Extended Analysis of Back Testing Framework for Value at Risk

Internship Rabobank International

author:	G.J. van Roekel
study:	Industrial Engineering and Management track Financial Engineering
department:	Finance & Accounting
supervisors:	Emad Imreizeeq MSc. (University of Twente) dr. Berend Roorda (University of Twente) dr. Viktor Tchistiakov (Rabobank International) Wei Lan MSc. (Rabobank International)
date:	August 2008

EXECUTIVE SUMMARY

This report is the result of the 'Extended Model Analysis' project at Rabobank International. The Value-at-Risk model and the back testing procedure are important parts of the banks market risk framework. The Value-at-Risk model provides a daily measure that describes a limit which the bank's trading portfolio daily loss is expected to exceed once every 100 days.

DNB and Rabobank agreed to perform a periodic analysis of the VaR model that goes beyond the regulatory guidelines. Every quarter Rabobank International tests the accuracy of its VaR model using the regulatory back test. This back test checks the number of times the VaR was breached (called exception). Based on this number of exceptions this test judges if the VaR model is accurate or not. The regulatory back test has its limitations.

Therefore, we conducted a literature research to investigate alternative back test methods. This resulted in a framework of five back tests that together test the most important properties of a VaR model:

- exception frequency: the number of realised exceptions
- exception clustering: independency of exceptions over the tested period.
- exception size: the size of the exception

We implemented the five back tests in a test framework that Rabobank International can use for the periodic back testing beyond regulation.

<<deleted, confidential>>

TABLE OF CONTENTS

1 INTRODUCTION

Rabobank International uses the Value at Risk (VaR) model to determine and control the exposure of the bank to market risk. De Nederlandsche Bank (DNB) expects Rabobank to use this model and to test its accuracy. Rabobank International agreed with DNB to perform periodical tests that move beyond strict regulatory guidelines. Next to that, the subprime crisis has led to very volatile financial markets. It is important for Rabobank International to know if their current model is good enough to represent this extraordinary market situation. During the internship I worked within Rabobank International to test the accuracy of the risk model.

The above description is technical and contains terms that need additional explanation. The next sections of the introduction provide an overview of the context of the project. It starts at a high level with a general description of financial risk management. Thereafter it explains market risk and the VaR framework for measuring market risk. It ends with the explanation of the tests to measure the accuracy of VaR.

1.1 Financial Risk Management

Risk plays an important role in life and especially in business activities. A general definition of risk is *"the possibility of something bad happening"*. The next step is to narrow the scope of the definition to the risk that organisations are exposed to. This risk can be split into two types: business risks and non-business risks. Business risks are risks that an organisation is willing to take for the creation of a competitive advantage and add value for shareholders. Non-business risks are risks that are not directly related to the core business.

For a financial firm a main part of business risk is financial risk. This is the risk that relates to possible losses in financial markets. Financial markets are the places for trading diverse financial products like FX, equity and credit spreads. In order to cope with this risk, financial firms have extensive and intensive accurate risk measurement and risk management.

1.2 Market Risk

1.2.1 Definition

If we look into more detail into financial risk, we can divide it in the following risk categories:

- market risk: risk that arises from movements in the level or volatility in market prices
- credit risk: risk caused by the fact that companies may be unwilling or unable to fulfil their contractual obligations
- liquidity risk: risk that the bank cannot do a trade because the size of the trade is too large relative to the market size (asset liquidity risk) or risk that the bank has not enough cash available to fulfil payments obligations (funding liquidity risk)

The purpose of the project is to test the used model for market risk. So we look in more detail at this specific category of financial risk. Rabobank International trades lots of different financial products. The bank categorises these products in trading books.

Changes in market prices influence the value of financial products and create market risk for the holder of the products. Market prices are for example interest rates, exchange rates, equity prices and commodity prices. It is important for a bank to know to what amount of market risk it is exposed to, because it wants to control the risks that the bank's traders take in their activities.

1.2.2 VaR history

The risk model at Rabobank International uses Value at Risk (VaR) as a measure for market risk. In this section we give a description of how Value at Risk over the years has become the leading market risk measure.

In the past, several techniques were used to measure market risk. These measurements all had clear limitations and were not capable of giving a single good indication of how big the bank's risk exposure was.

The pioneers in Value at Risk were working at J.P. Morgan during the late 1980's. They worked on a model that summarised the market risk of a certain portfolio in a single number. They used it several years internally before they made it available for free as RiskMetrics™, which used the VaR as a measure of market risk. The tool was adopted by many organisations, also because the Group of Thirty recognised VaR as a best practice. The Group of Thirty represents the largest banks in the world. Initially, their main goal was to develop a framework for best practices to deal with derivatives. A rapid growth was visible in the market for derivatives. A derivative is *a security whose price is dependent upon or derived from one or more underlying assets* (Investopedia, 2008). Well-known types of derivatives are options, swaps, futures and forwards. These products are traded very easily and provide the opportunity to take potentially large positions in the underlying assets with an investment that is relatively small. The value of the underlying assets is affected by changes in market prices. This value change can be very high in comparison to the investment in the derivative. So, the exposure to market risk can be quite high for derivatives. The Group's investigation appointed Value-at-Risk as the best method for measuring market risk.

Regulators of the financial industry were also investigating a uniform (market) risk management framework. They were thinking how to oblige banks and other financial institutions to carry enough capital to provide cover for incidental large losses.

A first step that has been taken by regulators, to unify rules and policy, is the development of the Basel Capital Accord (BCBS, 1988) by the Bank for International Settlements. The goal of this document was to define a standard for minimal capital requirements that banks should hold to cover their risk exposure. One of the limitations of this document was that it contained few guidelines or rules for market risk. This was solved by the Committee with a so-called Amendment to the 1988 Accord (BCBS, 1996). This document obliged the banks to either use a standard model or an internal model to determine the minimum amount of capital that has to be reserved to cover market risk. The internal model produces a VaR as measure of market risk. Because most banks adopted internal models and the regulators prescribed its use in that case, VaR has become the industry standard for measuring market risk. Rabobank International also implemented an internal VaR model.

1.2.3 Definition VaR

VaR is a value that summarizes the worst loss over a target horizon with a given level of confidence. VaR does not indicate how large the expected loss will be under the worst case scenario. Instead, it is a border value that will be crossed no more than a given number of times, depending on the confidence level.

By means of an example, we will try to explain this. The daily return of this portfolio is assumed to be normally distributed with mean 10,000 and standard deviation of 12,000.

The graph of the distribution and the VaR measure are shown in Figure 1.1.

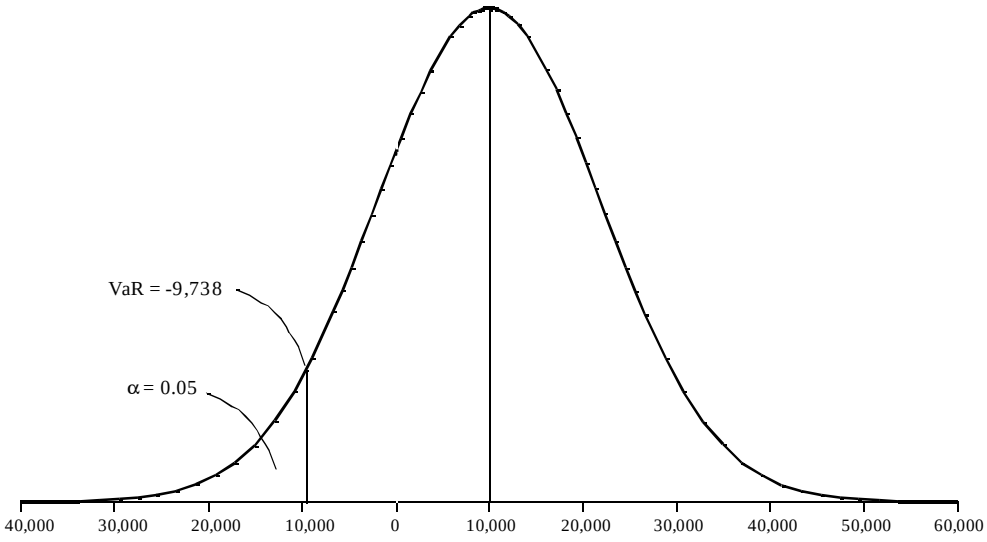


Figure 1.1 VaR portfolio

The figure shows the probability density function of the daily return of the portfolio. The figure indicates the surface corresponding with the lower 5% of this distribution by α. The VaR measure that is the border of this surface indicates that in 5% of all realised returns, the realised return will be smaller than or equal to -9,738. This value is the 95% confidence level VaR measure.

1.3 Testing VaR

We described in the introduction that DNB requires Rabobank International to test its VaR model. This section describes why DNB, in its function as a regulator, wants the banks to test the VaR model and it introduces the method that is used to perform this test.

Although many organisations and institutions accept and use VaR as a market risk measure, it does have its limitations (Jorion, 2007):

- VaR does not describe the worst loss, but represents a value that will be exceeded in a certain number of trading days over a given period (5 of 100 days in case of a 95% VaR confidence level).
- VaR does not give any information about the distribution of the losses in the lower tail. So if the VaR is exceeded only with a small amount or with a very large amount will not be shown.

Due to these limitations, DNB is very interested to see if the results of a VaR model are accurate. This has become even more important since the subprime crisis has had a large influence on the performance of banks. Several banks in the United States and Europe suffered large losses during the subprime period and it is very interesting for the central banks to see if the VaR models worked well in these circumstances.

