



Royal Institute of Technology
Department of Mathematics

Master Thesis, Financial Mathematics

**Operational Risk Modeling: Addressing
the Reporting Threshold Problem**

Authors:

Oscar Hallberg

Mattias Wärmlöf Helmrich

Supervisor:

Boualem Djehiche

June 3, 2015

Abstract

External loss data are typically left truncated at a reporting threshold. Ignoring this truncation level leads to biased capital charge estimations. This thesis addresses the challenges of recreating the truncated part of the distribution. By predicting the continuation of a probability density function, the unobserved body of an external operational risk loss distribution is estimated. The prediction is based on internally collected losses and the tail of the external loss distribution. Using a semiparametric approach to generate sets of internal losses and applying the Best Linear Unbiased Predictor, results in an enriched external dataset that shares resemblance with the internal dataset. By avoiding any parametrical assumptions, this study proposes a new and unique way to address the reporting threshold problem. Financial institutions will benefit from these findings as it permits the use of the semiparametric approach developed by Bolancé et al. (2012) and thereby eliminates the well known difficulty with determining the breaking point beyond which the tail domain is defined when using the Loss Distribution Approach. The main conclusion from this thesis is that predicting the continuation of a function using the Best Linear Unbiased Predictor can be successfully applied in an operational risk setting. This thesis has predicted the continuation of a probability density function, resulting in a full external loss distribution.

Keywords: *Operational Risk, Advanced Measurement Approach, Prediction, Semiparametric Modeling, Reporting Threshold.*

Acknowledgements

We would like to thank our supervisor at *KTH Mathematical Statistics*, Boualem Djehiche, for his feedback and guidance during the writing of this thesis.

Stockholm, June 3, 2015

Oscar Hallberg and Mattias Wärlöf Helmrich

Contents

1. Introduction	1
2. Regulations	3
2.1 International Convergence of Capital Measurement and Capital Standards ..	3
2.2 Methods for Quantifying Operational Risk	3
2.2.1 Basic Indicator Approach	3
2.2.2 Standardized Approach	4
2.2.3 Advanced Measurement Approach	4
3. Literature Review	7
4. Theory	11
4.1 Penalized B-splines	11
4.2 Principal Component Analysis	13
4.3 The Generalized Champernowne Distribution	14
4.4 Kernel Density Estimation	17
4.4.1 Boundary Correction	19
5. Methodology	21
5.1 Time Variant Factor Scaling	21
5.2 Semiparametric Density Estimation	21
5.3 Best Linear Unbiased Predictor	26
6. Data	27
6.1 Internal Data	27
6.1.1 Risk Cell 1	28
6.1.2 Risk Cell 2	28
6.1.3 Risk Cell 3	29
6.2 Simulated Data	29
6.3 External Data	30
6.3.1 Risk Cell 1	31
6.3.2 Risk Cell 2	31
6.3.3 Risk Cell 3	32

7. Results	33
7.1 Results from the Best Linear Unbiased Prediction.....	33
7.2 Accuracy of the Prediction	35
7.3 Comparing Risk Cells	37
8. Conclusions & Further Research	41
8.1 Further Research	42
9. Bibliography	43
10. Appendix	47
10.1 Detailed Figures.....	47

List of Figures

Figure 4.1 Thirteen basis functions defining an order four spline with nine interior knots, shown as vertical dashed lines (Ramsay & Silverman, 2005, p. 50)	12
Figure 4.2 Illustrative examples of the effect of parameter c . The cumulative distribution function and the probability density function of the Generalized Champernowne Distribution are plotted for different values of α and c , while $M=5$ in all plots. $c=0$ in the solid line and $c=2$ in the dotted line.	16
Figure 4.3 Illustrative example of kernel density estimation with boundary correction. The solid line is raw kernel density estimate and the dotted line is with boundary correction.	19
Figure 5.1 A probability density function of the Generalized Champernowne distribution, with the parameter found via Maximum Likelihood.	22
Figure 5.2 Histogram over a loss distribution, transformed using the cumulative distribution function of the Generalized Champernowne distribution with parameters estimated from Maximum Likelihood.	23
Figure 5.3 Kernel density estimation of a transformed dataset.	24
Figure 5.4 Resulting distribution from the semiparametric estimation of a dataset	25
Figure 6.1 Histogram over the internal dataset of operational risk losses used in this thesis. The last bar includes losses above 2μ	27
Figure 6.2 Histogram over the internal dataset of operational risk losses for risk cell 1. The last bar includes losses above 2μ	28
Figure 6.3 Histogram over the internal dataset of operational risk losses for risk cell 2. The last bar includes losses above 2μ	28
Figure 6.4 Histogram over the internal dataset of operational risk losses for risk cell 3. The last bar includes losses above 2μ	29
Figure 6.5 Histogram over the external dataset of operational risk losses. The last bar includes losses above 2μ	30
Figure 6.6 Histogram over the external dataset of operational risk losses for risk cell 1. The last bar includes losses above 2μ	31
Figure 6.7 Histogram over the external dataset of operational risk losses for risk cell 2. The last bar includes losses above 2μ	31
Figure 6.8 Histogram over the external dataset of operational risk losses for risk cell 3. The last bar includes losses above 2μ	32
Figure 7.1 The result from the prediction of the continuation of the loss distribution in risk cell 1. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.	34

Figure 7.2 The result from the prediction of the continuation of the loss distribution in risk cell 2. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.	34
Figure 7.3 The result from the prediction of the continuation of the loss distribution in risk cell 3. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.	35
Figure 7.4 The result from the prediction of the continuation of the loss distribution in risk cell 1 including confidence bands. The solid blue line is the observed part of the external dataset, the dashed red line is the forecasted continuation of the density function, and the solid green lines are the confidence bands at a 95% level.	36
Figure 7.5 The result from the prediction of the continuation of the loss distribution in risk cell 2 including confidence bands. The solid blue line is the observed part of the external dataset, the dashed red line is the forecasted continuation of the density function, and the solid green lines are the confidence bands at a 95% level.	36
Figure 7.6 The result from the prediction of the continuation of the loss distribution in risk cell 3 including confidence bands. The solid blue line is the observed part of the external dataset, the dashed red line is the forecasted continuation of the density function, and the solid green lines are the confidence bands at a 95% level.	37
Figure 7.7 The result from the prediction of the continuation of the loss distribution in risk cell 1. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.	38
Figure 7.8 The result from the prediction of the continuation of the loss distribution in risk cell 2. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.	38
Figure 7.9 The result from the prediction of the continuation of the loss distribution in risk cell 3. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.	39
Figure 10.1 A more detailed version of figure 7.4.....	47
Figure 10.2 A more detailed version of figure 7.5.....	48
Figure 10.3 A more detailed version of figure 7.6.....	48

List of Tables

Table 2.1 The eight different business lines defined by BCBS and corresponding fixed percentage multipliers used in the Standardized Approach for quantification of Operational Risk.	4
Table 2.2 The seven event type categories defined by BCBS used in the Advanced Measurement Approach for quantification of Operational Risk.	5
Table 4.1 Three commonly used kernel functions.	18

Chapter 1

Introduction

Under the current regulatory framework for the financial industry, referred to as Basel II (Basel Committee on Banking Supervision, 2006), banks are required to hold sufficient capital for their operational risks. The framework stipulates three different approaches for quantification of the capital charge. Out of the three is the Advanced Measurement Approach the most sophisticated and is expected to be closely related to the actual risk profile of the bank. Financial institutions intending to use the Advanced Measurement Approach will need to demonstrate the accuracy of their internal model where the use of historical losses is one of the factors. Historical losses must be collected internally and supplemented by external loss data for a more comprehensive view of the risk profile, since severe losses are rare or absent in the internal dataset. Modeling operational risk losses has traditionally been implemented using the Loss Distribution Approach, where practitioners model the frequency and severity distribution separately using parametric distributions (Aue & Kalkbrener, 2006). Furthermore, it is suitable to use one distribution for the body and another distribution for the tail when modeling the severity given the particular characteristics of operational risk losses. The Loss Distribution Approach leads to difficulties in defining the breaking point between the body and the tail domain, which is addressed by Bolancé et al. (2012) who develop a semiparametric approach, enabling the use of a single distribution over the entire loss dataset. However, fitting the semiparametric model to external losses is troublesome as external losses typically are reported above a threshold where it is only possible to observe the tail of the distribution, rendering it impossible to employ the method proposed by Bolancé et al (2012).

Recreating a left truncated dataset has not previously been addressed without dividing the distribution into a body and a tail domain with separate parametrical assumptions. By applying a procedure for predicting the continuation of a function on a truncated probability density function this thesis will recreate the unobserved body of an external loss distribution. Basing the prediction on the observed part of the density function and internally collected losses, makes this study unique by taking the characteristics of internal losses into account when recreating a truncated external loss distribution. Introducing a new approach for addressing the reporting threshold problem makes this research progressive.

This area within operational risk management is central since a truncated external loss distribution results in biased capital charge estimations. The result of this thesis is beneficial for financial organizations as it enriches the external database and simultaneously permits the use of the semiparametric approach developed by Bolancé et al. (2012) and thereby eliminates the well known difficulty with determining the breaking point beyond which the tail domain is defined. Moreover, this thesis generates an external dataset with desirable characteristics, without parametrical speculations. The purpose of this thesis is to recreate the unobserved body of an external loss distribution based on information from its tail and an internal database, by implementing a method for predicting the continuation of a function developed by Goldberg et al. (2014). The method will be applied on an external database of operational risk losses with a known reporting threshold.

Although operational risk is far from a new concept, it has been given more attention in recent years following the occurrence of large losses as the bankruptcy of Baring Bank due to internal fraud (Stevenson, 1995) and rouge trading at Société Générale in 2008 (Bittermann, 2008). Contrary to credit and market risks, which can be exploited to generate profit, operational risk is merely subject to risk minimization. Operational risks include every aspect of the business that can go wrong with both internal and external causes and as the drivers of operational risks are undefined, the risk modeling is cumbersome. The internal risks include everything from employees committing crimes to IT-system failures, while external risks ranges from bank robberies to earthquakes, which makes operational risk a widespread phenomenon. Events similar to the one in Baring Bank have resulted in an increase in effort to identify and measure operational risks.

Operational risk losses can generally be categorized into two groups: (1) events with high frequency and low severity, and (2) events with low frequency and high severity. Due to the reporting threshold in external databases, the low severity events are not present in the external datasets while the high severity events are rare in the internal loss data. The main reasons for a reporting threshold are cost reduction, the ability to hide insignificantly small losses, and incomplete ways for recording losses (Pirouz & Salahi, 2013). There are two categories of external databases: consortium data and public data. Consortium data are based on collaboration among financial institutions which commit to share their operational risk losses with each other under confidentiality. Public data record publicly released losses consisting of events too large or important to be concealed away from public eyes (Baud, et al., 2002a). In consortium databases, a reporting threshold is enforced and participants need to demonstrate their ability to report every loss above this level, while public databases do not specify a threshold but for natural reasons do not include small losses. In both of these cases, the external database will be left truncated.

Chapter 2

Regulations

2.1 International Convergence of Capital Measurement and Capital Standards

In 2006 the Basel Commission of Banking Supervision (BCBS) published guidelines for capital requirements, the Basel II Capital Accord (Basel Committee on Banking Supervision, 2006). The guidelines have since then been implemented in all member nations and in financial institutions operating on an international level. BCBS states in the first pillar of the accord that financial institutions should hold adequate capital for all different risk types. Operational risk is defined by BCBS as:

The risk of direct or indirect loss resulting from inadequate or failed internal processes, people and systems or from external events.

This definition includes a broad spectrum of events ranging from earthquakes to internal fraud and is therefore a hard subject to quantify and model.

2.2 Methods for Quantifying Operational Risk

The Basel II accord provides financial institutions with three different options to quantify their operational risks: the Basic Indicator Approach, the Standardized Approach, and the Advanced Measurement Approach (Basel Committee on Banking Supervision, 2006). The Basic Indicator Approach and the Standardized Approach are based on the use of gross income as a proxy for the financial institutions' operational risk exposure while the Advanced Measurement Approach allows financial institutions to construct their own model regulated by the national financial supervisory authority.

2.2.1 Basic Indicator Approach

The Basic Indicator Approach allows financial institutions to hold capital equal to a fixed percentage of the average non-negative gross income over the last three years. The required capital according to the Basic Indicator Approach is calculated as:

$$C_{BIA} = \frac{1}{n} \sum_{i=1}^n (GI_i \times \alpha), \quad (2.1)$$

where GI_i is the gross income for year i , α is the fixed percentage which is set by BCBS to 15% , and n is the number of years during the past three year period where the gross income is positive. Any years for which gross income is negative or zero should be excluded from both the numerator and denominator.

2.2.2 Standardized Approach

The Standardized Approach is closely related to the Basic Indicator Approach. Financial institutions following this approach are allowed to divide their gross income into eight business lines defined by BCBS. Within each business line the gross income is multiplied by a fixed percentage multiplier set by the committee to reflect the risk exposure corresponding to the business area. The required capital is calculated as:

$$C_{SA} = \left\{ \sum_{i=1}^3 \max \left[\sum_{j=1}^8 (GI_{i,j} * \beta_j), 0 \right] \right\} / 3, \quad (2.2)$$

where $GI_{i,j}$ is the yearly gross income in business line j during year i and β_j is the corresponding fixed percentage multiplier for business line j , which are shown in Table 2.1 below.

