Proportion	Cumulative
1	5.84138	2.72313	0.389	0.389
2	3.11824	0.55094	0.208	0.597
3	2.56730	1.55887	0.171	0.769
4	1.00842	0.05333	0.067	0.836
5	0.95508	0.42833	0.064	0.899
6	0.52674	0.18332	0.035	0.935
7	0.34341	0.10481	0.023	0.957
8	0.23860	0.02367	0.016	0.973

Table C.3: Proportion of variance - t_5 -marginals transformation

Factors	Eigenvalue	Difference	Proportion	Cumulative
1	5.84333	2.72796	0.390	0.390
2	3.11537	0.56069	0.208	0.597
3	2.55468	1.54811	0.170	0.768
4	1.00657	0.04219	0.067	0.835
5	0.96437	0.43477	0.064	0.899
6	0.52959	0.18340	0.035	0.934
7	0.34618	0.10731	0.023	0.957
8	0.23887	0.02354	0.016	0.973

C.2 Factor dependence structures

The factor dependence structures are presented in Tables C.4, C.5, and C.6. Recall that the diagonal as derived from the dependence generators will (almost) never be 1.00, and we therefore only report the upper triangle to avoid confusion.

References

- Adebambo, B., Brockman, P. & Yan, X. 2015. Anticipating the 2007-2008 financial crisis: who knew what and when did they know it?. *Journal of Financial and Quantitative Analysis*, 50(4):24.
- Afambo, E. F. 2006. Operational risk capital provisions for banks and insurance companies. Atlanta: Georgia State University (Thesis - PhD).
- Andersen, L. & Sidenius, J. 2004. Extensions to the Gaussian copula: random recovery and random factor loadings. *Journal of Credit Risk*, 1(1):42.
- Basel Committee on Banking Supervision 2006. International convergence of capital measurement and capital standards. Basel, CH. (Standard).
- Basel Committee on Banking Supervision 2011. Operational risk – supervisory guidelines for the advanced measurement approach. Basel, CH. (Standard).
- Basel Committee on Banking Supervision 2016. Standardised measurement approach for operational risk. Basel, CH. (Standard).
- Bluhm, C., Overbeck, L. & Wagner, C. 2003. An introduction to credit risk modelling. Chapman & Hall/CRC. Boca Raton, FL.
- Bouchaud, J.-P. 2001. Power laws in economics and finance: some ideas from physics. *Quantitative Finance*, 1(1):8.
- Brunel, V. 2014. New results on the correlation problem in operational risk. *The Journal of Financial Perspectives*, 2(2):22.
- Chapelle, A., Crama, Y., Hübner, G. & Peters, J.-P. 2004. Basel II and operational risk: Implications for risk measurement and management in the financial sector. Report. National Bank of Belgium.
- Chernobai, A. S., Rachev, S. T. & Fabozzi, F. J. 2007. Operational Risk: A Guide to Basel II Capital Requirements, Models and Analysis. John Wiley & Sons Inc. Hoboken, NJ.
- Cherubini, U., Gobbi, F., Mulinacci, S. & Romagnoli, S. 2012. Dynamic copula methods in finance. John Wiley & Sons Ltd. London, UK.
- Clauset, A., Shalizi, C. R. & Newman, M. 2009. Power-law distributions in empirical data. *SIAM Review*, 51(4):43.

- Cope, E. & Labbi, A. 2008. Operational loss scaling by exposure indicators: evidence from the ORX database. *Journal of Operational Risk*, 3(4):21.
- Dahen, H. & Dionne, G. 2008. Scaling models for the severity and frequency of external operational loss data. In Dahen, H. & Dionne, G., eds. *Performance Measurement in the Financial Services Sector: Frontier Efficiency Methodologies and Other Innovative Techniques*. Tanaka Business School, Imperial College.
- Dahen, H. & Dionne, G. 2010. Scaling models for the severity and frequency of external operational loss data. *Journal of Banking & Finance*, 34:13.
- Darsow, W. F., Nguyen, B. & Olsen, E. T. 1992. Copulas and Markov processes. *Illinois Journal of Mathematics*, 36(4):43.
- Davino, C., Furno, M. & Vistocco, D. 2014. *Quantile Regression: Theory and Applications*. John Wiley & Sons, Ltd. West Sussex, UK.
- Demarta, S. & McNeil, A. J. 2005. The t -copula and related copulas. *International Statistical Review*, 73(1):20.
- Embrechts, P. & Hofert, M. 2011. Practices and issues in operational risk modelling under Basel II. *Lithuanian Mathematical Journal*, 51(02):14.
- Fletcher, A. J. & Marchildon, G. P. 2014. Using the delphi method for qualitative, participatory action research in health leadership. *International Journal of Qualitative Methods*, 13:18.
- Frachot, A. & Roncalli, T. 2002. Mixing internal and external data for managing operational risk.
URL: http://www.maths-fi.com/malliavin/Mixing_internal_and_external_data_for-managing_operational_risk_01_29_2002.pdf [Date of access: 02 October 2013]
 Groupe de Recherche Opérationnelle, Crédit Lyonnais.
- Galloppo, G. & Prevati, D. 2014. A review of methods for combining internal and external data. *Journal of Operational Risk*, 9(4):22.
- Genest, C., Rémillard, B. & Beaudoin, D. 2009. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44:15.
- Giacometti, R., Rachev, S. T., Chernobai, A. S. & Bertocchi, M. 2008. Aggregation issues in operational risk. *Journal of Operational Risk*, 3(3):20.
- Groenewald, A. 2014. Practical methods of modelling operational risk. In Groenewald, A., ed. *Actuarial Society of South Africa's 2014 Convention*. ACM.
- Hull, J. & White, A. 2004. Valuation of a CDO and an n -th to default CDS without Monte Carlo simulation. *Journal of Derivatives*, 12(2):40.
- Joe, H. 2015. *Dependence modeling with copulas*. CRC Press. Boca Raton, FL.
- Kabacoff, R. I. 2011. *R in Action*. Manning Publications Co. New York, NY.