The technique that the regulator adopted for determining a VaR model's accuracy is back testing. We showed that a VaR model provides a loss limit that will be crossed a number of times in a certain time range. Regulators designed the back testing method to check if the number of realised exceptions (breaches of the VaR) over a given period of time is statistically not significantly different from the expected number of exceptions. For example, if the number of exceptions is too high, this can be reason for a regulator to ask a bank to provide additional insight in the model or force it to develop an improved model.

Next to this evaluation function, back testing also has implications for the capital requirement the regulator sets for the bank. The capital requirement is an amount of money a bank has to hold to cover its risk positions. The use

of this capital requirement should make sure a bank will almost never go into default in case it suffers any large losses due to market risk.

2 PROJECT DEFINITION

In this chapter we define exactly what the project is about and which steps will be taken to reach the goals set. The first section discusses the project objectives. Next we determine the scope. To reach the objectives, we develop a research model and research questions. The final section provides an overview of the structure of the thesis.

2.1 Project Objectives

The following statement represents the project objectives that we want to reach:

The main objectives of the research are to develop a tool for thorough back testing, to use the tool to analyse the performance of the VaR model and provide recommendations for improvements of the VaR model.

By fulfilling this objective we investigate the main problems mentioned in the introduction. First of all, the project provides Rabobank International with a tool that they can use periodically to report to DNB a thorough risk analysis and to gain additional insight in the performance of the VaR model. Next to that, the study provides an analysis of the current VaR model given the results of the back tests that we perform with the tool. We perform the back tests in such a way that we can analyse the influence of the subprime crisis. Finally, we make an initial investigation of alternative VaR models.

2.2 Scope

We perform the back tests on a number of trading books. Experts of Rabobank International make this selection such that it is a good representation of all major risk categories. The results of the back test provide insight in the performance of the current VaR model. Based on this, we investigate improvements for the VaR model and perform initial research into VaR model alternatives. A complete redesign of the VaR model is not within the scope of this project.

2.3 Project Structure

We reach the objectives by following a systematic and structured research flowchart. The flowchart, which forms the basic structure of the project, is shown in Figure 2.1.

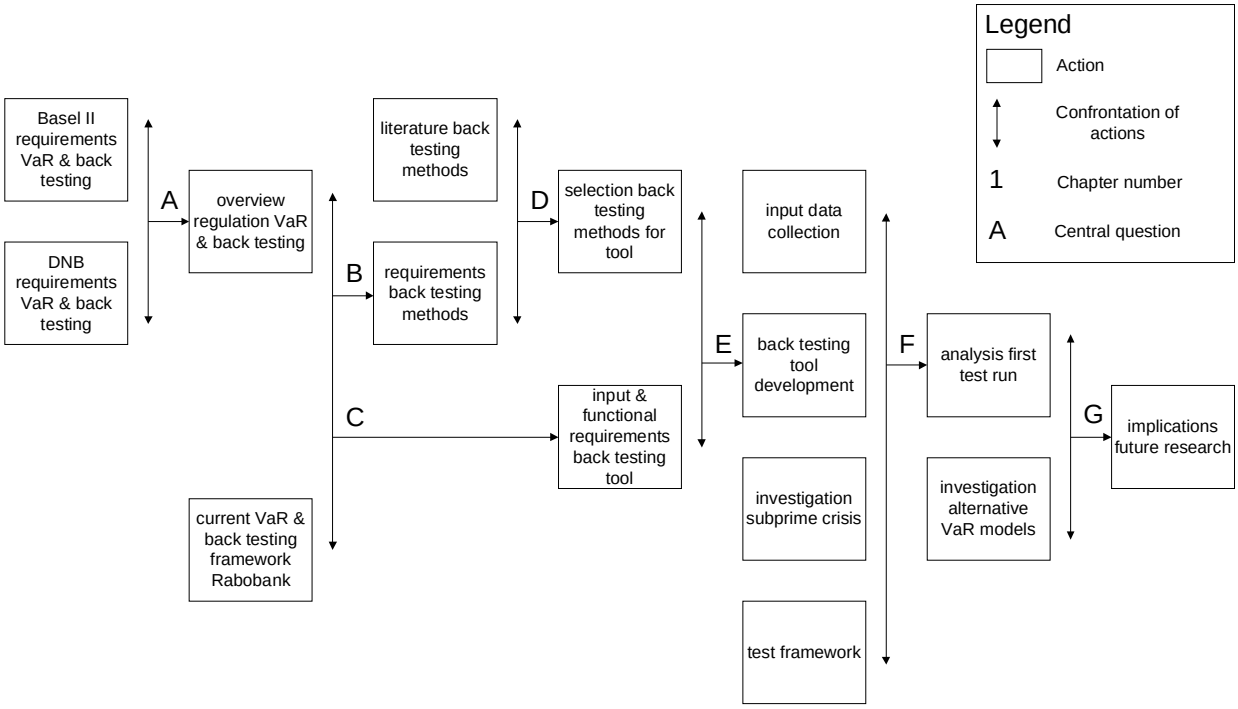


Figure 2.1 Research Flowchart

We specify this research flowchart by developing seven central research questions. Each specific research question connects to the corresponding letter in the research flowchart. The numbers above the action blocks give the chapter in which this issue is discussed.

A. What are the regulatory requirements for back testing and VaR? (Ch. 3 & 4)
B. What are the requirements for suitable back testing methods? (Ch. 5)
C. What are the input and functional requirement for the new back testing tool? (Ch. 5)
D. Which back testing methods are suitable for implementation in the new tool? (Ch. 6, 7, 8 & 9)
E. How can we develop the new tool? (Ch. 5)
F. What conclusions can we draw from the initial test run? (Ch. 10)
G. What implications for future research can we make up based on the initial investigation into alternative VaR models? (Ch. 12)

2.4 Thesis Structure

We derive the structure of the thesis in a straightforward manner. First of all, we provide a detailed description of the regulation for VaR and how it is implemented at Rabobank International in Chapter 3. Chapter 4 discusses the regulation for back testing VaR and its implementation. Chapter 5 gives an overview of the back testing framework's requirements. Chapter 6 discusses the exception frequency back tests. We describe the second category of back testing methods, exception clustering, in Chapter 7. Chapter 8 discusses the exceptions size back tests. We describe methods that do not fall in one of the previous categories in Chapter 9. Chapter 10 contains a summary of the conclusions we obtain from the first test run. Chapter 11 provides an overview of the alternative VaR models we investigated. Chapter 12 contains the implications for future research.

3 VALUE AT RISK

In this section we first give an overview of the regulation that relates to VaR. Next we describe the most used methods for determining the VaR.

3.1 Regulation

In the Basel Committee on Banking Supervision different institutions from the international banking world work together to overcome regulatory issues. The cooperation between banks is meant to upgrade the quality of worldwide regulation. This section provides an overview of the regulation, relating to VaR, that the Committee has developed in their most recent regulation document (BCBS, 2006). The numbers between brackets indicate the section number of the document. The document allows banks to use two methodologies to measure their market risk. The first is the implementation of a standard model that the document prescribes. The second choice is to use an internal model. Most banks prefer to use this internal model, since it is a better reflection of diversification that takes place in the bank's portfolio and it leads to a lower capital requirement for market risk (Hull, 2007). In order to use the internal model the bank has to fulfil a set of demands (701(ii)). Among these demands is the requirement to compute the VaR on a daily basis. The levels of VaR that the bank has to measure are the 10-day period 99% percentile and the 1-day period 99% percentile. The bank can use these VaR levels to respectively determine the capital requirement for market risk and obtain the multiplier for this capital requirement.

3.2 VaR methods

The regulator does not prescribe the method to compute the VaR. A bank is free to choose its own method. Many methods exist, but we can disseminate three main types of methods. We describe these in the next subsections.

3.2.1 Variance / Covariance method

J.P. Morgan adopted this method in their RiskMetrics database to compute the VaR. A bank's portfolio exists of a lot of different financial products. Market factors like interest rates, stock prices and currency rates influence the value of each of these financial products. The variance / covariance method uses estimates for the volatilities of the market factors and the correlations between market factors to obtain an estimation of the volatility of the bank's overall portfolio. The RiskMetrics method uses approximations to determine the volatility for more complicated financial products. The method assumes the overall portfolio has a normal distribution with mean zero and the estimated volatility. Finally, one estimates the VaR for the overall portfolio by taking the 99% percentile of this distribution.

3.2.2 Historical Simulation

Historical Simulation uses a set of data from the past to give a prediction of what will happen in the future. To be more specific, in this method we use a history of e.g. 250 hypothetical market factor shocks to determine what the portfolio VaR for tomorrow is. We compute the VaR by taking the 99% percentile of the hypothetical market shocks.

3.2.3 Monte Carlo Simulation

For this VaR method, we assume a distribution for the risk factors. After drawing from the market factor distributions, we can determine the impact on the Profit and Loss. By repeating this simulation millions of times, we are able to simulate millions of possible Profits and Losses. If we take the 99% percentile of this set, we obtain the VaR.

3.3 VaR at Rabobank

<<deleted, confidential>>

4 BACK TESTING VAR

Regulation prescribes that all banks, that use an internal model for the measurement of their market risk exposure, should verify the quality of this model using the so-called back testing procedure. This chapter gives an overview of the precise demands of the regulation of Basel II and DNB in the first and second section. The third section describes how the back testing procedure is applied at Rabobank.