Business Lines	Beta Factors
Corporate Finance (β_1)	18%
Trading and Sales (β_2)	18%
Retail Banking (β_3)	12%
Commercial Banking (β_4)	15%
Payment and Settlement (β_5)	18%
Agency Services (β_6)	15%
Asset Management (β_7)	12%
Retail Brokerage (β_8)	12%

Table 2.1 The eight different business lines defined by BCBS and corresponding fixed percentage multipliers used in the Standardized Approach for quantification of Operational Risk.

2.2.3 Advanced Measurement Approach

The Advanced Measurement Approach is the most sophisticated method for calculating the operational risk capital requirements for a financial institution. Under the Advanced Measurement Approach the required capital is quantified using a risk measure generated by the bank itself complying with the quantitative and qualitative criteria set by the BCBS.

A financial institution may apply to its financial supervisory authority to use an Advanced Measurement Approach given that they fulfill the requirements of the accord. The model specified by the financial institution must include credible, systematic, transparent, and verifiable approaches for weighting internal operational risk loss data, external risk loss data, scenario analysis, and factors reflecting the business environment and internal control systems.

Internal risk loss data are operational risk losses that the financial institution has recorded internally. External operational risk loss data are losses that other banks have encountered. These external losses are available either through public databases or through consortium data. Operational risk losses obtained through consortium databases are often considered more complete and inclusive. Due to the scarcity of internal risk loss data, the committee requires financial institutions to supplement their own internally collected data with external sources.

The quantitative requirements of the Advanced Measurement Approach are that method must ensure a capital allocation that holds for the Value-at-Risk at level 0.01 over a one year period. In addition to the business lines used in the aforementioned Standardized Approach, the Advanced Measurement Approach requires financial institutions to divide their operational risks into seven different event types, listed in table 2.2 below.

Event Type
Internal Fraud
Employment Practices and Workplace Safety
Execution, Delivery, and Process Management
Damage to Physical Assets
External Fraud
Clients, Products, and Business Practices
Business Disruption and System Failure

Table 2.2 The seven event type categories defined by BCBS used in the Advanced Measurement Approach for quantification of Operational Risk.

A correlation matrix is formed following the categorization of business line and event type and financial institutions employing the Advanced Measurement Approach are allowed to model correlation between the different intersections where all correlations assumptions must be described in detail and approved by the financial supervisory authority. The reason behind this categorization is that the loss distributions tend to differ substantially between different intersections, referred to as risk cells, of business line and event type. It is therefore not viable to analyze the aggregated loss distribution stemming from all cells. A detailed description of the categorization can be found in Basel II Capital Accord (Basel Committee on Banking Supervision, 2006).

Chapter 3

Literature Review

One common method for quantifying operational risk capital is using the actuarial approach called the Loss Distribution Approach where the financial institution estimates the distribution of the loss events for the next year for each intersection of business line and event type. The computation is usually performed by simulating the loss frequency and the loss severity separately before finding the compound distribution using Monte Carlo simulation (Aue & Kalkbrener, 2006). The frequency distribution is commonly found using a Poisson process, while the severity distribution is estimated using a parametric distribution such as Log-Normal or Weibull for the body, while Extreme Value Theory is used to capture tail events. The capital requirements are calculated as the 99.9th quantile of the compound distribution. However, separating the severity distribution at the breaking point between the body and the tail may result in flawed estimations due to the uncertainty in determining where the two domains are defined. Other practical issues regarding this method were highlighted in an early stadium by Frachot et al. (2001) who state that common difficulties are: missing data for some combinations of business line and event type, internal data being biased towards low-severity losses, and external losses being biased towards high-severity losses following recording thresholds. These issues are however not addressed in their paper.

Bolancé, Guillén, Gustafsson, and Nielsen (2012) propose a new approach for modeling the severity distribution when introducing a semiparametric approach, enabling the use of a single distribution over the entire loss domain. The semiparametric approach uses the Generalized Champernowne distribution to transform the loss data to the $(0, 1)$ interval, followed by kernel density estimation and back transformation to obtain the severity distribution. Bolancé et al. (2012) eliminate the problems associated with choosing the breaking point between the body and the tail domain under the loss distribution approach, and the method has proven to estimate operational loss distribution data well. Furthermore, the semiparametric approach results in a probability density that is able to produce estimates of tail probabilities beyond the range of the sample data. However, the proposed method requires a full loss distribution which is not available when

modeling external loss data, which typically is left truncated at a reporting threshold.

A method for addressing the data truncation problem was developed by Baud et al. (2002a) who attempt to pool internal and external loss data after considering the threshold level. Discussing different assumptions regarding the threshold, the authors conclude that the reporting threshold is known or unknown, as well as constant or stochastic. Under the unknown threshold assumption, the authors discard the naïve approximation of the threshold level as the smallest recorded loss as external data may contain badly recorded losses and therefore underestimate the true threshold. Instead the authors propose to determine the threshold level H by estimating the set of parameters θ for a fitted truncated Log-Normal distribution over different $H \in [0, \infty)$ using Maximum Likelihood. The authors plot H as a function of θ and choose the threshold level where the parameter set θ stabilizes. Furthermore, the article presents an alternative approach by including the threshold level H as a parameter in the Log-Likelihood function that is to be maximized when fitting the parametric distribution to loss data.

In addition, the authors comment that the method proposed is based on the assumption that external data follow the same distribution as internal data except that the external data are left truncated at a threshold H . Furthermore, they state that, even though not investigated in this paper, their methodology is able to provide a statistical test of the equality of the two distributions. According to the authors, it can serve as a reliable indicator of whether internal losses of a specific bank are comparable with losses from other banks.

Building on these principles Baud et al. (2002b) discuss different methods for incorporating external loss data to the internal dataset by addressing the fact that the external data have gone through a truncation process. The authors use three simulated datasets, truncated at different threshold levels to develop an approach and assess the accuracy of their principles. The simulated datasets are supposed to resemble data encountered in real life, where two of the datasets have a single threshold level, representing internal and consortium loss data, and the third dataset is drawn from distributions truncated at three different levels, representing a public loss database. The authors merge the three datasets and use three different Maximum Likelihood approaches to fit a truncated Log-Normal distribution to the resulting dataset. The results indicate that ignoring any truncation level or assuming a common level for all datasets overestimate the capital requirement. Merging the loss data, and fitting a truncated Log-Normal distribution under the assumption that the number of thresholds is unknown, yields consistent results. The most accurate result is obtained when the threshold levels are known in two of the datasets, and is only estimated in the third, but this method is demanding in terms of prior information.

As an alternative to the fitting of a truncated distribution many authors, including Dutta & Perry (2007), use a simplified approach referred to as the Shifted Approach. In this model the dataset is shifted to the left by subtracting the threshold level to each data point before fitting a parametric distribution and estimating its parameters. The resulting fitted distribution is then shifted back before calculation of the capital charge is performed. This approach has been shown by Lou et al. (2007) to provide flawed estimations when examining the results from different threshold levels. The Shifted Approach results in both underestimations as well as overestimations depending on the proportion of data being truncated. These results agree with the critique presented by Chernobai et al. (2004) on ignoring data truncated below the threshold level when quantifying operational risks and spurred a debate among US regulators on the validity of the Shifted Approach. The debate resulted in a recommendation in 2009 to the financial institutes in the United States to employ a truncated distribution instead of the Shifted Approach (Davis, 2011).

Estimating parameters for a truncated distribution is much more complicated than for a complete distribution and while the Maximum Likelihood approach gained popularity among researchers and practitioners, some authors claim that certain drawbacks with the method exist as the likelihood surface often is flat and the global maximum may be impossible to find. Zhou et al. (2013) explain two approaches to circumnavigate the problem, the Expectation-Maximization algorithm and Penalized Likelihood estimate but advocates for a third approach: the Bayesian Method. In the Bayesian Method, both data and parameters are considered stochastic (Shevchenko & Temnov, 2009). A parameter can be considered a random variable and the true value of the parameter is a realization of this random variable. The density of the parameter is called a prior density and the selection of a proper prior distribution is an important aspect when employing the Bayesian Method. Zhou et al. (2013) assess the accuracy of the Bayesian Method by simulating a sample of size 100 from a Log-Normal distribution truncated at 20 000 and estimate the parameters of the truncated distribution for each sample size from 2 to 100, drawn as a subsample from the original sample. The authors conclude that the Bayesian Method is less sensitive to sample size than Maximum Likelihood estimation, and produces more stable estimates, although the Bayesian estimates converges to the ML-estimates as sample size increases. The authors apply the different methods on empirical risk loss data and find that the Maximum Likelihood estimates are inconsistent for small sample sizes as they sometimes yield negative location parameters.

In recent research authors have tried to address the issues of Maximum Likelihood estimates. Ergashev et al. (2014) acknowledge the problems with a flat likelihood surface and the difficulties in finding a global maximum. The authors present the regularity condition which is a specific, necessary, and sufficient condition for the existence of a global solution to the severity parameter

estimation problem. They show that a violation of the regularity condition is the main reason behind unstable parameter estimates under the Maximum Likelihood approach. In their article the authors use the Method of Moments estimation technique to derive their parameter estimates and show analytically that the method yields the same results as Maximum Likelihood estimation given the existence of a global solution. They compare the capital bias of the truncated approach under the regularity condition with the shifting approach and conclude that the truncated approach induces a capital bias that converges to zero as sample size increases.

The data truncation problem has primarily been addressed with the use of parametric distributions, involving assumptions that may not be valid or desirable, yet research on non-parametric approaches is limited. Goldberg et al. (2014) develop a model to forecast the continuation of a function using functional data techniques. The authors construct the Best Linear Unbiased Predictor and apply their method on data from a call center at a bank, and forecast the arrival and workload process. The prediction model builds on the fact that the curves are governed by a small number of factors, which are found through Principal Component Analysis. The curves are represented using B-spline basis functions and the procedure involves computation of the mean function on the observed segment and the segment the authors wish to predict. Calculation of the covariance operators between the mean functions is then performed. These functions are expressed in terms of a B-spline basis and the corresponding coefficients. Prediction is obtained through computing the representation of the coefficients on the second part of the curve. The authors yield satisfactory result and claim that forecasting the continuation of a function can be achieved successfully by the proposed Best Linear Unbiased Predictor.

Chapter 4

Theory

4.1 Penalized B-splines

Many datasets consist of a number of multidimensional observations that reflect the underlying smooth curve that we assume generates them. In such cases it may be preferable to treat the data as functional rather than as multiple series of data points. With functional data it is possible to analyze smooth curves or surfaces that vary over a continuum. The use of functional data analysis is proven useful in noise reduction, producing robust estimates, and dealing with missing values (Ramsay & Silverman, 2005). In order to reconstruct the smooth characteristics of the data that may have been lost in the observation phase, a powerful tool is penalized B-splines. Spline curves were initially used as a drafting tool for aircrafts and shipbuilding industries, where a flexible strip of material was clamped or weighted to pass through a number of points with smooth deformation. Penalized B-splines are used to convert discrete measurements to a function with values computable for any desired argument.

When using B-splines, one represents a smooth function as a linear combination of basis functions defined on subsets of the whole domain. A basis function system is a set of known functions represented via a recursive formula derived by De Boor (1978). The formula for B-spline basis functions is:

$$N_i^k(u) = \frac{u - u_i}{u_{i+k-1} - u_i} N_i^{k-1}(u) + \frac{u_{i+k} - u}{u_{i+k} - u_{i+1}} N_{i+1}^{k-1}(u), \quad (4.1)$$

where

$N_i^1(u) = \begin{cases} 1, & u_i \leq u < u_{i+1} \\ 0, & \text{otherwise} \end{cases}$, N_i^k is the i th B-spline basis function of order k , u_i is a non-decreasing set of real numbers, also called the knot sequence, and u is the parameter variable.