- Kühn, R. & Neu, P. 2003. Functional correlation approach to operational risk in banking organizations. *Physica A: Statistical Mechanics and its Applications*, 322:650–666.
- Koenker, R. 2005. Quantile Regression. Cambridge University Press. New York, NY.
- Koenker, R. 2017. quantreg: Quantile Regression.
URL: <https://CRAN.R-project.org/package=quantreg> [Date of first access: 18 April 2014] R package version 5.34.
- Koenker, R. & Machado, J. A. 1999. Goodness-of-fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):16.
- Marsaglia, G. & Marsaglia, J. C. W. 2004. Evaluating the Anderson-Darling distribution. *Journal of Statistical Software*, 9(2):5.
- Mazo, G. 2014. Construction et estimation de copules en grande dimension.
URL: <https://tel.archives-ouvertes.fr/tel-01130963v2/document> [Date of access: 11 December 2015] Grenoble: Université de Grenoble (Thesis - PhD).
- Na, H. S. 2004. Analysing and Scaling Operational Risk Loss Data. Rotterdam: Erasmus University (Thesis - Master's).
- Na, H. S., Miranda, L., Van den Berg, J. & Leipoldt, M. 2006. Data scaling for operational risk modelling.
URL: <https://repub.eur.nl/pub/7234> [Date of access: 21 July 2015] ERIM report series.
- Nelsen, R. B. 2006. An introduction to copulas. Springer. New York, NY.
- Neugebauer, M. 2007. A comprehensive analysis of advanced pricing models for collateralised debt obligations. Oxford: University of Oxford (Thesis - Master's).
- Osborn, T. & Haritonova, A. 2017. New Basel delay throws SMA into doubt.
URL: <https://www.risk.net/risk-management/operational-risk/2410697/scaling-operational-loss-data-and-its-systemic-risk-implications> [Date of access: 17 October 2017] *Risk.net*.
- Palaro, H. P. & Hotta, L. K. 2006. Using conditional copula to estimate value-at-risk. *Journal of Data Science*, 4:23.
- Panjer, H. H. 2006. Operational Risk: Modeling Analytics. Wiley Series in Probability and Statistics. John Wiley & Sons. New Jersey, NJ.
- Powell, J. L. 1984. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25:23.
- Schmidt, T. 2006. Coping with copulas. In Rank, J., ed. Copulas: From theory to application in finance. 1st ed. Risk Books. electronic book section 1, p. 302.

- Shih, J., Samad-Khan, A. & Medapa, P. 2000. Is the size of an operational loss related to firm size?.
URL: [http://www.opriskadvisory.com/docs/Is_the_Size_of_an_Operational_Loss_Related_to_Firm_Size_\(Jan_00\).pdf](http://www.opriskadvisory.com/docs/Is_the_Size_of_an_Operational_Loss_Related_to_Firm_Size_(Jan_00).pdf) [Date of access: 07 August 2013] OpRisk Advisory - Stamford Risk Analytics.
- Tong, L., Yang, J. & Cooper, R. S. 2010. Efficient calculation of p -value and power for quadratic form statistics in multilocus association testing. *Annals of human genetics*, 74(3):10.
- Torresetti, R. & Nordio, C. 2014. Scaling Operational Loss Data and Its Systemic Risk Implications.
URL: <https://www.risk.net/risk-management/operational-risk/2410697/scaling-operational-loss-data-and-its-systemic-risk-implications> [Date of access: 24 May 2015] *Risk.net*.
- Trivedi, P. K. & Zimmer, D. M. 2005. Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1(1):111.
- Vašíček, O. A. 2002. Loan portfolio value. *Risk.net*, (December 2002):3.
- Venter, G. G. 2002. Tails of copulas. *Proceedings CAS*, LXXXIX:46.
- Wang, D., Rachev, S. T. & Fabozzi, F. J. 2009. Pricing tranches of a CDO and a CDS index: recent advances and future research. Physica-Verlag Heidelberg. p. 289.
- Woodall, L. 2017. Revised SMA could allow banks to ignore past op risk losses.
URL: <https://www.risk.net/risk-management/5292146/revised-sma-could-allow-banks-to-ignore-past-op-risk-losses> [Date of access: 17 October 2017] *Risk.net*.
- Zhang, H. & Jiao, F. 2012. Factor copula models and their application in studying the dependence of the exchange rate returns. *International Business Research*, 5(2).