4.1 Basel II Regulation

The first subsection summarises why back testing is necessary. Subsection two contains a description of the back testing framework prescribed by the Committee. Finally, the third subsection describes how banks should interpret the results of the back tests. The numbers between brackets refer to the paragraph numbers of the Basel II regulation document (BCBS, 2006).

4.1.1 Need for Back Testing

If a bank chooses to develop an internal model for market risk, one of the requirements of the Basel II Accord is that they have to implement a back testing procedure (718(LXXIV)(b)). Through this method, the regulators can gain insight in and judge the performance of the internal models used at the banks.

Many methods exist for back testing; no uniform method gives the best results. This is something the Committee has taken into account while developing the regulation. The goal of the back testing procedure is to find a balance between its performance in measuring power and its imperfections (Annex 10a.6).

Back testing is especially important since it is the most important factor in determining the capital requirement that banks have to hold to cover market risk of their trading portfolio. The size of the capital requirement is equal to the higher value of the VaR of the day before and the average of the VaR values of the previous sixty days multiplied by a factor (718(LXXVI)(i)). The multiplication factor S_t has a minimum value of 3 and can be as high as 4, depending on how good or bad the results of the back tests for the actual P&L are (718(LXXVI)(j)). The amount of regulatory capital for market risk can be calculated with the following formula:

$$RC_t = S_t \cdot \max\left(VaR_t(0.01), \frac{1}{60}\sum_{i=0}^{59} VaR_{t-i}(0.01)\right) \qquad (4.1)$$

The VaR measure used here is the 10-day 99% confidence level VaR.

The Basel II document also prescribes additional tests that banks have to perform next to the standard back test (718(XCix)). Firstly, they must demonstrate that all assumptions in the model are appropriate. Examples of assumptions are the use of a normal distribution and the use of the square root of time rule for scaling from a one-day to a ten-day holding period of the VaR. Tests for model validation should go beyond the standard Basel II back test. Specific examples that the regulation gives are:

- Perform back tests with hypothetical changes in portfolio value.
- Perform back tests over a longer look back period.
- Use other confidence intervals than the 99% interval required.
- Test portfolios below the bank level.

The third rule of the regulation concerns the use of hypothetical portfolios to ensure that the model is able to account for particular structural features that may arise. This might, for example, occur when historical data is not complete enough to map the required look back period.

4.1.2 Basel II Back Testing Framework

The Basel Committee developed a standard minimum test that banks should perform in measuring their market risk management system's performance.

Basically, back testing is simply about a periodic comparison of the banks daily VaR measures and the actual trading outcome for that day (Annex 10a.8). The framework requires banks to compute VaR at a confidence level of 99% (Annex 10a.10). Given this confidence level we expect that once every 100 days the actual trading loss is larger than the VaR. We call this breach of the VaR an exception. By simply comparing the realised number of exceptions with the expected number we can draw conclusions upon the performance of the VaR model.

An important limitation of the described back testing method is the fact that it uses the actual trading result of a day in the comparison with the VaR estimate. This assumes that the only changes that take place during the day are due to price and rate movements. This does not happen, since portfolios change during the day. So the actual trading results include fee income and trading gains and losses. These values contaminate the back test results (Annex 10a.12). Because of this reason the framework uses VaR with a one-day holding period. A ten-day holding period would include even more trading events and portfolio changes.

The framework suggests some solutions that might (partially) solve the contamination problem. The first one is eliminating the contamination by carefully identifying the contaminating values and leaving them out of the back test. The second solution consists of using hypothetical instead of actual trading results. The bank computes hypothetical results under the assumption that during a trading day the positions in the portfolio do not change. By using this method, all changes in portfolio value happen due to changes in market factors like interest rates.

Regulation requires banks to perform the back test quarterly using at least a year of trading data (Annex 10a.22).

A limitation of the back tests formulated by the Committee is that they cannot distinguish accurate and inaccurate models extremely well (Annex 10a.26). On the other hand it is very easy to implement and perform.

4.1.3 Interpretation Results

Now that the back testing method is clear, we address how the results should be interpreted according to the Basel II Accord (Annex 10a.27-59).

Since the back test does have its limitations, one cannot implement very strict rules for judging the model. Otherwise the probability of a type 1 error (rejecting an accurate model) or a type 2 error (accepting an erroneous model) would become too large. Instead, regulation prescribes three result zones: green, yellow and red. The green zone indicates that the number of exceptions generated by the model is acceptable and suggests the model is accurate. The yellow zone will start a discussion of the results. Exceptions might be attributed to multiple causes (Annex 10a.48):

- incorrect model
- instruments' risk is not assessed correctly
- random chance
- unexpected market movements
- large intra-day trading caused loss

In case of the yellow zone, the bank will get a chance to prove that the high number of exceptions has another cause than an inaccurate model, before the regulator raises the multiplication factor. Finally, the red zone involves such a high number of exceptions that the probability of an accurate model is very low. In that case, the regulator penalises the bank with an increase of the multiplication factor to 4 and the requirement to develop an improved model. Table 4.1 contains an overview of the zones, number of exceptions and the size of the multiplication factor.

Zone	# Exceptions	Multiplication factor
Green	0	3.00
Green	1	3.00

Green	2	3.00
Green	3	3.00
Green	4	3.00
Yellow	5	3.40
Yellow	6	3.50
Yellow	7	3.65
Yellow	8	3.75
Yellow	9	3.85
Red	> 10	4.00

Table 4.1 – Basel II Zone Classification

4.2 European and DNB Regulation

The European Union implemented much of the Basel II requirements in law in the Capital Requirements Directive (CRD). This again is transferred in Dutch Law into the Financial Supervision Act (FSA). Finally, DNB transferred the annexes of the CRD into regulation. Of special interest to this paper is regulation that is prescribed by DNB (DNB, 2006).

4.2.1 Regular Validation

DNB wants the banks to validate its internal model both periodically and in special cases. The periodical requirement states that the banks should validate its model at least once a year or in special cases. Special cases involve significant changes to the model or market events that are likely to have a large influence on the model and maybe even make the model inaccurate.

To validate the internal model, banks should use other techniques besides back testing. The DNB requirements oblige the bank to perform at least:

- tests to demonstrate that any assumptions made within the internal model are appropriate and do not under- or overestimate the risk;
- model validation tests in relation to the risks and structure of the financial undertaking's portfolio;
- the usage of hypothetical portfolios to ensure that the internal model is able to account for particular structural features that may arise.

DNB also recognises the contamination problem which we mentioned in subsection 4.1.2, because they require the banks to include both actual and hypothetical profits and losses in the back testing procedure.

The last part of the regulation that is important for back testing has to do with the results of the back test. DNB also allows banks to ask for dispensation from capital requirement increases. But regulation states very clearly that this can happen only under exceptional circumstances. Next to that, if a bank experiences inaccuracies in the VaR model through back testing, DNB should be notified in five days.

4.3 Back Testing at Rabobank International

<<deleted, confidential>>

5 BACK TESTING FRAMEWORK REQUIREMENTS

This chapter gives an overview of the back testing framework that we develop in the project. We divide the methodology part of the project in three major fractions, which are described in the three sections of this chapter. The first part is the back test method research, in which we investigate which back testing methods are most suitable for implementation. The second part is the tool development for which we set up a list of requirements. Finally, the third part is the test framework which describes how we design the first test run.

5.1 Back Test Method Research

To obtain the most suitable back testing methods for the new back testing framework, we perform a literature study. The first two subsections describe how we perform the research and what scope we use. We set up a number of requirements for the back testing methods that we discuss in the third section. Finally, the fourth section describes how we make the selection.

5.1.1 Research Method

In order to provide a decent overview of suitable methods, we use a structured approach. The starting point is the detailed overview of VaR by Jorion (Jorion, 2001). This book dedicates a chapter to back testing. This provides us insight in the more basic tests and types of methods. Next to that it gives references to basic articles on back testing (Kupiec, 1995), (Christoffersen, 1998), (Crnkovic and Drachman, 1997) (Lopez, 1999). The next step we perform is searching the articles citing these authors. We find existing literature reviews on back testing methods (Campbell, 2005), (Haas, 2001), (Blanco and Oks, 2004). Also, we discover an extensive list of articles describing one or more back testing methods. To provide a thorough literature research, we also use two major search engines (Scopus, 2008) (ISI, 2008) with a list of keywords (found in 0). Together, these search engines cover almost 25.000 journals (Scopus, 2008) (ISI, 2008). Finally, we use the web site of GloriaMundi, containing a list of 57 articles on back testing VaR (GloriaMundi, 2008). Our analysis consists of scanning, selecting and summarising back testing methods from this extensive collection of articles. This results in a division of the back testing methods into three different types: exception frequency, exception clustering and exception size. Each type tests a different property of the VaR model. In the Chapters 6, 7 and 8 we discuss each of these types. Finally, Chapter 9 contains methods that do not fall under one of the other three types.

5.1.2 Scope

A distribution back test that compares the realised and hypothetical distribution has more power in detecting an inaccurate model than the methods that only address a quantile of the distribution (Campbell, 2005). But the increased power comes at a cost. Campbell states that a VaR model can excel in describing extreme losses but be less accurate on moderate profits and losses. In that case, we could judge the model as inaccurate, while the model is accurate from a risk management perspective. For this purpose, we are first of all interested in the properties of the distribution's tail. Since the main goal of this project is to find out if the VaR model used at Rabobank accurately models exceptions, the use of distribution forecast tests has limited added value. Next to that, this type of test comes with an informational burden, since we would have to estimate a hypothetical distribution for every trading book. The last reason for not including this type of test is the fact that we will include multiple confidence levels for the VaR in the back testing framework. This already provides insight into the accuracy of a larger part of the P&L distribution.