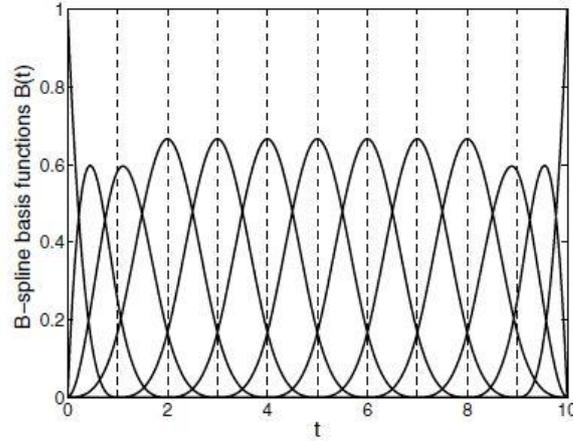


Figure 4.1 Thirteen basis functions defining an order four spline with nine interior knots, shown as vertical dashed lines (Ramsay & Silverman, 2005, p. 50)

Equation 4.1 shows that the B-spline basis function can be expressed by a linear combination of basis functions of a lower degree. One apparent defining feature of the basis functions is the knot sequence u_i . Over each interval where a basis function is defined, the basis function is a polynomial of order k . The different polynomials join smoothly at the breakpoints that separate them, so the function values are constrained to be equal at the breakpoints. Furthermore, the derivatives of order $k - 2$ are also constrained to match. In addition to the breakpoints, the knot sequence affects the characteristics of the basis functions. The knot sequence is a series of non-decreasing real numbers defined on the domain of the basis functions, and the knots will be located at the breaking points. It is possible to define a knot sequence with multiple knots at some break point. In this scenario one loses some continuity restrictions and it is possible to engineer abrupt changes in the derivative or function value of the spline. A basis function of degree k is defined over $k + 1$ knots, or k knot intervals, and since the basis functions are based on the knot difference, the functions are dependent on the knot spacing and not the knot values. A basis function of degree k will consist of a polynomial segment of order $k - 1$ (Ramsay & Silverman, 2005). The sum of the B-spline basis function values at any point is equal to 1. There may also be knots positioned outside the domain of the spline function. These knots are irrelevant for the definition of the spline; however they affect the basis functions (Höllig & Hörner, 2013). The spline curve $s(u)$ is represented as a linear expansion of the basis functions:

$$s(u) = \sum_{i=0}^n d_i \times N_i^k(u), \quad n \geq k - 1 \quad (4.2)$$

where $s(u)$ are points along the curve as a function of the parameter u , d_i is the point coefficient, and N_i^k is the i th B-spline basis function of order k .

The coefficients are determined by partly by the data to be fitted and partly by an added penalty function aiming to impose smoothness and avoid over fitting.

4.2 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical method that employs orthogonal transformation in order to translate a set of possibly correlated variables to a set of linearly uncorrelated variables called principal components. The aim of PCA is to be able to explain the variation in an observed dataset by the principal components, which are considerably fewer than the number of original variables. The principal components are orthogonal since they are the eigenvectors of the covariance matrix, which is symmetric (Ramsay & Silverman, 2005).

The procedure is built on the central concept of expressing functions as linear combinations of variable values,

$$f_i = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, N \quad (4.3)$$

where β_j is a weighting coefficient applied to the observed values x_{ij} of the j th variable. Principal component analysis is performed by finding the set of normalized weights that maximizes the variation in the f_i 's. The procedure can be explained through the following steps (Ramsay & Silverman, 2005):

1. Find the weight vector $\boldsymbol{\varepsilon}_1 = (\varepsilon_{11}, \dots, \varepsilon_{p1})'$ for which the linear combination values

$$f_{i1} = \sum_j \varepsilon_{j1} x_{ij} \quad (4.4)$$

have the largest possible mean square $N^{-1} \sum_i f_{i1}^2$ under the constraint

$$\sum_j \varepsilon_{j1}^2 = \|\boldsymbol{\varepsilon}_1\|^2 = 1. \quad (4.5)$$

2. Carry out second and subsequent steps, possibly up to a limit of the number of variables p . On the m th step, compute a new weight vector $\boldsymbol{\varepsilon}_m$ and new values $f_{im} = \boldsymbol{\varepsilon}_m' \boldsymbol{x}_i$. The values f_{im} will have maximum mean square, under the constraint $\|\boldsymbol{\varepsilon}_m\|^2 = 1$, and the $m - 1$ additional constraint(s)

$$\boldsymbol{\varepsilon}_k' \boldsymbol{\varepsilon}_m = 0, \quad k < m. \quad (4.6)$$

The unit constraint on the weights is essential to the procedure. Without this constraint, the problem is ill defined since the mean squares could take arbitrarily large values. The motivation behind the first step is to find the strongest and most important mode of variation in the variables. In the second and following steps,

one again seeks the most important mode of variation but require the weights to be orthogonal to the ones previously found, so they are indicating something new. The amount of variation in terms of mean square will decrease for each step, and one can expect to be able to explain the major part of the variance between variables well before the maximum step p . The vectors $\boldsymbol{\varepsilon}_i$ are called the loading vectors and the f_i 's contain the principal component scores.

4.3 The Generalized Champernowne Distribution

The original Champernowne distribution was first proposed by D. G. Champernowne in 1937 to describe a family of curves for the graduation of pre-tax income distributions. It was further developed and discussed by the author in 1952, where methods are described for fitting the distribution parameters (Champernowne, 1952). The probability density function for the original Champernowne distribution is:

$$f(x) = \frac{c_*}{x \left(\frac{1}{2} \left(\frac{x}{M} \right)^{-\alpha} + \lambda + \frac{1}{2} \left(\frac{x}{M} \right)^{\alpha} \right)} \quad x \geq 0, \quad (4.7)$$

where c_* is a normalizing constant, and α, λ and M are nonnegative parameters.

The characteristic features of the Champernowne distribution is the convergence to a Pareto distribution in the tail, while looking like a Log-Normal distribution near 0 when $\alpha > 1$. In the case when $\alpha \neq 1$, the density is either 0 or infinity at 0. One of the benefits of using the Champernowne distribution instead of Extreme Value Theory is that one does not have to choose the starting point from where the tail domain is defined. In order to avoid the inflexibility of the distribution at 0, Bolancé et al. (2012) present the Generalized Champernowne distribution, which includes an additional parameter c ensuring the possibility of a positive finite value at 0 for all α . The cumulative distribution function of the Generalized Champernowne distribution is:

$$F_{\alpha, M, x}(x) = \frac{(x+c)^\alpha - c^\alpha}{(x+c)^\alpha + (M+c)^\alpha - 2c^\alpha}, \quad x \geq 0, \quad (4.8)$$

where $\alpha > 0, M > 0$ and $c \geq 0$. The probability density function of the Generalized Champernowne distributions is:

$$f_{\alpha, M, c}(x) = \frac{\alpha(x+c)^{\alpha-1}((M+c)^\alpha - c^\alpha)}{((x+c)^\alpha + (M+c)^\alpha - 2c^\alpha)^2}, \quad x \geq 0. \quad (4.9)$$

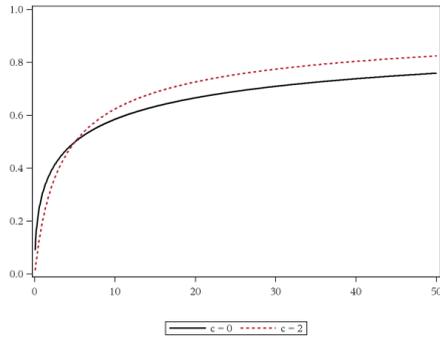
The Generalized Champernowne distribution does equivalently to the Champernowne distribution, converge to a Pareto distribution in the tail. Let us define

$$g_{\alpha, M, c}(x) = \frac{\alpha \left(((M + c)^\alpha - c^\alpha)^{\frac{1}{\alpha}} \right)^\alpha}{x^{\alpha+1}}, \quad (4.10)$$

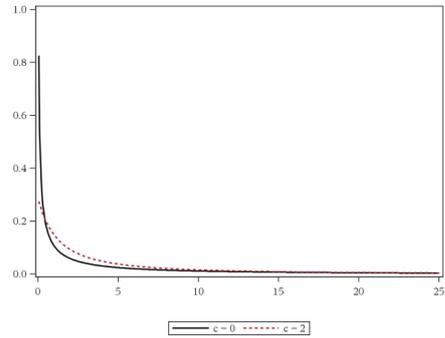
which is the density function of the Pareto distribution with the mode of the Generalized Champernowne distribution inserted as x_m . Then the convergence imply that

$$\lim_{x \rightarrow \infty} \frac{f_{\alpha, M, c}(x)}{g_{\alpha, M, c}(x)} = 1. \quad (4.11)$$

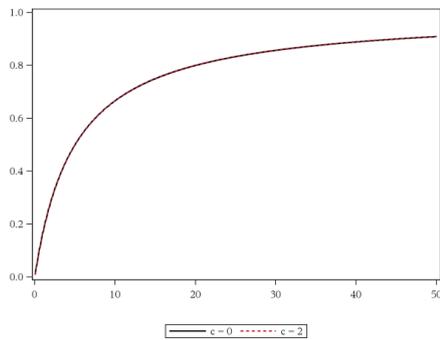
The effect of the additional parameter c introduced in the generalized version of the distribution is different for different values of α . The parameter c experiences some scale parameter properties, materializing as the derivative of the cumulative distribution function becoming larger with an increasing c when $\alpha < 1$. Conversely, the derivative of the cumulative distribution function decreases with increasing c when $\alpha > 1$. When $\alpha \neq 1$, the parameter c affects the density in different ways. When $\alpha < 1$, a positive c leads to lighter tails and the opposite when $\alpha > 1$. Furthermore, a positive c ensures a finite density at 0. The parameter also has a shifting effect as a positive c shifts the mode to the left when $\alpha > 1$. In the scenario where $\alpha = 1$, the parameter c has no effect on the density. The influence of the parameter c is visualized in Figure 4.2.



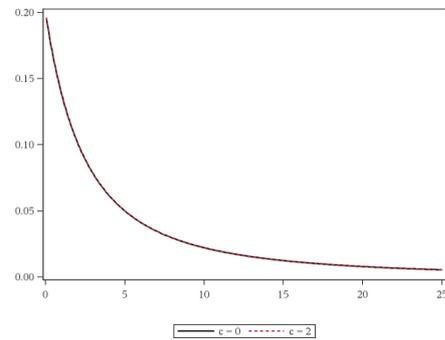
CDF, $\alpha = 0.5$



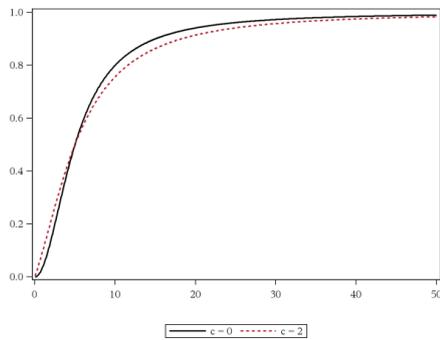
PDF, $\alpha = 0.5$



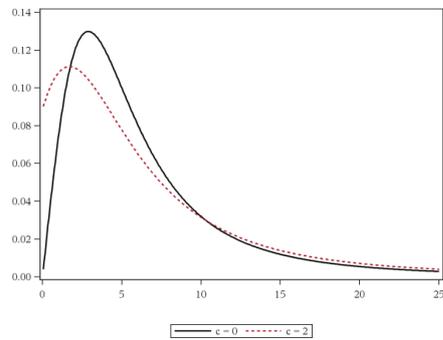
CDF, $\alpha = 1$



PDF, $\alpha = 1$



CDF, $\alpha = 2$



PDF, $\alpha = 2$

Figure 4.2 Illustrative examples of the effect of parameter c . The cumulative distribution function and the probability density function of the Generalized Champernowne Distribution are plotted for different values of α and c , while $M = 5$ in all plots. $c = 0$ in the solid line and $c = 2$ in the dotted line.

4.4 Kernel Density Estimation

Kernel density estimation is a nonparametric method for approximation of the probability density function of a random variable (Bolancé, et al., 2012). The method will provide an estimation of the density function at every point in the domain by using sample information in the neighborhood of the point where the density function is estimated. The influence of empirical data will be large from nearby points and small from points far away. How the sample information is combined and weighed is a function of both the kernel function as well as the bandwidth parameter.

For a sample of n independent identically distributed observations X_1, X_2, \dots, X_n , the kernel density estimator of the density function f is defined as:

$$\hat{f}(x) = \frac{1}{nb} \sum_{i=1}^n K\left(\frac{x - X_i}{b}\right), \quad (4.12)$$

where $K(\cdot)$ is the kernel function and b is the bandwidth parameter. Both $K(\cdot)$ and b have to be chosen. The kernel function determines the shape of the weighting, while the bandwidth parameter determines the width and thereby the smoothing factor. The shape of the kernel function does not influence the shape of the estimated density. Kernel density estimation can also be implemented if one is interested in estimating a multivariate density.

The kernel function is a function of a single variable and must integrate to 1. The functions is usually symmetric with zero mean and commonly used kernel functions are Gaussian, Epanechnikov, and Uniform although the choice of kernel functions is not limited to density functions. The kernel functions determine the way in which empirical data is handled in the estimation function, where for instance a Gaussian kernel weights data according to a normal probability distribution and the density estimation will therefore be influenced by data points far from the point of estimation, while an Epanechnikov kernel will not use information from data points outside its domain. Table 4.1 displays some of most commonly used kernel functions. The bandwidth parameter b controls the width of the kernel and determines thereby the amount of smoothing. For a smaller choice of b , the density estimation will experience larger fluctuations while a larger b results in a smoother density estimation, but details from the data points may be lost.

One potential drawback of using kernel density estimation when using heavy tailed data is that the density will be biased in the mode, implying that the density will be underestimated in this area (Bolancé, et al., 2012). The smoothing parameter will also tend to be large. When using a smaller bandwidth, the density may result in a bumpy shape due to the presence of scarce large observations.