5.1.3 Requirements

An ideal back testing method does not exist. In order to make a good selection of back testing methods we set up a number of requirements. We investigate for each back testing methods how it performs for each of the requirements.

The requirements we use are goals, power, size and feasibility. The following subsections provide an explanation of each of these four.

Goals

For each method we define what the goal of the test is. The requirement 'Goals' indicates how well a test fulfils this goal.

Power

Statistical tests always are a trade off between the two types of errors that can arise. A good statistical test has both a low type 1 and type 2 error. A type 1 error occurs if the null hypothesis is true, but rejected by the test. The type 2 error occurs if the null hypothesis is false, but accepted by the test.

In case of most back tests the errors are rejecting an accurate model (type 1) and accepting an inaccurate model (type 2). The goal of back testing is to judge a VaR model's accuracy. The power criterion indicates how good the back test is in separating inaccurate and accurate models.

Size

The sample size is the look back period, measured in trading days that we use as input for the back tests. Some of the back tests require large sizes in order to make sure the results of the test are reliable in separating inaccurate and accurate models. This is not very convenient, since we would like to see reliable test results also for short look back periods.

If a test needs a large sample size for accuracy, we give a low score for the size criterion. If it needs only a small sample size, we give it a high score.

Feasibility

The feasibility criterion covers some topics that we cannot measure easily:

- If the added value of the test is large enough to overcome the implementation effort.
- If the back testing method tests a property of the VaR model that is not or partially covered by other methods.

5.1.4 Selection Procedure

The selection of the back testing methods that we consider suitable for implementation in the back testing tool is not a straightforward procedure. The judgement how good a back test performs for each requirement is subjective. Still, we use explicit scores in the selection procedure to indicate the performance, since this provides a much more convenient overview. In order to give insight in the scoring, we provide argumentation for the scoring decision. Table 5.1 shows the scores we use in the selection procedure. We provide the argumentation and the selection in the next three chapters.

Score	Explanation
--	very bad
-	bad
+ / -	moderate
+	good
++	very good

Table 5.1 – Score Range

5.2 Requirements Back Testing Tool

This section describes the requirements that we set up for the back testing tool. The first subsection contains the domain analysis which describes the environment in where expert will use the tool. In the next subsection we determine the input requirements for the tool. In the section after that we discuss the functional requirements, which describe the user settings that have to be available in the tool. Finally, we give an overview of the output the tool has to generate.

5.2.1 Domain analysis

Experts within Rabobank will use the tool for two main purposes. The first one is the reporting Rabobank International has to fulfil to De Nederlandsche Bank (DNB). This has to show how well the VaR model for market risk within Rabobank International performs. The current procedure measures this model performance by back testing. Every quarter DNB requires Rabobank International to deliver a report containing a description of the results of the Basel II back test for individual trading books within Rabobank International and at an overall group level. Besides this quarterly report, Rabobank International promised DNB *to provide regular additional insight in the VaR model's performance by using alternative back testing methods*. Experts can use the tool that we develop in this project to provide this additional insight.

Secondly, experts can use the information that the back testing tool provides to judge the performance of the VaR model and the individual trading books.

5.2.2 Input Requirements

We have to make decisions on what data we will use as input for the tool. In this section we give an overview of the data that we will use in the first test run. We create this overview with a description of the requirements concerning the selection of the trading books and the input data structure.

1. **The input data for the first test run must contain a representative selection of Rabobank International's trading books.**

We perform the initial test run over a number of trading books that is representative for the group level portfolio. In section 1.2 we mentioned four categories of market prices: equity, interest rate, commodity and FX. Rabobank International has also divided its trading books in these four categories. We do not select the trading books ourselves, but a representative selection of the categories has been made by experts of Global Market Risk. The books that they selected have the highest contribution to the group level VaR.

2. **The tool must be able to handle different trading books.**

We want to use the first selection of the trading books to investigate how well the VaR model performs for trading portfolios from different risk categories, especially during the subprime crisis. But Rabobank International should be able to use the tool after the project. So the tool must be able to cope with other trading books as well.

3. **Both hypothetical and actual profits and losses have to be tested.**

We include this requirement because of regulation. Basel II requires a bank to use both hypothetical and actual profits and losses in their back testing procedure.

4. Four types of VaR coverage have to be tested (1-day 95%, 97.5% and 99% & 10-day 99% coverage)

The data that we need for the tool is extracted from a VaR database, risk engines and a control database within Rabobank for each trading book. The VaR coverage levels that we use are the estimated 1-day VaR values at 95%, 97.5% and 99% coverage and the 10-day VaR at 99% coverage.

We need to include the 1-day VaR at 99%, since this is the level that banks have to use for back testing due to regulation.

We include the 97.5% because Rabobank International's internal limit setting framework incorporated this measure. The back testing might be used for internal model control, so inclusion of this measure is appropriate.

We include the 95% because we decided to leave out back testing methods that use distribution forecasts. One argument for this decision was that we would include multiple VaR levels in the tool.

We include the 10-day 99% VaR measure since the bank uses this value for determining the Basel II capital requirement as mentioned in section 4.1.1.

5. The look back period taken into account in the back tests should range from 250-1250 days.

The look back period is the number of days we take into account in the back testing procedure (sample size). We want to test for different periods to see if the length of the period influences the test results.

To determine the input data requirements we need to set an upper limit to the number of days that can be judged by the back testing tool. This limit is set to 1250 trading days, equivalent to 5 years. Since both short and long look back periods have drawbacks, testing scenarios have to include periods ranging from 250 to 1250 days.

6. The initial test data set will include records until April 1st 2008.

By using this end date, we make sure that recent data is tested. The subprime crisis started around August 2007. Using the 1st of April 2008 as end date, the amount of data representing the sub-prime crisis is large enough to test model accuracy during that period.

5.2.3 Functional Requirements

The tests that we will implement in the tool have many different parameters. The tool's end user should have the possibility to set the values of several of these parameters. This makes the tool very flexible and allows for extensive scenario testing. This section describes the options that the end user has in selecting the test parameters.

7. The user should be able to select back tests to be performed individually (optional)

All of the tests we will select are separately selectable. It is not necessary to include all back tests in each test run.

8. The user can select the size of the look back period in the range of 250-1250 days. (optional)

For each test run the user selects a single amount of days. This means that if a user wants to test multiple look back periods of different length, it will be necessary that he performs multiple test runs. This reduces flexibility, but it also makes the implementation simpler. This requirement is optional, since the user can also influence the size of the look back period using the input data.

9. The user can select an end date for the look back period. (optional)

We include this option so that the user is able to select look back periods that end before April 1st. This is especially useful to test different scenarios in- and excluding the sub-prime crisis. This requirement is optional, since the user can change the end date by altering the input data.

10. The user has the option to exclude actual or hypothetical profits and losses. (optional)

Although regulation requires the inclusion of both P&L types, the tool will leave the choice of excluding one of them to the user. For internal purposes it might be more appropriate to use only one of the types.

5.2.4 Output Requirements

The tool will provide an overview of the results. The following requirements indicate what outputs the tool should provide and why.

11. The outputs should contain summary statistics of the selected P&L types and VaR coverage levels.

The statistics can provide quick insight into the size, volatility and distribution of the P&Ls and VaR levels.

12. The selected P&L types and VaR coverage levels have to be represented in graphs.

A graph in which the profits and losses and VaR are set out over the look back period can provide quick insight into the development of these values over time.

13. The output overview should provide the number of exceptions.

The back testing tool is all about the exceptions, so the number of exceptions that occurred should be part of the output. We make a distinction between exceptions larger than €50,000 and smaller than €50,000. Rabobank International also uses this limit in the current back testing tool. The tool uses all exceptions to measure the accuracy of the VaR model, but experts investigate the exceptions larger than €50,000 more thoroughly and they provide argumentation to explain why it occurred.

14. The tool should present the results of all tests using a zone classification.

No matter what tests we select, the tool should present the test results in such a way that the user can see immediately how the test classified the model's performance. The Basel II zone classification in a green, yellow and red zone is very convenient for this. If applicable, we will use the same kind of zone classification for the tests that we select for implementation.

5.3 Test Framework

The research objective states that the tool developed should be *"used to analyse the performance of the VaR model"*. Given this research goal, we set up a test framework to judge the model's performance. The first section describes the test procedure that we followed. The second section describes how we interpret the results of tests.

5.3.1 Test procedure

The input data set for the first test run contains the P&L and VaR vectors for seven trading books. There is not enough trading book data available to test all look back periods ranging from 250 to 1250 data points. Table 5.2 gives an overview which look back periods we test for the books.

	excluding subprime					including subprime				
book	250	500	750	1000	1250	250	500	750	1000	1250
1										
2										
3										
4										
5										
6										
7										

Table 5.2 – Look back periods

Next to the different look back period lengths, we make a distinction between periods including and excluding the subprime crisis period. The period "excluding subprime" ends at 1st July, 2007 and the test period "including subprime" ends at 1st April, 2008.

We test the trading books for four VaR levels (99% 1-day, 97.5% 1-day, 95% 1-day and 99% 10-day).

We test the trading books for the actual and hypothetical P&L.

5.3.2 Results analysis

In order to test more specifically how the VaR model performs, we set up several detailed analyses. Although the individual book results will be available from the output of the tool, we do not discuss the performance of the individual books. Since the first goal of the thesis is to judge the VaR model's performance, the individual book results are out of the scope. Nevertheless, these results can be important for internal purposes within Rabobank International, so we still include these in the tool's output.