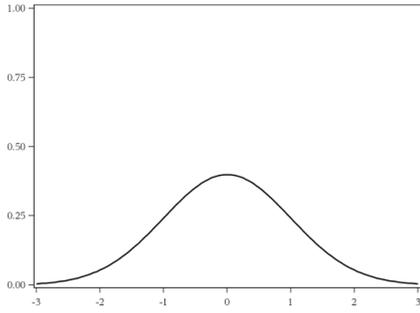
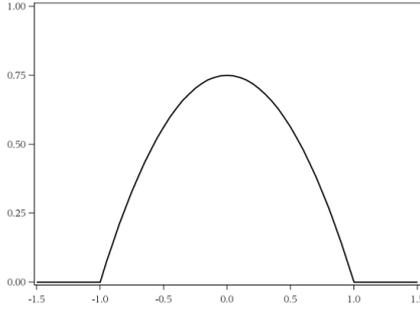
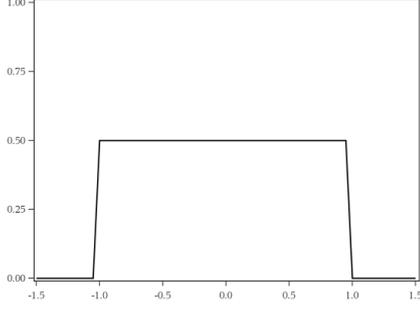
Kernel	$K(u)$	Plot
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ for $ u \leq \infty$	
Epanechnikov	$\begin{cases} (3/4)(1-u^2) & \text{for } u \leq 1 \\ 0 & \text{otherwise} \end{cases}$	
Uniform	$\begin{cases} 1/2 & \text{for } u \leq 1 \\ 0 & \text{otherwise} \end{cases}$	

Table 4.1 Three commonly used kernel functions.

4.4.1 Boundary Correction

The estimation of a density function introduces restrictions as the resulting estimate must integrate to one. The classical kernel density estimation is developed with unbounded support with no knowledge of boundaries, resulting in a density with probability mass outside the support and an estimate without unit integration. Figure 4.3 illustrates one scenario where boundary correction is needed. This would result in a considerable bias of the estimator, and simply truncating the distribution would be inappropriate and insufficient (Jones, 1993). In this thesis boundary correction is needed as data are transformed into the $[0, 1]$ interval. Many approaches to boundary correction are available and this thesis will use the one presented by Bolancé et al. (2012). The method forces the integration of each kernel overspilling the boundaries to unity by renormalizing. The following function gives the asymptotic properties of the kernel density estimator around boundaries.

$$a_{kl}(y, b) = \int_{\max\{-1, \frac{y-1}{b}\}}^{\min\{1, \frac{y}{b}\}} u^k K(u)^l du \quad (4.13)$$

for $y \in [0, 1]$, where $K(\cdot)$ is the kernel function.

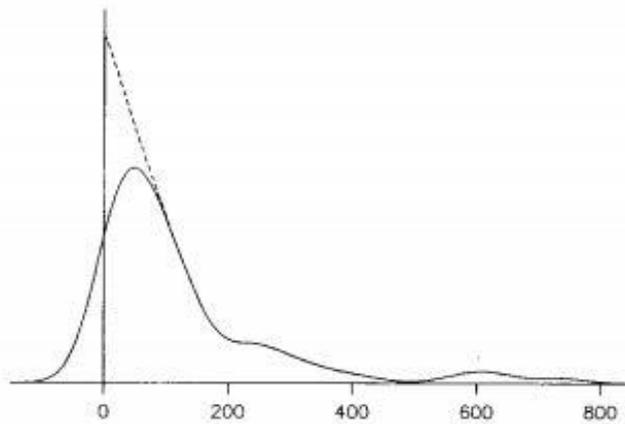


Figure 4.3 Illustrative example of kernel density estimation with boundary correction. The solid line is raw kernel density estimate and the dotted line is with boundary correction. (Jones, 1993, p. 136)

Note that $a_{11}(y, b) = 0$ and $a_{01}(y, b) = 1$ for the interior points of y , that is on the interval $[b, 1 - b]$. The integral only takes nontrivial values within one bandwidth away from the boundary points.

Using equation 4.13, the kernel density estimator with boundary correction will then be defined as:

$$\hat{f}(y) = \frac{1}{nba_{kl}(y,b)} \sum_{i=1}^n K\left(\frac{y - Y_i}{b}\right). \quad (4.14)$$

Equation 4.14 is simply equation 4.12 with the introduced correction term.

Chapter 5

Methodology

5.1 Time Variant Factor Scaling

In order to analyze the internal dataset from the bank together with external data from an operational risk data consortium, the internal operational risk losses have been converted to Euro. The conversion has been done according to historical exchange rates provided by European Central Bank (European Central Bank, 2015). As stated by Schevchenko & Temnov (2009), in practice, losses are often scaled using time variant factors before used in modeling. One such factor is inflation which in this thesis has been adjusted for using the historical euro inflation reported by the European Central Bank (European Central Bank, 2015). Each loss has been adjusted as if it occurred in January 2015. Furthermore, the reporting threshold is scaled correspondingly, resulting in a known threshold varying in time as suggested by Schevchenko & Temnov.

5.2 Semiparametric Density Estimation

This thesis will operate under the assumption that internal losses follow the same distribution as external losses conditioned on external losses being truncated below a threshold value. This assumption results in the approach to predict the continuation of the probability density function of the external loss data based on information from both its tail as well as information from the internal loss dataset. Therefore the prediction will be based on previously observed probability density functions of the internal loss dataset, but given that only one observation is available we will estimate the distribution of the internal losses to generate more observations. The estimation of the distribution will be performed using a semiparametric model presented by Bolancé (2012).

The method will find parameters of the Generalized Champernowne distribution, presented in chapter 4.3, that fit the dataset and will then continue with kernel density estimation. The first step in finding the estimated distribution is to estimate the parameters α, M, c of the cumulative distribution function. For the Generalized Champernowne distribution it holds that $F_{\alpha, M, c}(M) = 0.5$, suggesting the parameter M should be estimated as the empirical median of the

dataset. The parameters α and c are then estimated using Maximum Likelihood estimation with the Log-Likelihood function following equation 4.9:

$$\begin{aligned}
l(\alpha, c) = & n \log \alpha + n \log((M + c)^\alpha - c^\alpha) \\
& + (\alpha - 1) \sum_{i=1}^n \log(X_i + c) \\
& - 2 \sum_{i=1}^n \log((X_i + c)^\alpha + (M + c)^\alpha - 2c^\alpha).
\end{aligned} \tag{5.1}$$

For a fixed M , the Maximum Likelihood function is concave and has a maximum. The maximization is performed numerically using the Newton-Rhapson method. Figure 5.1 below shows the probability density function for a Generalized Champernowne distribution evaluated with the parameters found via Maximum Likelihood.

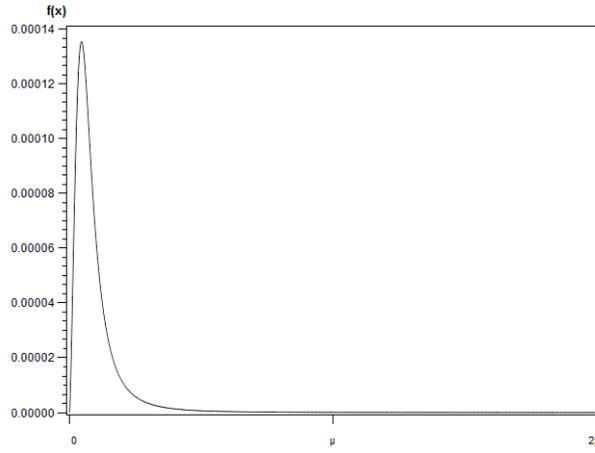


Figure 5.1 A probability density function of the Generalized Champernowne distribution, with the parameter found via Maximum Likelihood.

The next step in the process is to transform the dataset into the $(0, 1)$ interval. Using the estimated parameters $\hat{\alpha}$, \hat{M} , and \hat{c} previously found, the data are transformed using the cumulative distribution function of the Generalized Champernowne distribution. Transformation of the data is based on the probability transform that says that if X is a random variable with a continuous distribution function F , then $F(X)$ is uniformly distributed on the interval $(0, 1)$ (Hult, et al., 2012). In this case F will be the Generalized Champernowne distribution, and we will receive the transformed variable

$$Y_i = F_{\hat{\alpha}, \hat{M}, \hat{c}}(X_i), \tag{5.2}$$

where $i = 1, \dots, n$.

The transformation is designed to make the transformed data as close to a uniform distribution as possible. However, even if the transformed dataset does

not appear to be uniformly distributed, it still contains the correct characteristics. Figure 5.2 below shows a transformed dataset.

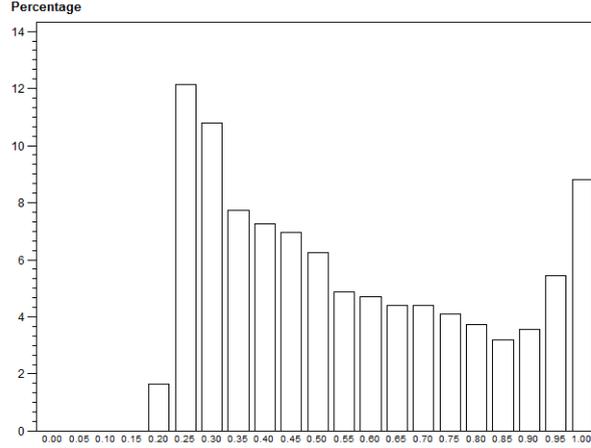


Figure 5.2 Histogram over a loss distribution, transformed using the cumulative distribution function of the Generalized Champernowne distribution with parameters estimated from Maximum Likelihood.

Following the transformation, the next step is to implement kernel density estimation, where boundary correction is added. The kernel density estimation is calculated using equation 4.14, with the transformed variables and the accurate boundary correction limits. We now receive the following density estimation:

$$\hat{f}_{transformed}(y) = \frac{1}{nba_{0,1}(y,b)} \sum_{i=1}^n K\left(\frac{y - Y_i}{b}\right), \quad (5.3)$$

where $a_{0,1}(y,b)$ is the boundary correction term from equation 4.13. The boundary correction term allows the kernel density estimation to reach unit integration, which is a central aspect in transforming operational losses. By transforming the original dataset to a Uniform distribution, and require the density to integrate to one, one is able to estimate beyond the data and extrapolate past the maximum observed value in the original dataset (Bolancé, et al., 2012). Even if one does not have full information on the tail, one can use the integration restriction to extract valuable information from the density estimation.

The kernel function used is the Epanechnikov since it is bounded and the most efficient. Silverman's rule of thumb is used to calculate the bandwidth (Silverman, 1986). Silverman's rule of thumb calculates a bandwidth aiming to minimize the Mean Integrated Square Error. The Bandwidth is calculated as follows:

$$b = \hat{\sigma} \left(\frac{40\sqrt{\pi}}{n} \right)^{\frac{1}{5}}, \quad (5.4)$$

where $\hat{\sigma}$ is the standard deviation estimated from the sample. Given that the bandwidth is proportional to $n^{-1/5}$, an increase in sample size will decrease the bandwidth and thereby reduce the smoothing effect. Figure 5.3 below shows the kernel density estimation of a transformed dataset.

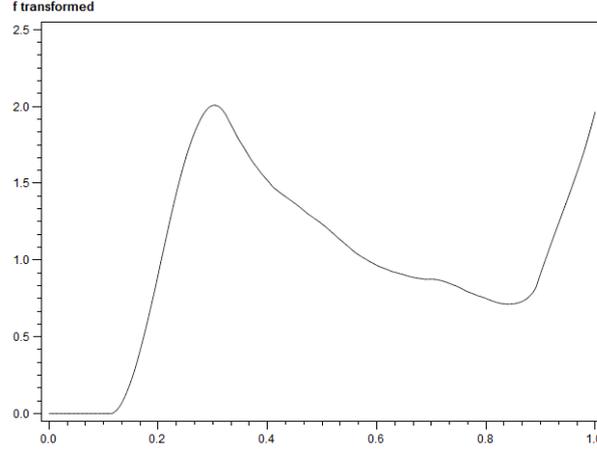


Figure 5.3 Kernel density estimation of a transformed dataset.

The final step of the semiparametric approach is to convert the probability density function from the kernel estimation, by a back-transformation to its original scale. Following this transformation, we will have a smoothed probability density function that we can interpret and use properly. The transformed data is converted using $F_{\hat{\alpha}, \hat{M}, \hat{c}}^{-1}(y)$. We find the expression for the back-transformation as:

$$\begin{aligned} \hat{f}(x) &= \hat{f}_{transformed}(F_{\hat{\alpha}, \hat{M}, \hat{c}}(x)) \left| \frac{dF_{\hat{\alpha}, \hat{M}, \hat{c}}(x)}{dx} \right| \\ &= \hat{f}_{transformed}(y) F'_{\hat{\alpha}, \hat{M}, \hat{c}}(x). \end{aligned} \quad (5.5)$$

Exchanging the expression for $\hat{f}_{transformed}(y)$ with equation 5.3 yields the final expression for the semiparametric process as:

$$\hat{f}(x) = \frac{F'_{\hat{\alpha}, \hat{M}, \hat{c}}(x)}{nb\alpha_{01}(F_{\hat{\alpha}, \hat{M}, \hat{c}}(x), b)} \sum_{i=1}^n K\left(\frac{F_{\hat{\alpha}, \hat{M}, \hat{c}}(x) - F_{\hat{\alpha}, \hat{M}, \hat{c}}(X_i)}{b}\right), \quad (5.6)$$

where $F'(x)$ is the derivative of $F(x)$.

The result from all the proceeding steps is presented in Figure 5.4 below.

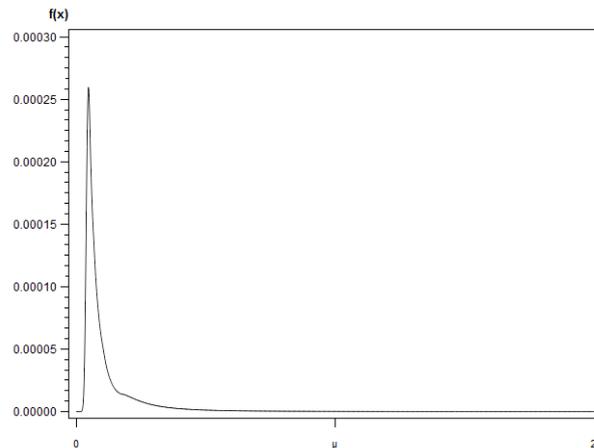


Figure 5.4 Resulting distribution from the semiparametric estimation of a dataset

Following this semiparametric process, a probability density function explaining the behavior of the internal loss process is obtained. The procedure is termed semiparametric as it provides a link between parametric and nonparametric methodologies, posing certain advantages. When data are sparse, one does not want to rely solely on nonparametric estimation, and the estimation will therefore be close to the parametric model. When available data increases, the model will converge to its nonparametric counterpart. Additionally, this model will solve the problems associated with kernel density estimation in the tail, which often results in a density estimation with a bumpy shape. The proposed model also provides advantages over extreme value theory, as it enables the fitting of a Pareto-like tail without the need to identify the domain in which the tail distribution is defined (Bolancé, et al., 2012).

The probability density function obtained from this approach will be used in the subsequent step to generate observations of the internal loss dataset. The next step is to generate datasets with a ranging amount of observations, as described in section 6.2, which will be used to forecast the continuation of the probability density function of the external dataset. The approach will be applied to three risk cells, following the categorization described in chapter 2.2.3. Risk cells have been chosen as cells with a large sample size in both the internal and external dataset, since the semiparametric approach to fit a Generalized Champernowne distribution will yield more accurate results when using more observations. For confidentiality reasons, this thesis will not disclose which intersections the risk cells represent. The choice of risk cells will in no way affect the results of this research, as the approach is applicable to every cell with sufficient data. Since the external dataset is more extensive than the internal counterpart, we must be able to generate observations containing larger losses than the ones encountered in the internal dataset. Using the proposed semiparametric approach, this is achieved, while still using information from internal losses to predict external losses according to the assumption of common distribution

5.3 Best Linear Unbiased Predictor

Predicting the probability density function of the external operational risk loss data will be performed using the Best Linear Unbiased Predictor (BLUP), presented by Goldberg et al. (2014). The approach uses information from previously collected functions, obtained from the semiparametric approach. The initial step of the prediction is to represent the curves generated from the internal dataset with a basis expansion with respect to a B-spline basis $\mathbf{b} = (b_1, \dots, b_N)'$, defined on a fixed knot sequence $\boldsymbol{\tau}$. The previously collected curves can now be represented through coefficients of the B-spline basis. The next step is to express the B-spline curves with a small number of variance factors, found via Principal Component Analysis. From this analysis eigenvectors, eigenvalues, and mean function are extracted, all expressed in the basis \mathbf{b} .

Once extracted, the eigenvectors and mean function are restricted to the two intervals defined by the part observed in the external dataset, part 1, and the part to predict, part 2. The eigenvectors and mean functions are expressed in terms of two new B-spline bases defined on the two subintervals. If we define the function to predict as X , one can divide it into two parts X_1 and X_2 . X_1 is the beginning part, which have already been observed and is therefore not stochastic and hence defined as \mathbf{x}_1 . The mean functions of part 1 and 2, defined as $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, are based on the previously observed curves. The prediction of X_2 is then performed by adjusting $\boldsymbol{\mu}_2$ by the deviation of \mathbf{x}_1 from $\boldsymbol{\mu}_1$, scaled with the observed variance. The formula for the prediction is as follows:

$$\hat{X}_2(t) = \mathbf{b}_2(t)'(\boldsymbol{\mu}_2 + A_2(A_1'A_1 + \sigma^2 L^{-1})^{-1}A_1'(\mathbf{x}_1 - \boldsymbol{\mu}_1)) \quad (5.7)$$

Where L is a diagonal matrix with the eigenvalues from the principal component analysis on the diagonal, σ^2 is the scaled sum of the eigenvalues, the A_i 's are the coefficients of the restricted eigenvectors, and \mathbf{b}_2 is the B-spline basis restricted to part 2. For a more extensive explanation of the prediction, please see the article by Goldberg et al (2014).

The proposed method will enable the recreation of the truncated body of a loss distribution, by using information from its tail and previously collected losses. The method will not be inferring a parametric distribution on the operational risk losses, which is a desirable feature as previous research has failed to reach consensus regarding which parametric distribution that best fit operational risk losses. The proposed method of linear prediction has been developed and validated by Goldberg et al. (2014) and verified to work on call center data. The same dataset has been tested by this thesis with good results. By employing the same procedure as Goldberg et al (2014), this thesis will also compute confidence bands for the prediction.

Chapter 6

Data

6.1 Internal Data

An internal dataset on operational risk losses is used in this thesis and contains fewer observations than the external dataset. Internal losses are reported above a threshold too low to influence the data characteristics or total loss amount. For confidentiality reasons, this thesis will display two scale parameters μ and θ for the axes. The parameters are constant through the entire thesis.

Losses have been converted to Euro and adjusted for inflation, to be comparable with the external loss data. The distribution is positively skewed, indicating that the mass of the distribution is concentrated on the left of the distribution and that the right tail is heavier than the left tail. A stochastic variable that is symmetric will have skewness of zero. The kurtosis of the distribution is positive, which is indicative of a distribution with heavier tail and more peakedness than a normal distribution. The measure of kurtosis is centered on the normal distribution which has adjusted kurtosis of 0. Figure 6.1 below displays a histogram over the internal loss dataset, where the last bar is includes all losses above 2μ , as the tail is long.

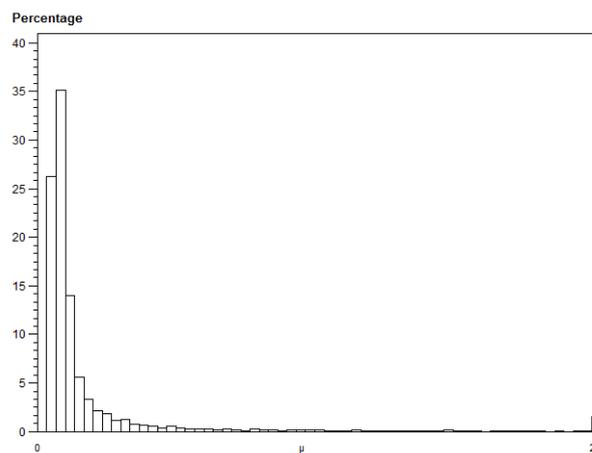


Figure 6.1 Histogram over the internal dataset of operational risk losses used in this thesis. The last bar includes losses above 2μ

6.1.1 Risk Cell 1

Three risk cells have been chosen for analysis in this thesis, and are therefore presented. A histogram over the losses within risk cell 1 is displayed below.

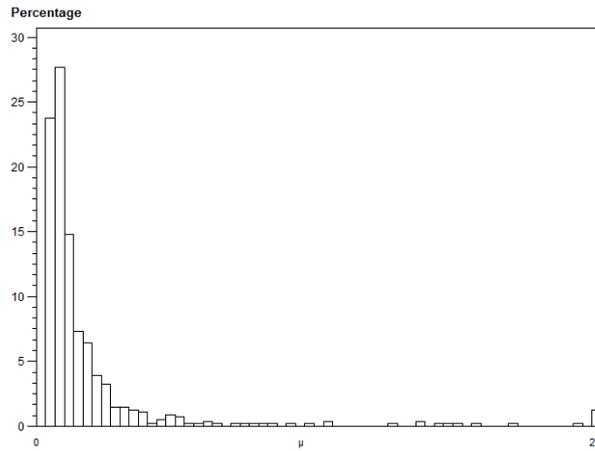


Figure 6.2 Histogram over the internal dataset of operational risk losses for risk cell 1. The last bar includes losses above 2μ

6.1.2 Risk Cell 2

A histogram over the losses in risk cell 2 displayed below. The distribution of this risk cell is the least skewed and has lowest kurtosis, indicating a lighter tail than the other two datasets.

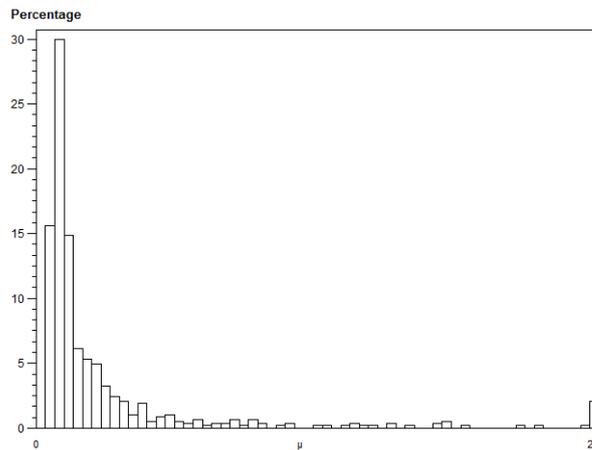


Figure 6.3 Histogram over the internal dataset of operational risk losses for risk cell 2. The last bar includes losses above 2μ

6.1.3 Risk Cell 3

A histogram over the losses in risk cell 3 displayed below. The distribution of this risk cell is the least skewed and has lowest kurtosis, indicating a lighter tail than the other two datasets. The distribution of this risk cell is more skewed and has higher kurtosis than the other two risk cells, indicating a heavier tail. This is the risk cell with the largest observation.

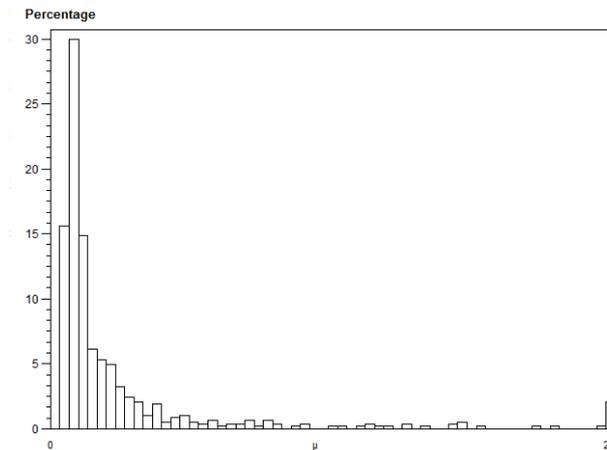


Figure 6.4 Histogram over the internal dataset of operational risk losses for risk cell 3. The last bar includes losses above 2μ

6.2 Simulated Data

In order to apply the prediction method proposed by Goldberg et al. (2014) a number of observed complete datasets are required. The internal dataset only consists of a single observation and datasets are therefore simulated to be used in this thesis.

Building on the semiparametric approach presented by Bolancé et al. (2012), datasets are simulated by drawing from the density distribution resulting from the semiparametric approach based on the internal data. For chosen risk cells in the correlation matrix between business line and event type, datasets are drawn with a ranging number of losses up to twice the amount of expected losses in the external data, thereby placing the mean of the historical number of losses close to the amount in the external dataset, which is to be predicted.

6.3 External Data

External operational risk loss data can be classified into two main categories: public data and consortium data, where consortium data are reported above a stated threshold. A threshold is introduced as a mean to ensure the financial institutions' ability to report every operational risk loss as well as to reduce costs associated with collecting and reporting. This thesis will use consortium data where the threshold is known.

The distribution of the dataset used in this thesis is positively skewed, indicating that the mass of the distribution is concentrated on the left of the distribution and that the right tail is heavier than the left tail. The kurtosis of the distribution is positive, which is indicative of a distribution with heavier tail and more peakedness than a normal distribution. Figure 6.5 below shows a histogram over the losses where the last bar is cumulative for presentational purposes. All losses above 2μ are shown in the rightmost bar of the chart. Noteworthy is the reporting threshold and the very large last bar in Figure 6.5, indicating a heavy tail. Furthermore, the external dataset is larger than the internal counterpart and the largest observation is much bigger in the external data.