The first analysis that we conduct for each implemented back test simply creates an overview of all observations.

As tests are run on a wide variety of portfolios and VaR percentiles, the graphs in which we present the results do not show individual outcomes. Instead, they present the percentages of the observations that fall in a particular zone.

The number of observations that we present in the results depends on the number of parameters that is taken into account. We provide the number of observations for each test in the graph, since the number of observations is not equal for all tests. For example, each book has a different maximum look back period (see Table 5.2), so the comparison amongst look back periods will show a different number of observations for each look back period.

We assume that Rabobank International's VaR model is correct, so we can compare the percentage of realised results in the different zones with the expected percentage. For each back test we add a table that describes the zone classification in the section that gives the results overview.

We perform additional analyses to test more detailed factors that can influence the performance of the VaR model. The next sections describe these additional analyses.

5.3.3 Difference between actual and hypothetical P&L

In the tests we use both actual P&L's and hypothetical P&L's to check the VaR model's performance. The two P&L types have a very different interpretation. Hypothetical P&L is based on the same market data, position data and pricing models as the VaR computations. So the back testing results of the hypothetical P&L explicitly show how good the model used for VaR calculation is.

The actual P&L is influenced by portfolio changes during a trading day. Back testing provides insight in the accuracy of position data, market data and pricing models combined.

Due to the large differences in interpretation, each of the additional analyses is split into actual and hypothetical P&L results.

5.3.4 Influence of the subprime period

We make a comparison between the test results for the period preceding July 2007 (excluding the subprime crisis) and the period preceding April 2008 (including the subprime crisis).

In the tool requirements we mentioned that we include a zone classification for all tests we select. We compare results from both periods with the same parameters (look back period and VaR level). We do this by checking the zone in which the results fall and scoring the difference in zones. For example, if the result excluding subprime falls into the green zone and the result including subprime is in the red zone the comparison score is 2. Table 5.3 represents the scores that we attached to each difference in zone classifications.

zone classification (excl. subprime)	zone classification (incl. subprime)	comparison score

green	green	0
green	yellow	1
green	red	2
yellow	green	-1
yellow	yellow	0
yellow	red	1
red	green	-2
red	yellow	-1
red	red	0

Table 5.3 – Scoring procedure subprime comparison

5.3.5 VaR percentage level influence on results

We made the decision to exclude back tests that test the distribution of the underlying P&L distribution. To compensate for this exclusion, we include different VaR levels in the back testing tool. By including 99%, 97.5% and 95% VaR levels, we can test a larger part of the tail of the VaR model. If the results are very different for each VaR level, this might indicate a weak VaR model.

We take all the model results of the individual trading books together and then split according to the four VaR levels. Recall that we will give the number of observations for each VaR level in the result charts. Since every VaR level has an equal number of model results, we graph the test results in a stack diagram summing to a total of 100%.

5.3.6 Look back length influence on results

High confidence level VaR models like the 99% model generate only few exceptions. We expect that a larger look back period will generate more reliable test results. So, we are interested to see if the length of the look back periods influences the test results. That is why we compare several look back period lengths.

As we mentioned before, not all trading books have enough historical data available to include them in all look back periods. As a result of this, the number of observations we include in the results is not equal for all look back periods. But we still want to compare the results of the different periods. In order to do this we present the model results in a stack diagram with percentages. For the look back length comparison we also mention the number of included observations in the result charts.

6 EXCEPTION FREQUENCY TESTS

The most basic type of back testing checks the unconditional coverage or frequency property of the VaR models (Campbell, 2005). This type of test considers the frequency of exceptions that was realised during a period and compares this with the number of exceptions that one would expect given the confidence level of the VaR model. This type of test only considers the number of exceptions. It does not make a difference how the exceptions are divided over time or how large the exceptions are.

For example, if a VaR model has a confidence level of 99% and we consider a period of 250 days, one would expect that during this period, in 2.5 days the model would realise a loss that is larger than the VaR. If the realised frequency of exceptions is six, exception frequency tests will analyse if an inaccurate VaR model caused that number.

This chapter gives an overview of the exception frequency tests that we discuss in this project. The first section describes the details about the methods we encountered during the literature research. The second section contains the selection of the methods that we implement in the tool. The third section describes what choices we made in the implementation of the tests. Finally, the fourth section gives an overview of the results of the initial test run.

6.1 Test Descriptions

In this section we describe the exception frequency tests we investigated in the literature research. For each back test we first give a general description of the test. The goal is one of the requirements that we use in the test selection. Next, we indicate what the underlying distribution of the exceptions is. After that we indicate what the test measurements (e.g. hypotheses) are. Power is the requirement that indicates how large the test's statistical errors are. Large errors mean low power of the test. Finally, we describe what the influence of the length of the look back period on the test results is.

6.1.1 Basel II Back Test

Description

The Basel II back test is the most commonly used back test. The regulator requires banks to use this method with a look back period of 250 days. For each of these days the bank has calculated a VaR and P&L. With the 99% confidence interval level that this test uses, the expected number of exceptions is 2.5 during the look back period of 250 days.

Goal

This test's goal is to find out if the VaR model is accurate by testing the number of exceptions that is generated. It regards a model as accurate if the realised number of exceptions is not significantly larger than the expected number of exceptions.

Exception Distribution

Since the test uses a 99% confidence level, the probability of an exception occurring during a given day is 1%. The P&Ls are assumed to be independent. So we can see the occurrence of exceptions at a given day as a Bernoulli experiment with a binomial distribution. We consider a model accurate if it generates an exception on 1% of the trading days.

We can calculate the probability that an accurate model generates k exceptions in $n = 250$ days using the properties of a binomial distribution:

$$P(N = k) = \binom{250}{k} p^k (1-p)^{250-k} \; for \; k = 0,..,250 \qquad (6.1)$$

Table 6.1 shows the probabilities for the accurate model.

# Exceptions (k) (n=250)	P(N=k)
0	8.11%
1	20.47%
2	25.74%
3	21.49%
4	13.41%
5	6.66%
6	2.75%
7	0.97%
8	0.30%
9	0.08%
10	0.02%

Table 6.1 – Exception probabilities accurate model

Test measurements

The input for the test consists of the number of exceptions that a model realised in 250 trading days. Based upon that result the test classifies the model in one of the three zones mentioned in Table 4.1.

Given the binomial distribution, we compute the expected number of exceptions during 250 days as:

$$E(X) = np = 250 \times 0.01 = 2.5 \tag{6.2}$$

Regulation formulates the null hypothesis of the test as *'the actual probability of an exception occurring is equal to the expected probability of 0.01'*. Or, if formulated in terms of probability:

$$H_0: p = 0.01 \tag{6.3}$$

The alternative hypothesis is *'the actual probability of an exception occurring is significantly higher than the expected probability of 0.01'*. This can put in a formula as:

$$H_1: p > 0.01 \tag{6.4}$$

Power

The Basel II back test has a small type 1 error. Recall that a type 1 error is rejecting the null hypothesis while it is true. So, a small type 1 error means that an accurate model is judged as inaccurate. Basel II rejects a model if it falls in the red zone. The red zone starts at 10 exceptions. The probability that an accurate model generates 10 or more exceptions is 0.03 %. This is the size of the type 1 error for the Basel II test.

The Type 2 error concerns the acceptance of an erroneous model. This type 2 error corresponds to accepting the null hypothesis, while the alternative hypothesis is true. The Basel II back test suffers from type 2 errors. For example, suppose we have an inaccurate model with a 98% confidence level VaR or $p = 0.02$. This means that we have an expected number of exceptions generated in 250 trading days equal to:

$$E(X) = np = 250 \times 0.02 = 5 \tag{6.5}$$

Since the Basel II back test checks if the model generates a 99% confidence level VaR, it should reject the null hypothesis, because the model has a probability p of 0.02 instead of the 0.01 under the null hypothesis. Table 6.2 shows what the probability of k exceptions is, given that the tested model is inaccurate (having p values of 0.02 and 0.03 respectively). From the table we conclude that for the model with p of 0.02 the probability of an actual number of exceptions of four is 17.65% ($P(k = 4)$). If we sum all the probabilities of the green zone, we can conclude that in 43.87% ($P(k < 5)$) of all cases for the 0.02 and in 12.82%($P(k < 5)$) for the p of 0.03 models, the test accepts the null hypothesis while it should reject it (type 2 error).

# Exceptions (k) (n=250)	p = 0.02 P(N=k)	p = 0.03 P(N=k)
0	0.64%	0.05%
1	3.27%	0.38%
2	8.30%	1.47%
3	14.01%	3.75%
4	17.65%	7.17%
5	17.72%	10.91%
6	14.77%	13.77%
7	10.51%	14.85%
8	6.51%	13.95%
9	3.57%	11.60%
10	1.76%	8.65%

Table 6.2 – Exception probabilities inaccurate models (p = 0.02 and p = 0.03)

The two problems we mentioned above make it impossible to set strict limits to model acceptation. This is the reason why Basel II regulation uses the 3-zone approach.

Size

The Basel II back test becomes more powerful if we increase the look back period. A small look back period and a high VaR confidence level will cause few exceptions. If the number of exceptions is small, the zones of Basel II are close together. So, the test will more easily classify a model in the wrong zone.

6.1.2 Kupiec Proportion of Failure Test

Description

Kupiec describes one of the first and best known back test alternatives (Kupiec, 1995). It is an extension of the 'standard' Basel II back test. It also tests specifically the number of expected exceptions versus the realised number of exceptions. But this test judges a model as inaccurate if the number of exceptions is significantly higher or lower than the expected number. So the test is two tailed.