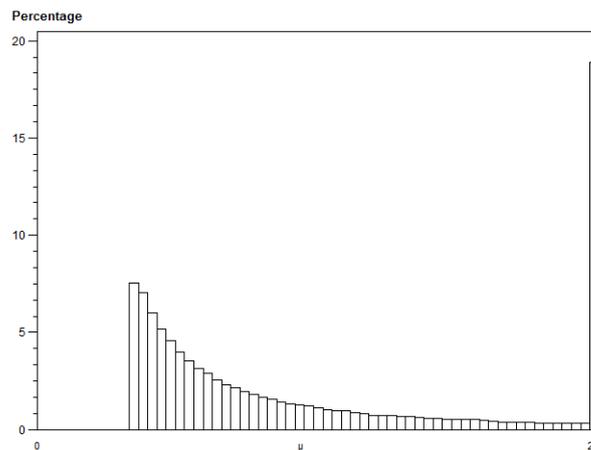


Figure 6.5 Histogram over the external dataset of operational risk losses. The last bar includes losses above 2μ

6.3.1 Risk Cell 1

A histogram over the external losses in risk cell 1 is presented below.

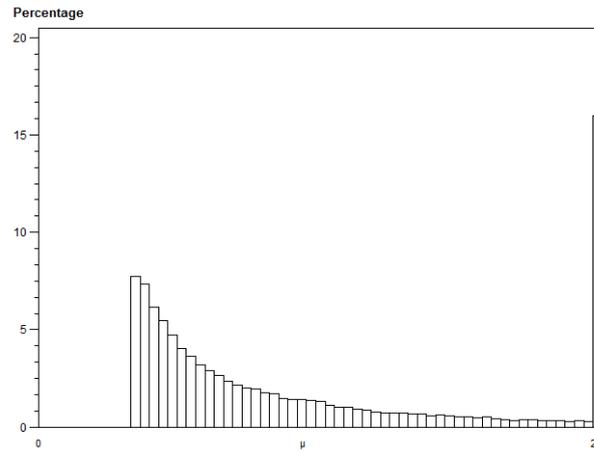


Figure 6.6 Histogram over the external dataset of operational risk losses risk cell 1. The last bar includes losses above 2μ

6.3.2 Risk Cell 2

A histogram over the external losses in risk cell 2 is presented below. This risk cell contains relatively few small observations, and many large losses. This is confirmed by the high skewness and kurtosis.

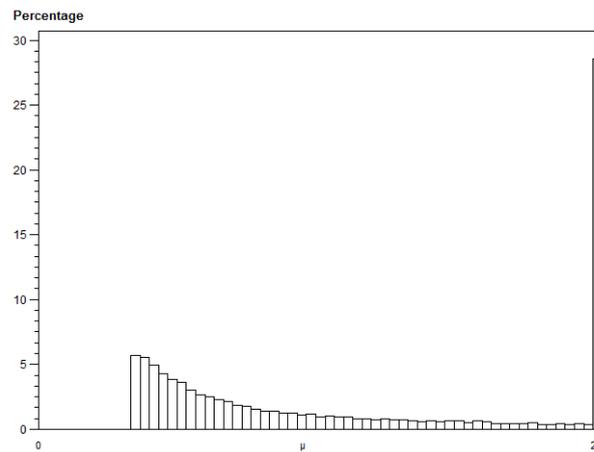


Figure 6.7 Histogram over the external dataset of operational risk losses for risk cell 2. The last bar includes losses above 2μ

6.3.3 Risk Cell 3

A histogram over the external losses in risk cell 3 is presented below. This risk cell contains many small observations, and relatively few large losses.

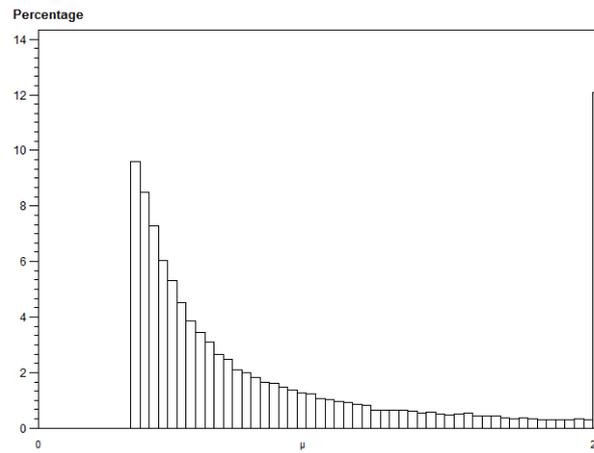


Figure 6.8 Histogram over the external dataset of operational risk losses for risk cell 3. The last bar includes losses above 2μ

Chapter 7

Results

7.1 Results from the Best Linear Unbiased Prediction

The prediction approach presented in the methodology chapter is performed on three risk cells in the correlation matrix between business line and event type. The following three figures demonstrate the results from the prediction in the different risk cells where the solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function. In order to give the best examination of the prediction part the entire loss domain is not shown in the figures. Evident is the discreteness in the observed part, as is the smooth nature of the prediction part, following the use of penalized B-splines.

The figures indicate the validity of the prediction method, where the curve representing the observed part smoothly transition to the curve representing the forecasted continuation, demonstrating that the mean function of the previously observed curves have been properly adjusted to fit the external dataset, according to equation 5.7. The study results in prediction of an external dataset on operational risk losses, resembling the internal loss dataset.

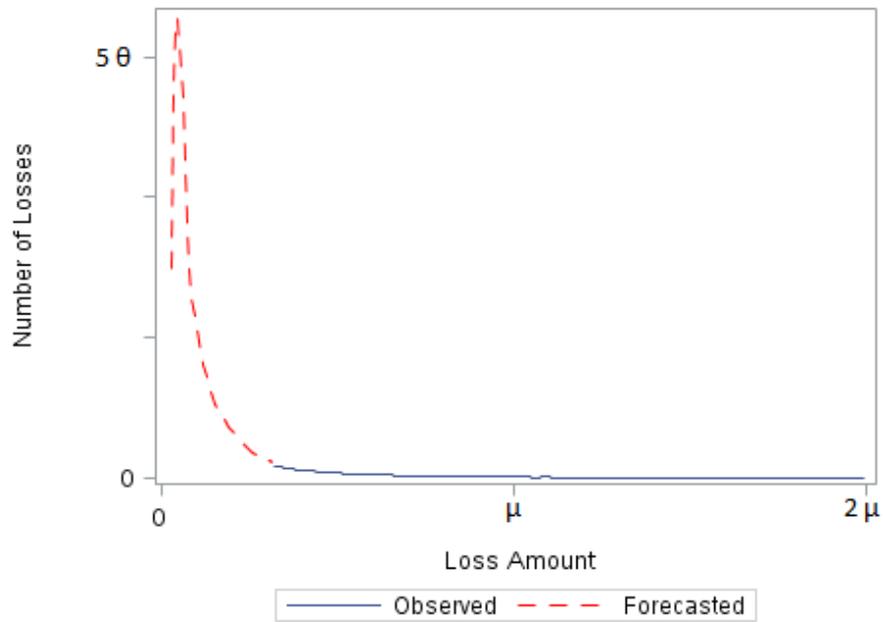


Figure 7.1 The result from the prediction of the continuation of the loss distribution in risk cell 1. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.

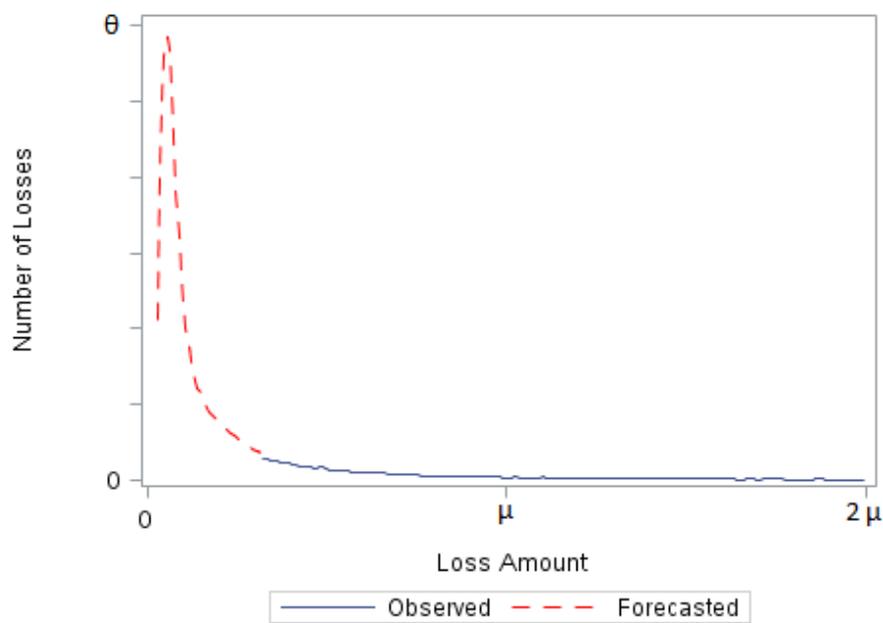


Figure 7.2 The result from the prediction of the continuation of the loss distribution in risk cell 2. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.

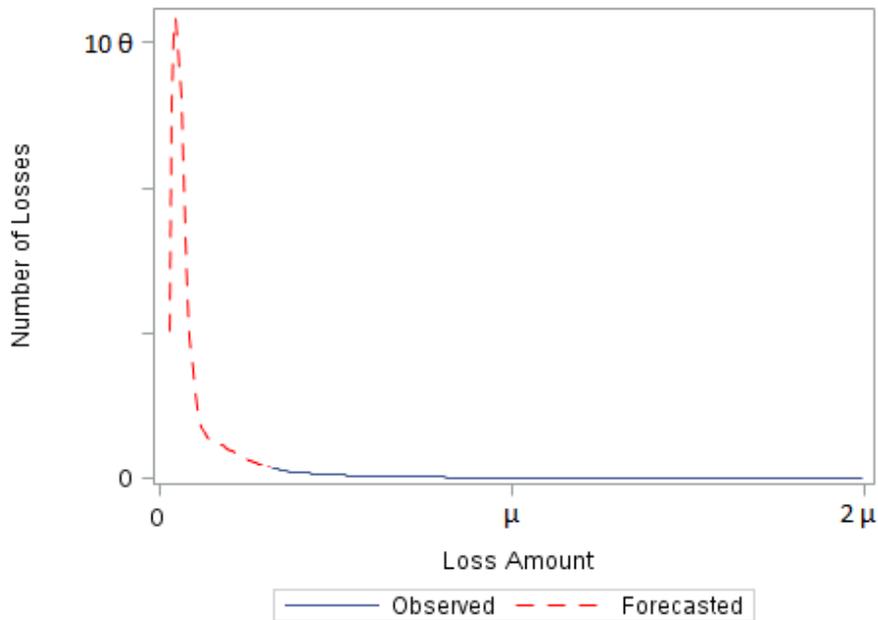


Figure 7.3 The result from the prediction of the continuation of the loss distribution in risk cell 3. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.

7.2 Accuracy of the Prediction

One important consideration is that the predicted continuation part of the function involves uncertainty, and therefore requires confidence bands to assess the accuracy. The following figures will display the same figures as in chapter 7.1 but with confidence bands included. The interested reader is referred to the appendix for more detailed pictures. The solid blue line is the observed part of the external dataset, the dashed red line is the forecasted continuation of the density function, and the solid green lines are the confidence bands at a 95% level. In order to give the best examination of the prediction part the entire loss domain is not shown in the figures.

The figures illustrate the precision of the prediction, and notable is high precision, visualized as narrow confidence bands, in the prediction part close to the threshold value as well as the leftmost part. The confidence bands are wider at the mode of the density, illustrating a less precise prediction. Wider confidence bands in this region are to be expected given that this area experiences more variations following the peakedness of the distribution. Furthermore, the accuracy of the prediction does not vary significantly between the different risk cells, indicating a robustness of the prediction method.

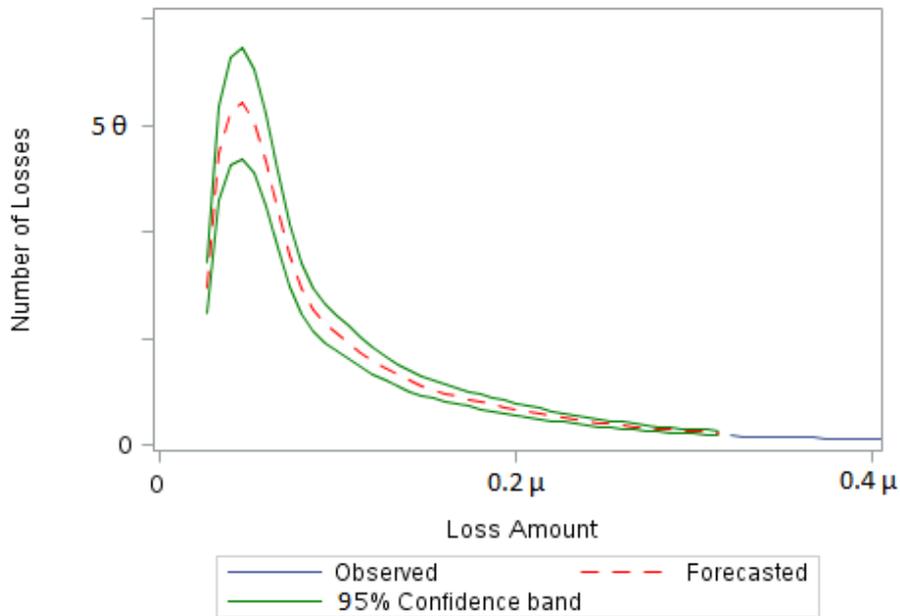


Figure 7.4 The result from the prediction of the continuation of the loss distribution in risk cell 1 including confidence bands. The solid blue line is the observed part of the external dataset, the dashed red line is the forecasted continuation of the density function, and the solid green lines are the confidence bands at a 95% level.