Goal

The goal of the Kupiec proportion of failure test is to determine if a VaR model is accurate by testing if the realised number of exceptions is not significantly different from the expected number of exceptions.

Exception Distribution

The distribution of the exceptions under the null hypothesis is the same as in the Basel II back test, the binomial distribution:

$$P(N = k) = \binom{250}{k} p^k (1-p)^{250-k} \; for \; k = 0,..,250 \qquad (6.6)$$

Test measurements

The null hypothesis presumes that the empirically realised probability $(\tilde{p})$ is equal to the theoretical probability (p):

$$H_0: p = \tilde{p} = \frac{x}{n} \tag{6.7}$$

The alternative hypothesis presumes that these probabilities are not equal:

$$H_1: p \neq \frac{x}{n} \tag{6.8}$$

Again, the test represents the exceptions by a random variable N with a binomial distribution. The most suitable test for comparing a theoretical and realised value is the likelihood ratio test. This type of test computes a test statistic for each number of realised exceptions. The following formula represents the test statistic:

$$LR_{POF} = -2\ln\left(\frac{p^x(1-p)^{n-x}}{\tilde{p}^x(1-\tilde{p})^{n-x}}\right) \tag{6.9}$$

The test statistic has a chi-square distribution with one degree of freedom $(\chi^2(1))$. For a confidence level α this test statistic has a critical value. If the test statistic is bigger than this value, the actual number of exceptions is significantly differen. from the expected number at confidence level α. In other words, the VaR model is inaccurate.

Power

To give an impression of the power of the proportion of failure test, we determine the acceptance regions for a VaR models with 99% coverage level. This way we can compare the zone classification for this test with the one for the Basel II test. The acceptance region consists of the number of realised exceptions for which the test does not reject the null hypothesis, while the rejection region consists of the number of realised exceptions which the test rejects. We determine these zones by checking if the test statistic for a certain number of realised exceptions is larger than the critical value. The critical values have the chi square distribution with 1 degree of freedom. Finally we compute the acceptance region.

The main difference with the Basel II back test is the fact that the acceptance zone (green zone in Basel) does not start at zero exceptions. The Kupiec back test rejects a model if it generates too few exceptions. Next to that, the acceptance zone for 250 days of data with a 99% VaR is 1 – 6 (see Appendix B), where the green zone for Basel II is 0 4. But Basel II also has a yellow zone which ranges from 5 – 9 exceptions. So the Kupiec POF test rejects a model faster than the Basel II model. So it would suffer from less type 2, but more type 1 errors. Please note that the critical value we used for the Kupiec test is at 95% confidence level. But even for a 99% critical value the Kupiec test rejects models earlier than the Basel II back test.

Size

Kupiec indicates that his new test has problems with rejecting inaccurate models if the look back period is small. Similar to Basel II, a small number of exceptions will often cause the model to accept inaccurate models.

6.1.3 Kupiec Time Until First Failure Test

Description

This test closely resembles the previous Kupiec test. The only difference is that the likelihood ratio test will now measure the time until the first exception.

Goal

The goal of this test is to test if the underlying VaR model is accurate by checking if the realised time until the first failure is significantly different from the expected time until the first failure.

Exception Distribution

Under the null hypothesis the exceptions have a binomial distribution:

$$P(N = k) = \binom{250}{k} p^k (1-p)^{250-k} \; for \; k = 0,..,250 \tag{6.10}$$

Test measurements

If v represents the time until the first exception, the test considers the following null hypothesis (if we use a 99% VaR level):

$$H_0: p = \hat{p} = \frac{1}{v} = 0.01 \tag{6.11}$$

And the alternative hypothesis is:

$$H_1: p \neq 0.01 \tag{6.12}$$

The following formula now defines the test statistic as:

$$LR_{TUFF} = -2ln\left(\frac{p(1-p)^{v-1}}{\hat{p}(1-\hat{p})^{v-1}}\right) \tag{6.13}$$

The chosen confidence level again determines the critical value of this test. Again, if the realised value of the test statistic is bigger than the critical value, the test judges the model as inaccurate.

Power

This test has lower power than the previous Kupiec test. This is caused by the fact that it only tests the period until the first occurrence of an exception.

Size

Since the underlying assumptions are the same as in the POF test, the sample size again needs to be large to provide powerful results.

6.1.4 Quality control of risk measures test

Description

This method, proposed by de la Pena et. al. is an alternative to the regular Basel II and Kupiec methods (de la Pena, Rivera and Ruiz-Mata, 2007). It recognises the low power of the Basel test concerning the type 2 error, leading to the acceptance of an inaccurate VaR model.

Goal

The goal of this method is to judge whether or not a VaR model is accurate in the same way Basel II does while controlling the type 2 error of Basel II.

Exception Distribution

The test uses the same assumptions as the Basel II method, so again the exceptions have a binomial distribution.

Test measurements

In essence, the quality control of risk measures test simply switches the hypotheses of the Basel II test. Let p be the probability of an exception occurring during any given day. The hypotheses are defined as:

$$H_0 = p > 0.01$$
$$H_1 = p \leq 0.01 \tag{6.14}$$

Hence if we accept the null hypothesis, we reject the VaR model. Inverting the hypotheses also switches the type 1 and type 2 errors of the Basel test. The test wants to control the type 2 error of the Basel II test, so it should control its own type 1 error. This error is rejecting the null hypothesis while it is true.

In Basel II the probability of rejecting an accurate model (99% VaR) is only 0.03% in case of a look back period of 250 days. This comes at a cost, because the probability that the test accepts (ending in the green zone) an inaccurate model is relatively large.

The confidence intervals in Basel II are such that the yellow zone starts at the point where the cumulative probability of the number of exceptions equals or exceeds 95%, and the red zone begins at the point where the cumulative probability equals or exceeds 99.99%.

The QCRM test computes the green, yellow and red zone classification while it makes sure that the type 1 error of their test does not become larger than 1%. De la Pena uses a numerical optimisation structure for this, which is described in Appendix E.1.

Table 6.3 shows the QCRM zone classification. The only difference is that they use a 99% limit to the red zone instead of the 99.99% limit used in Basel II.

Zone	Number of exceptions
Green	0-5
Yellow	6-7
Red	>8

Table 6.3 – Zone classification (De la Pena, Rivera, Mata, 2007)

As we can see, the new test has a smaller yellow zone, compared to the Basel II back test So a 99% coverage VaR model will be accepted for 0-5, questioned for 6 or 7 and rejected for 8 or more exceptions.

Power

The designers of the QCRM test perform a formal power test to compare the quality of their test with the Basel II test. This shows that rejecting a 99% coverage VaR model while it is correct will happen in less than 0.03% of the cases for Basel II and in less than 0.4% of the cases for the QCRM test. So the QCRM test is a bit less powerful on this subject, but the probability of accepting an inaccurate model is reduced to (less than) 1%.

Size

This test does not suffer too much from size problems since its power is high.

6.2 Selection

This section describes which of the exception frequency tests we include in the implementation of the new tool.

6.2.1 Score sheet

The following table contains the scores that we assigned to each of the requirements for the tests.

Back testing method	Goals	Power	Size	Feasibility	Selected
Frequency tests					
Basel II	-	-	--	+	yes
POF	+ / -	+ / -	-	+	yes
TUFF	-	-	-	-	no
QCRM	+	++	+	++	yes

Table 6.4 – Score sheet tests

6.2.2 Argumentation

Basel II

The Basel II test often accepts inaccurate models, so its power is not very high. It does not reach its goal very well, because of this power problem. But the tests feasibility is quite high since the test is the standard. We can also use it as a benchmark for the other tests in the process. Next to that, it is quite easy to implement. So, we select the Basel II test.

Kupiec Power Of Failure (POF)

The POF test reaches its goal reasonably well. It has a small acceptance region compared to Basel II and it can provide more insight if a model is over- or underestimating risk.

The power of the POF test cannot be directly compared to the power of the Basel II model. Since a likelihood ratio statistic is used, it is not possible to compute the type 1 and 2 errors. The probability distribution of the number of exceptions as given in Appendix B is not cumulative. All the test can do is either accept or reject the accuracy of a model at a given confidence level. But, the acceptance region for Kupiec is smaller than the Basel II back test, so its power score is better than for the Basel II back test. The same reasoning holds for the size score.

The feasibility of this test is moderate. It is easy to implement and tests the number of exceptions in a slightly different way and stricter compared to the Basel II test, which might provide additional insight. The most interesting feature of this test is its ability to detect VaR models that overestimate risk or produce too few exceptions. If this happens, the Basel II capital requirement is low, since no or few exceptions are produced by the model.

So, we select the POF test.

Kupiec Time Until First Failure (TUFF)

The goal of TUFF is to test VaR model accuracy by judging the time until first failure. The likelihood ratio test reaches this goal fairly well.

The power of the test is low, compared to the POF test, since it only tests for the time until the first exception occurs and does not look at the remainder of the period.

The size receives the same score as the POF test since the underlying assumptions are the same as in the POF test, so the sample size again needs to be large provide powerful results.

The added value of the test lies in the time until the first failure property. But it tests only for the first failure and does not say anything about the distribution of exceptions over time. This does not provide much insight in VaR model accuracy.

So, we do not select the TUFF test.