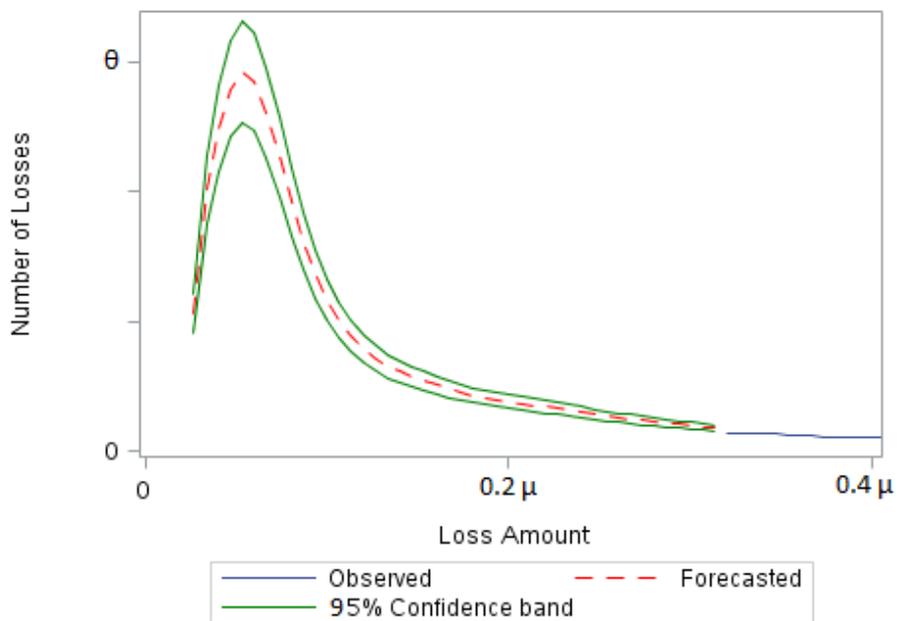


Figure 7.5 The result from the prediction of the continuation of the loss distribution in risk cell 2. The solid blue line is the observed part of the external dataset, the dashed red line is the forecasted continuation of the density function, and the solid green lines are the confidence bands at a 95% level.

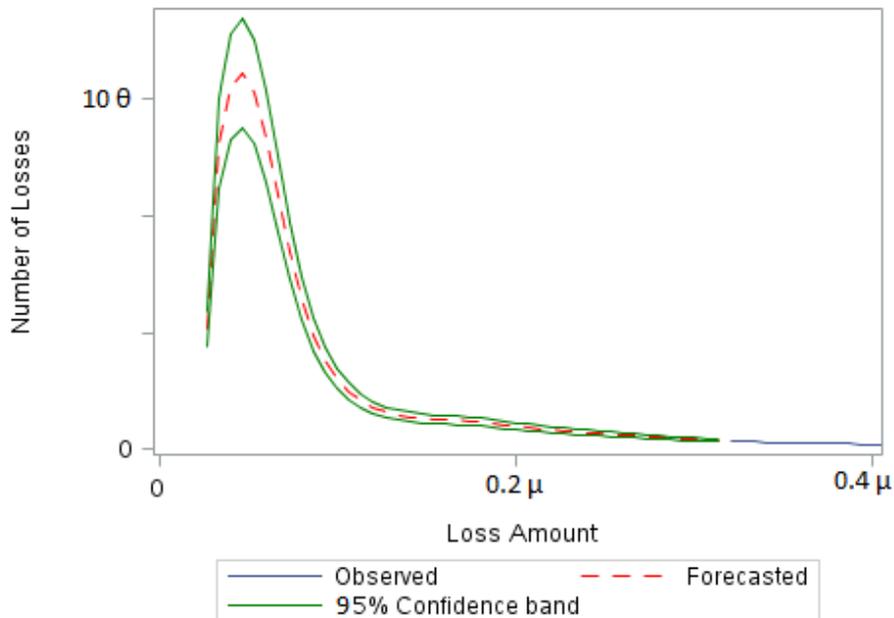


Figure 7.6 The result from the prediction of the continuation of the loss distribution in risk cell 3 including confidence bands. The solid blue line is the observed part of the external dataset, the dashed red line is the forecasted continuation of the density function, and the solid green lines are the confidence bands at a 95% level.

7.3 Comparing Risk Cells

To illustrate the difference between risk cells, the same figures are displayed with equal axes. The following three figures display the result from the prediction on the different risk cells, all in the same scale. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.

The figures indicate that the number of losses differs significantly between the cells, as well as the shapes of the resulting distributions, suggesting that different factors influence both loss severity as well as loss frequency in different business lines and event types. The result confirms the need to categorize the losses.

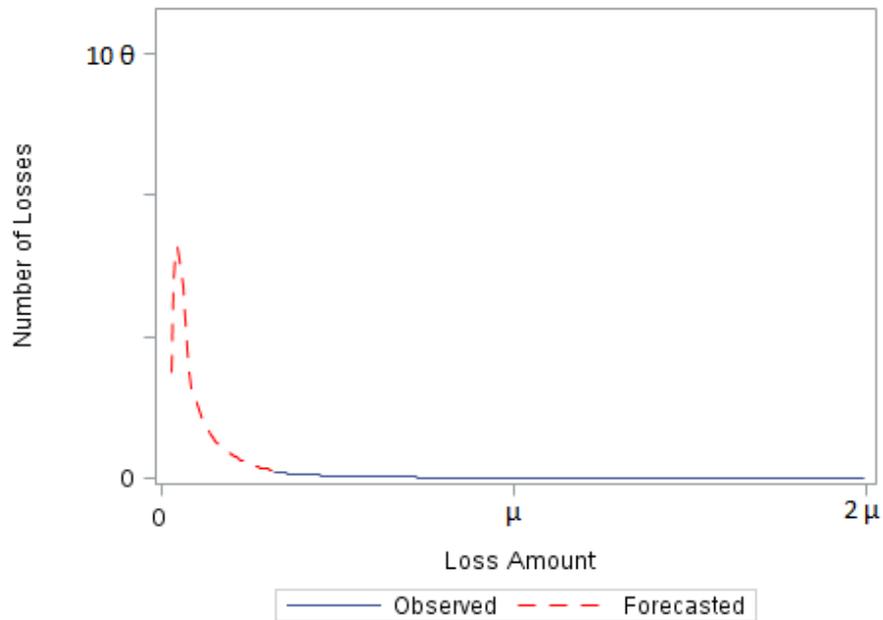


Figure 7.7 The result from the prediction of the continuation of the loss distribution in risk cell 1. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.

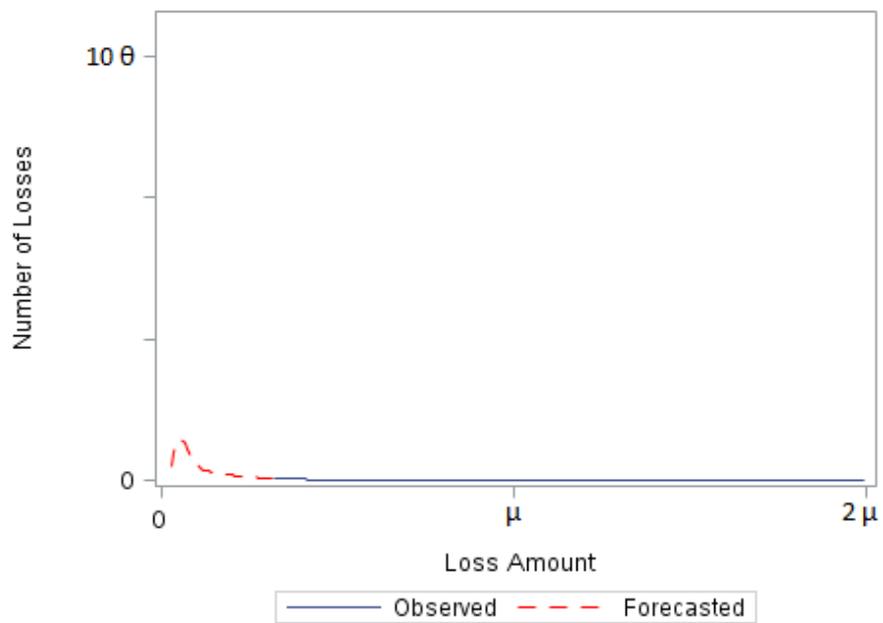


Figure 7.8 The result from the prediction of the continuation of the loss distribution in risk cell 2. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.

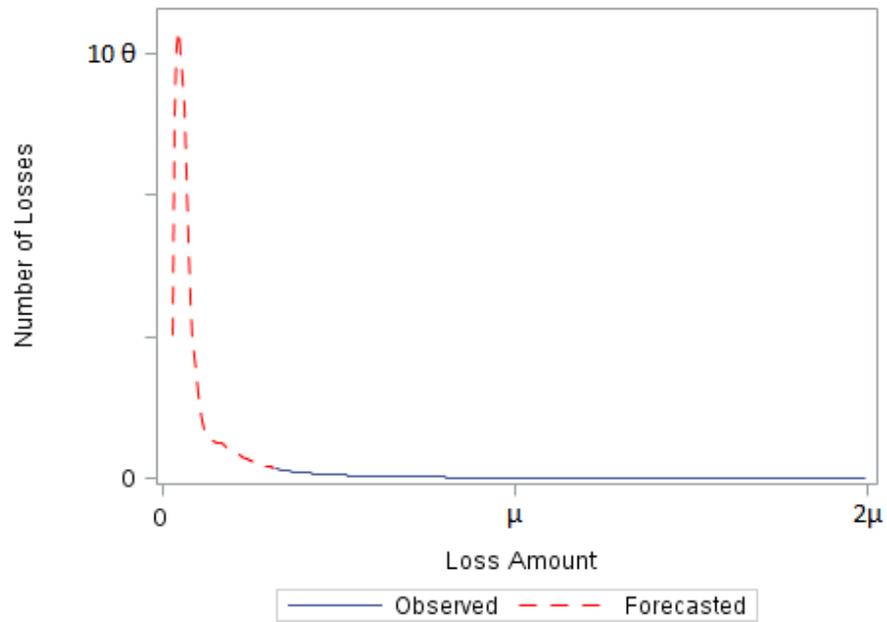


Figure 7.9 The result from the prediction of the continuation of the loss distribution in risk cell 3. The solid blue line is the observed part of the external dataset and the dashed red line is the forecasted continuation of the density function.

Chapter 8

Conclusions & Further Research

Internal data on operational risk losses are limited, and financial institutions are required by Basel Committee on Banking Supervision to use external data when using an Advanced Measurement Approach for operational risk quantification. The Loss Distribution Approach usually models the loss severity using one distribution for the body and another distribution for the tail, introducing problems in determining the breaking point beyond which the tail domain is defined. Bolancé et al. (2012) address this issue when presenting the semiparametric approach, enabling the use of a single distribution over the entire loss domain. However, this methodology requires a full loss distribution and external data are typically left truncated where only large losses are visible. To address the problem with truncated external datasets, this thesis has predicted the continuation of a probability density function, in order to recreate the unobserved body of a loss distribution.

Using Best Linear Unbiased Prediction, the truncated body of an external loss distribution is recreated, based on information from its tail and probability density functions from a set of internally collected losses, and the purpose of the thesis is reached. By demonstrating the results from implementing the Best Linear Unbiased Predictor on operational risk loss data, this thesis illustrates the viability and utility of the proposed method, and provides risk managers with a methodology to address one of the many challenges with operational risk modeling. Predicting the continuation of a function, using the methodology developed by Goldberg et al. (2014), can be successfully achieved on a probability density function of external losses, which our study shows. Furthermore, the distribution is recreated based on an internal dataset and without introducing parametrical assumptions, making this research unique and pioneering. The findings of this thesis will be beneficial for financial institutions as it enables risk managers to employ the semiparametric approach introduced by Bolancé et al. (2012) and thereby eliminates the problem in defining the breaking point between the tail and the body when employing the loss distribution approach. The study generates an enriched external dataset and in addition, confirms the need to categorize operational risk losses into intersections between business line and event type when recreating truncated data, as the loss distribution differs

significantly between risk cells. Additionally, the confidence bands for the prediction indicate a precise prediction in almost the entire loss domain. The confidence bands are wider near the mode of the distribution as a result of the characteristics of the distribution, which experiences a large peak in this area. Following this study, practitioners within risk management are able to incorporate both internal and external operational risk loss data when employing the approach proposed by Bolancé et al. (2012) and furthermore avoid the difficulties associated with the loss distribution approach.

Worth mentioning is that the approach proposed by this thesis is viable under the assumption that the external loss distribution experiences a similar shape as the internal distribution, and that the reporting threshold of the external loss data is known. The shape of the internal loss dataset illustrates the bank's historical losses and since the objective when using external loss data is to complement the oftentimes limited internal loss dataset, is it desirable that the two datasets experience similar characteristics. When using this methodology, the external loss dataset is recreated without any parametrical assumptions using the information from its own tail and the internally collected losses. This is a unique approach where all available data is being used without the need to divide the external loss distribution into a body and a tail domain, avoiding the uncertainty in determining the breaking point between the two. As a final point, the use of the internally collected losses when recreating the truncated part of the external database tailors the result to the financial institution making the approach more suitable to use in the modeling of operational risk.

8.1 Further Research

The prediction model proposed in this thesis is based on the assumption that internal losses follow the same distribution as external losses under the condition that external losses are truncated below a threshold value. An assessment of this assumption is a recommended following step, along with an evaluation of the possible errors this assumption would induce if not valid.

As this thesis generates a full dataset of external operational risk losses, the natural subsequent step is to assess how to merge the dataset with the internally collected losses without introducing any bias caused by the difference in sample size.

A final suggestion for further research in the field of operational risk is to develop a method to fill the empty cells in the correlation matrix between business line and event type. There exists dependency between the risk cells, and as our research shows, the loss distribution varies significantly in the different cells confirming the necessity to categorize the losses. This thesis has employed prediction in the cells with many observations as the semiparametric approach may yield flawed results when the data are scarce. By exploring and modeling the dependency, it may be possible to enrich the datasets in the cells with fewer observations.

Chapter 9

Bibliography

Aue, F. & Kalkbrener, M., 2006. LDA at work, Deutsche Bank's approach to quantifying Operational Risk. *Journal of Operational Risk*, 1(4), pp. 49-93.

Aue, F. & Kalkbrener, M., 2007. *LDA at Work*, Frankfurt: Deutsche Bank.

Basel Committee on Banking Supervision, 2006. *International Convergence of Capital Measurement and Capital Standards*, Basel: Bank for International Settlements.

Baud, N., Frachot, A. & Roncalli, T., 2002a. *Internal data, external data and consortium data for operational risk measurement: How to pool data properly?*, s.l.: Groupe de Recherche Opérationnelle, Crédit Lyonnais.

Baud, N., Frachot, A. & Roncalli, T., 2002b. *How to Avoid Over-estimating Capital Charge for Operational Risk?*, Lyon: Groupe de Recherche Opérationnelle, Crédit Lyonnais.

Bittermann, J., 2008. *Report: Trader had drawn red flags*. [Online]

Available at:

<http://edition.cnn.com/2008/WORLD/europe/01/29/rogue.trader/>

Bolancé, C., Guillén, M., Gustafsson, J. & Perch Nielsen, J., 2012. *Quantitative Operational Risk Models*. London: Chapman & Hall.

Champernowne, D. G., 1952. The Graduation of Income Distribution. *Econometrica*, 20(4), pp. 591-615.

Chernobai, A., Menn, C., Trück, S. & Rachev, S. T., 2004. *A note on the estimation of the frequency and severity distribution of operational losses*, : Applied Probability Trust.

Davis, E., 2011. *Truncated or Shifted: US debates loss data collection threshold in op risk modelling*, : Operational Risk & Regulation.

De Boor, C., 1978. *A practical guide to splines*. New York: Springer.

Dutta, K. & Perry, J., 2007. *A Tale of Tails: An Empirical Analysis of Loss Distribution Models for Estimating Operational Risk Capital*, Boston: Federal Reserve Bank of Boston.

Ergashev, B., Pavlikov, K., Uryasev, S. & Sekeris, E., 2014. *Estimation of Truncated Data Samples in Operational Risk Modeling*, Washington D.C: The Office of the Comptroller of the Currency.

European Central Bank, 2015. *www.ecb.europa.eu*. [Online]
Available at:
<https://www.ecb.europa.eu/stats/prices/hicp/html/inflation.en.html>
[Accessed 15 April 2015].

European Central Bank, 2015. *www.ecb.europa.eu*. [Online]
Available at:
<https://www.ecb.europa.eu/stats/exchange/eurofxref/html/index.en.html>
[Accessed 15 April 2015].

Frachot, A., Georges, P. & Roncalli, T., 2001. *Loss Distribution Approach for Operational Risk*, Lyon: Groupe de Recherche Opérationnelle, Crédit Lyonnais.

Goldberg, Y., Ritov, Y. & Mandelbaum, A., 2014. Predicting the continuation of a function with applications to call center data. *Journal of Statistical Planning and Inference*, Volume 147, pp. 53-65.

Hult, H., Lindskog, F., Hammarlid, O. & Rehn, C. J., 2012. *Risk and Portfolio Analysis*. New York: Springer .

Höllig, K. & Hörner, J., 2015. *Approximation and Modeling with B-splines*. s.l.:SIAM.

Jones, M. C., 1993. Simple boundary correction for kernel density estimation. *Statistics and Computing*, Volume 3, pp. 135-146.

Luo, X., Shevchenko, P. V. & Donnelly, J. B., 2007. Addressing the Impact of Data Truncation and Parameter Uncertainty on Operational Risk Estimates. *The Journal of Operational Risk*, 2(4), pp. 3-26.

Pirouz, M. & Salahi, M., 2013. Modeling Truncated Loss Data of Operational Risk in E-Banking. *I.J. Information Technology and Computer Science*, 12(), pp. 64-69.

Ramsay, J. O. & Silverman, B. W., 2005. *Functional Data Analysis*. 2nd ed. New York: Springer.

Shevchenko, P. V. & Temnov, G., 2009. Modeling operational risk data reported above a time-varying threshold. *Journal of Operational Risk*, 4(2), pp. 19-42.

Silverman, B. W., 1986. *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Stevenson, R., 1995. *The Collapse of Barings: The Overview; Young Trader's \$29 Billion Bet Brings Down a Venerable Firm*. [Online]
Available at: <http://www.nytimes.com/1995/02/28/us/collapse-barings-overview-young-trader-s-29-billion-bet-brings-down-venerable.html>

Zhou, X., Giacometti, R., Fabozzi, F. J. & Tucker, A. H., 2013. Bayesian estimation of truncated data with application to operational risk measurement. *Quantitative Finance*, 14(5), pp. 863-888.

Chapter 10

Appendix

10.1 Detailed Figures

Here are enlarged and more detailed figures presented from section 7.2.

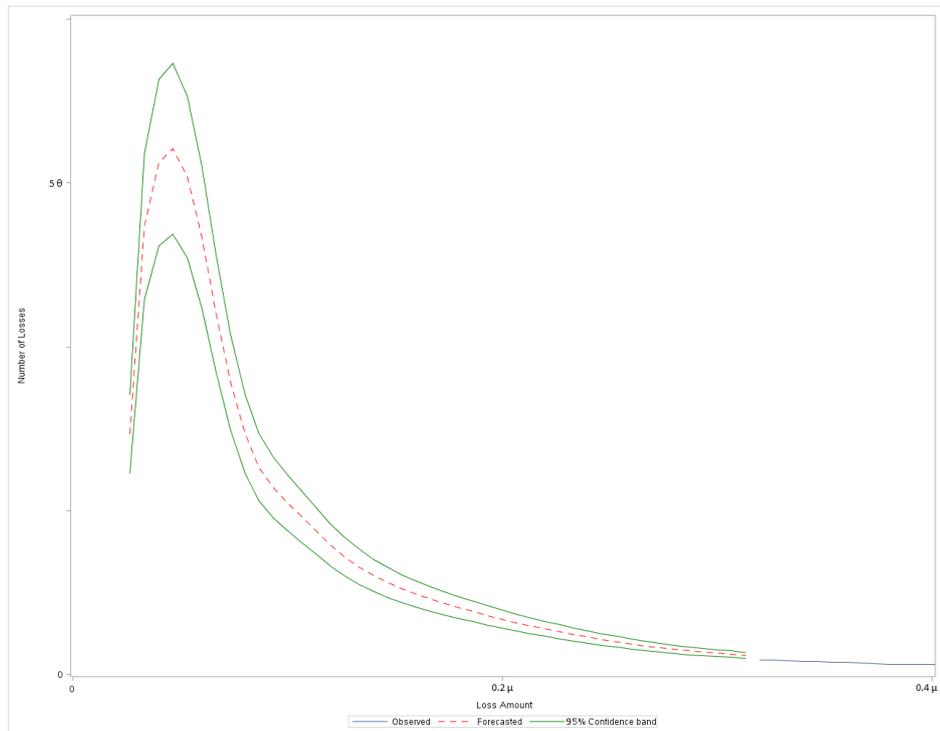


Figure 10.1 A more detailed version of figure 7.4

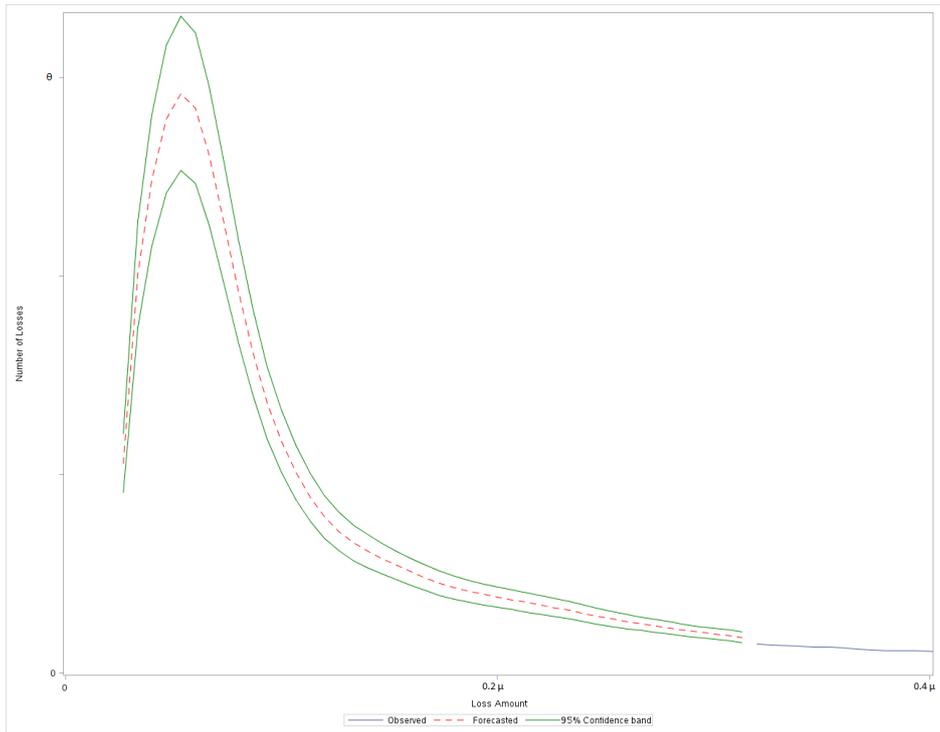


Figure 10.2 A more detailed version of figure 7.5

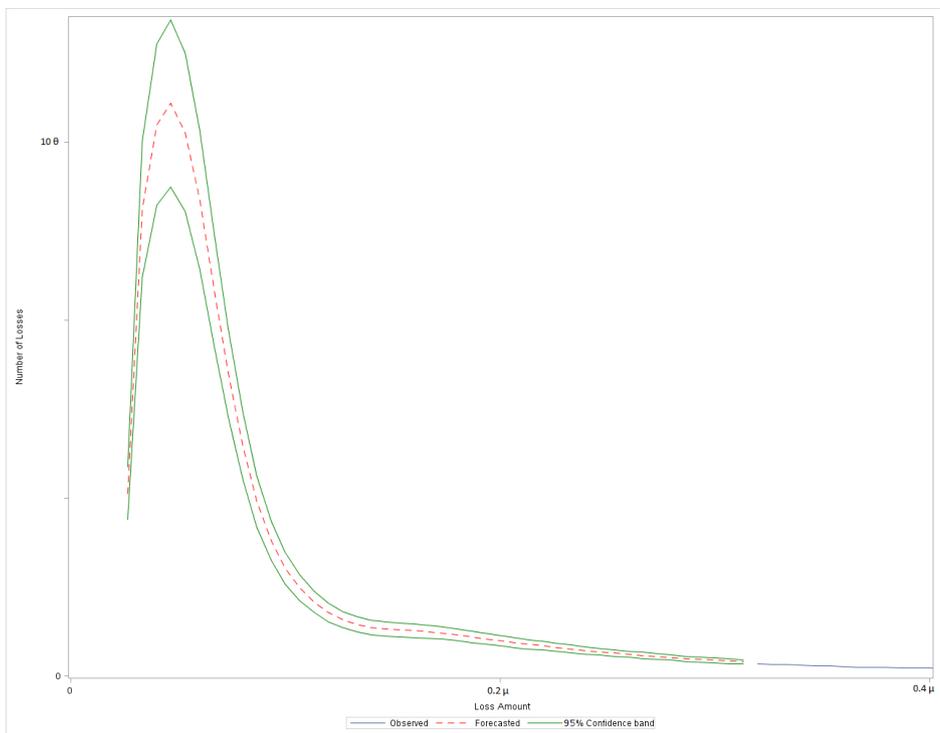


Figure 10.3 A more detailed version of figure 7.6