QCRM

We select the quality control of risk measures method, because of its high power. It reduces the important limitation of the Basel II back test, considering the type 2 error or accepting inaccurate models. At the same time,

the costs for reaching these results are limited. The QCRM test has a slightly increased error on rejecting accurate models compared to the Basel II back test.

Since the power of the test increased, the influence of the sample size is also less important. For lower sample sizes, the QCRM test will have better results compared to Basel II, since its zone classification is stricter.

6.3 Implementation

One of the requirements of the tool states that we should present the test results for each back testing method with a zone classification similar to the Basel II back test. The QCRM back test also uses a zone classification which is described in section 6.1.4. The following section describes how this classification is implemented for Kupiec test.

6.3.1 Zone Classifications Kupiec

In order to make the results of this test easy to interpret, we introduce a zone classification similar to the Basel II back test. The green, yellow and red zones again have the same interpretation as in the Basel II back test. Next to that, we introduce two new zones (dark blue and light blue) that have a similar interpretation as the yellow and red zone. But where the yellow and red zones indicate that the model is possibly inaccurate in that it generates too many outliers, the dark and light blue zones indicate that the model is possibly inaccurate because it generates too few outliers. In order to give an idea what the Kupiec zone classification looks like we determine the zone classification for a 99% VaR level and a look back period of 250 days.

Zone	Number of exceptions
Dark blue	0
Light blue	1
Green	2-5
Yellow	6
Red	≥7

Table 6.5 – Zone classification Kupiec test

So if such a model generates 0 or 1 exceptions the Kupiec test indicates that the model overestimates risk. If it generates 2-5 exceptions, the test accepts the model as accurate. We question the model's accuracy if it generates 6 exceptions. If the number of realised exceptions is 7 or larger, the Kupiec test indicates that the model underestimates risk.

Appendix D.5 provides a more detailed explanation of the implementation of the zone classification for the Kupiec test.

7 EXCEPTION CLUSTERING

A disadvantage of testing exception frequency concerns the independence of exceptions over time. Consider the example of the introduction of Chapter 6 again: if the number of exceptions over the 250 days is 4, the exception frequency test will judge the VaR model as accurate. But if all four exceptions were during the last 20 days, it is likely there is some market condition that the VaR model cannot cope with: hence the model is inaccurate. On top of that, if clustering of exceptions happens at multiple banks at the same time, this can have large consequences for the industry. In this situation, the independence test proves its usefulness. In an accurate VaR model, the exceptions should be independent of each other. In other words, the probability that an exception occurs during a given day should be independent of the history of exceptions before that day. The independence tests consider this property of the exceptions. This type of test only considers the occurrence of the exceptions over time. The tests do not consider the number of exceptions or the size of the exceptions.

This chapter gives an overview of the exception frequency tests that we discuss in this project. The first section describes the details about the methods we encountered during the literature research. The second section contains the selection of the methods that we implement in the tool. The third section describes what choices we made in the implementation of the tests. Finally, the fourth section gives an overview of the results of the initial test run.

7.1 Test Descriptions

In this section we describe the exception clustering tests we investigated in the literature research. For each back test we first give a general description of the test. The goal is one of the requirements that we use in the test selection. Next, we indicate what the underlying distribution of the exceptions is. After that we indicate what the test measurements (e.g. hypotheses) are. Power is the requirement that indicates how large the test's statistical errors are. Large errors mean low power of the test. Finally, we describe what the influence of the length of the look back period on the test results is.

7.1.1 Likelihood Ratio Test for Independence

Description

The likelihood ratio test for independence is an addition to the exception frequency tests by testing if no clustering of exceptions over time occurs. (Christoffersen, 1998)

Goal

The independence test checks the accuracy of the VaR model by investigating if the occurrence of exceptions over time is independently distributed.

Exception Distribution

The basic assumption of the test is that an accurate VaR model will generate an independent series of exceptions. This is reasonable since an accurate VaR model should generate an exception on any given day with a probability p. It does not depend on the results of previous days.

The following formula represents the series of results showing if exceptions occurred or not.

$$I_{t+1}(\alpha) = \begin{cases} 1, & x_{t,t+1} \leq -VaR_t(\alpha) \\ 0, & x_{t,t+1} > -VaR_t(\alpha) \end{cases} \tag{7.1}$$

This vector simply contains a string of zeros and ones that describes how often and when the VaR was crossed over time. The independence property of the exceptions over time means in this context that each pair of elements from the result vector should be independent of each other. Christoffersen suggests checking this by showing that the history of previous results $\{\dots, I_{t-1}(\alpha), I_t(\alpha)\}$ does not influence the expected value of the result of tomorrow:

$(I_{t+1}(\alpha))$. He wants to proof this by showing that the sequence of results is independently Bernoulli distributed with parameter p (= probability of an exception on a single day).

Test measurements

Christoffersen describes a test that is capable of testing independence using a first order Markov chain $\{I_t\}$ for two successive results. The transition probability matrix shows what the probabilities are of either an exception or no exception given that the day before an exception or no exception has taken place:

$$\Pi_1 = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix} \tag{7.2}$$

Here $\pi_{ij} = P(I_t = j | I_{t-1} = i)$ is the probability that j occurs at time t given that i occurred at time $t-1$. The following formula shows the likelihood function for this function:

$$L(\Pi_1; I_1, I_2, .., I_T) = (1-\pi_{01})^{n_{00}} \pi_{01}^{n_{01}} (1-\pi_{11})^{n_{10}} \pi_{11}^{n_{11}} \tag{7.3}$$

Here n_{ij} is the number of observations with value i followed by j. We estimate the Markov transition matrix by simply computing the ratios of the appropriate cells (which are the maximum likelihood estimates for the values in matrix (7.2)):

$$\hat{\Pi}_1 = \begin{bmatrix} \frac{n_{00}}{n_{00}+n_{01}} & \frac{n_{01}}{n_{00}+n_{01}} \\ \frac{n_{10}}{n_{10}+n_{11}} & \frac{n_{11}}{n_{10}+n_{11}} \end{bmatrix} \tag{7.4}$$

In the next step, we take a look at the result vector $\{I_t\}$. If the elements of the result vector are independent, the Markov transition probability should look like:

$$\Pi_2 = \begin{bmatrix} 1-\pi_2 & \pi_2 \\ 1-\pi_2 & \pi_2 \end{bmatrix} \tag{7.5}$$

In other words, there is no difference between the probabilities for an exception or no exception for a certain day, no matter what the result was on the day before. Hence we define the null hypothesis as:

$$H_0: \pi_{01} = \pi_{11} = \pi_2 \tag{7.6}$$

The maximum likelihood estimator for π_2 is:

$$\hat{\pi}_2 = \frac{(n_{01}+n_{11})}{(n_{00}+n_{10}+n_{01}+n_{11})} \tag{7.7}$$

We compute the likelihood function under the null hypothesis by:

$$L(\Pi_2; I_1, I_2, .., I_T) = (1-\pi_2)^{(n_{00}+n_{10})} \pi_2^{(n_{01}+n_{11})} \tag{7.8}$$

Finally, we compute the test statistic by the following formula:

$$LR_{IND} = -2log\left[L(\hat{\Pi}_2; I_1, I_2, .., I_T) / L(\hat{\Pi}_1; I_1, I_2, .., I_T)\right] \tag{7.9}$$

We obtain the formulas for the maximum likelihood estimates for $\hat{\Pi}_1$ and $\hat{\Pi}_2$ by filling in the maximum likelihood estimates that we saw in formulas (7.4) and (7.7). The test has an asymptotically $\chi^2(1)$ distribution.

Power

We do not compare the power of the likelihood ratio test for independence to the power of the exception frequency tests. The reason for this is that the methods test completely different properties of the VaR model. We showed in the introduction to the back testing types, that the change of one property does not influence the back test of the other property. So it is not useful to compare the power of these tests. But Christoffersen uses Monte Carlo simulations to test how often his method rejects an inaccurate model. He concludes that the test has problems with rejecting inaccurate models for smaller sample sizes.

Size

This exception independence test does not perform very well if the number of exceptions is low. So, this test needs a large look back period to have high power.

7.1.2 Extended Likelihood Ratio Test for Independence

Description

In a previous internship performed at Rabobank, the researcher proposed a technique to cope with a shortcoming of the independence test of Christoffersen (Ermshaus, 2001). The likelihood ratio test of independence only checks for independence between two successive days. Since the probability of two successive exceptions is relatively small, much correlation between exceptions over larger time periods passes this test unnoticed. Ermshaus proposes to use the Markov chain of a ten-day period. Ten trading days are comparable with a period of two weeks.

Goal

The extended likelihood ratio test for independence tests the accuracy of the VaR model by checking if the occurrence of exceptions over time has an independent distribution.

Profit and Loss Distribution

The method uses the same assumptions as Christoffersen.

Test measurements

This test again uses the result vector $I_t(\alpha)$, only this time an element of the vector represents a period of ten days instead of one. Hence the following formula represents the vector:

$$I_t(\alpha) = \begin{cases} 1, & exception\ occurred\ during\ 10\ day\ period \\ 0, & no\ exception\ occurred \end{cases} \qquad (7.10)$$

In order to represent this vector in the Markov matrix, Ermshaus changes the interpretation of π_{ij} and n_{ij} is changed. The i stands again for the occurrence of exceptions, but this time for the occurrence of an exception in the 10 day period, while j stands for the occurrence of an exception at the first day after the 10 day period.

Power

Ermshaus makes no comparison of the power of this test to the original Christoffersen test. Nevertheless we state that the test's power to detect inaccurate models is higher, since it checks for bi-weekly dependency of exceptions instead of two-day dependency.

7.1.3 Duration-Based Test

Description

Christoffersen published another article which provides a back test method that tests the duration of days between the exceptions of the VaR (Christoffersen and Pelletier, 2004)). They use the term duration for the number of days in between two exceptions. the test with the idea in mind that the clustering of exceptions will result in an excessive number of relatively short and relatively long no-exception durations, corresponding to market turbulence and market calmness, respectively.

Goal

The goal of this test is to test the occurrence of exceptions over time to check their independence.

Exception Distribution

This test assumes that that an accurate VaR model will generate a result vector $I_t(\alpha)$ in which the occurrences of exceptions are independently distributed. Christoffersen and Pelletier define the no-exception duration as the period between two exceptions:

$$D_i = t_i - t_{i-1} \tag{7.11}$$

The null hypothesis presumes the VaR model is accurate. In that case the no-exception duration should have no memory and a mean duration of $1/p$ days, because the exceptions are independently distributed over time. Again, p is the probability of an exception occurring at a specific day. The following formula represents the distribution of the durations.

$$f_G(d; p) = p(1-p)^{d-1}, \qquad d \in \mathbb{N} \tag{7.12}$$

This is a geometric distribution. It can be represented by a hazard function that indicates the probability of an exception occurring at day D, given that $D-1$ days there has been no exception:

$$\lambda(d) = P(D = d | D \geq d) = \frac{P(D=d)}{P(D \geq d)} = \frac{f(d)}{\sum_{j=d}^{\infty} f(j)} = \frac{f(d)}{S(d)} \tag{7.13}$$

Here $S(d)$ is the survivor function. If we now insert formula (7.12) in (7.13), this leads to the following simplification (Haas, 2006):

$$\lambda_G(d) = \frac{f_G(d)}{\sum_{j=d}^{\infty} f_G(j)} = \frac{p(1-p)^{d-1}}{\sum_{j=d}^{\infty} p(1-p)^{j-1}} = \frac{p(1-p)^{d-1}}{p(1-p)^{d-1}\sum_{j=0}^{\infty}(1-p)^{j}} = p \tag{7.14}$$

Christoffersen and Pelletier use an exponential function that is the continuous variant of function (7.12). In order to perform the test of independence, they propose a Weibull distribution which is memory free and can be used for the actual test. The following formula presents the probability distribution function:

$$f_{CW}(d; a, b) = ba^b d^{b-1} \exp\left(-(ad)^b\right), \qquad a, b > 0, d > 0 \tag{7.15}$$

This Weibull function becomes the exponential function for $b = 1$. What makes this distribution convenient for testing for independence is the fact that the hazard function can be represented by:

$$\lambda_{CW}(d) = \frac{f_{CW}(d)}{1 - F_{CW}(d)} = ba^b d^{b-1} \tag{7.16}$$

The hazard function is flat for $b = 1$. Or in other words, for $b = 1$, the exceptions over time are independent of each other. If $\lambda > 1 (\lambda < 1)$, the number of exceptions is increasing (decreasing) over time, which suggests their occurrence is not independent.

Test measurements

Based upon the above hazard function, the test formulates the following hypotheses:

$$\begin{aligned} H_{0,IND}&: b = 1 \\ H_{1,IND}&: b \neq 1 \end{aligned} \tag{7.17}$$

In the next steps, the test calculates the durations and the vector C indicating censored or non censored durations:

$$C_i = \begin{cases} 1, & if\ D_i\ is\ censored \\ 0, & if\ D_i\ is\ uncensored \end{cases} \tag{7.18}$$

A duration i is censored if it is at the beginning (end) of the sequence and the first (last) day is not an exception. The following log-likelihood function becomes:

$$\begin{aligned} \ln L(D;\Theta) = C_1 \ln S(D_1) + (1 - C_1)\ln f(D_1) + \sum_{i=2}^{N(T)-1} \ln\big(f(D_i)\big) \\ + C_{N(T)} \ln S\big(D_{N(T)}\big) + \big(1 - C_{N(T)}\big)\ln f(D_{N(T)}) \end{aligned} \tag{7.19}$$

Here $S(D_i) = 1 - F(D_i)$ is the survivor function of (7.15).

The next step the Christoffersen and Pelletier perform is to find the maximum likelihood estimates by numerical optimisation. The test uses these to obtain a value for the test statistic, which it compares to the critical value of the likelihood ratio test.

Power / Size

Christoffersen and Pelletier perform multiple tests to check the power of their test in rejecting inaccurate models compared to the power of the likelihood ratio test for independence of subsection 7.1.1. Their main findings are that their test performs better than the previous likelihood ratio test for independence in almost all situations. Especially if the sample size is large, the difference in power is large. For a small sample of 250 days the test does not have much power. The likelihood ratio test for independence performs better at low sample sizes, but still too bad to draw reliable conclusions. This result emphasizes the need for relatively large look back windows for the independence tests, if the goal of the back test is to detect inaccurate models.

From this result we conclude that the test is not very feasible for rejecting inaccurate models. But if this test rejects a model this is a meaningful result. Also for short look back windows.

7.1.4 Improved Duration-Based Test

Description

Another article underwrites the usefulness of the duration approach we discussed in the previous section, but it also discusses some improvements (Haas, 2006). Haas researches discrete alternatives for the continuous Weibull distribution of (7.15), which gives a better representation of the underlying model which is also discrete.

Goal

This test again tests for independence of exceptions over time. It assumes that if the model is accurate and the probability of an exception at any given day is p, the conditional expected duration between two exceptions will significantly be equal to $1/p$ days.

Exception Distribution

The Weibull distribution that Haas uses has the following probability density function, survivor and hazard functions:

$$f_{DW}(d; a, b) = \exp\left(-a^b(d-1)^b\right) - \exp\left(-a^b d^b\right), \qquad a, b > 0, d \in \mathbb{N} \tag{7.20}$$

$$S_{DW}(d) = \exp\left(-a^b(d-1)^b\right) \tag{7.21}$$

$$\lambda_{DW}(d) = 1 - \exp\left(-a^b\left(d^b - (d-1)^b\right)\right) \tag{7.22}$$

As in the approach of the previous duration method, this test indicates a correct VaR model and hence a flat hazard function by $b = 1$.

Test measurements

Haas derives the following hypotheses:

$$\begin{aligned} H_{0,IND}&: b = 1 \\ H_{1,IND}&: b \neq 1 \end{aligned} \tag{7.23}$$

Haas states that he only tests for independence and not for exception frequency. So he states that the parameter a has the following value:

$$a = -\log(1-p) \tag{7.24}$$

Under the null hypothesis Haas derives a log likelihood function that is similar to the one of the duration-based test (see formula (7.19)). But in this formula one should replace the probability density function and survivor function by (7.20) and (7.21). Next, one should derive the log likelihood function maximum likelihood estimate ($\hat{b}$) for the parameter b using numerical procedures. In the final step one should calculate the following test statistic which he can compare to the critical value of the likelihood ratio test with 2 degrees of freedom.

$$LR_{ID}(d) = -\frac{\exp\left(-a^b(d-1)^b\right) - \exp\left(-a^b d^b\right)}{\exp\left(-\hat{a}^{\hat{b}}(d-1)^{\hat{b}}\right) - \exp\left(-\hat{a}^{\hat{b}} d^{\hat{b}}\right)} \tag{7.25}$$

Power

Haas addresses the problem that the sample size must be relatively large to reliably reject inaccurate models. In order to make sure the tests performed have enough power, he uses the Monte Carlo method of (Dufour, 2005). This test determines the power of statistical tests by performing Monte Carlo simulation over different confidence levels and sample sizes.

The next step that he takes, is to test the power of the improved duration test against the power of the (Christoffersen and Pelletier, 2004)duration test. He performs this test with different look back windows and models with VaR confidence levels of 95 and 99%. All results show that the power of the discrete Weibull test is higher than the continuous Weibull test.

Size

This test also suffers from the necessity of large look back periods.

7.2 Selection

This section describes which of the exception dependency tests we include in the implementation of the new tool.

7.2.1 Score sheet

The following table contains the scores that we assigned to each of the requirements for the tests.

Back testing method	Goals	Power	Size	Feasibility	Selected
Independence tests					
LRT	+ / -	+ / -	--	+ / -	no
ELRT	+ / -	+	--	+ / -	no
Duration	+	+	-	+	no
Improved Duration	+	++	-	++	yes

Table 7.1 – Score sheet tests

7.2.2 Argumentation

Likelihood Ratio Test for Independence

The goal of the test is to judge VaR model accuracy by testing the independence of exceptions over time. It reaches this goal reasonably well, but the Markov test uses only tests for independence between two successive days.

(Christoffersen, 1998) showed in his research that the likelihood ratio test for independence has large power in rejecting inaccurate VaR models if the sample size is large enough. The problem is again that a VaR model with high coverage (>95%) will generate few exceptions, so in order to accurately judge properties of these exceptions, a large sample is needed. We do not include the Likelihood Ratio Test for Independence method. Although it addresses another property of VaR models, there are methods available that test for higher order independence.

Extended Likelihood Ratio Test for Independence

This test obtains almost the same scores as the previous one. This is logical since this test is almost the same. Its added value is that it checks a 10 day or bi-weekly period for independence which is a more useful period. Although this is an improvement, the test still tests only one period. If, for example within this 10-day period 5 exceptions occur, this cannot be classified by this test. So, we do not select this test, since better methods are available for testing exception clustering.

Duration

The goal of the test is to judge VaR model accuracy by testing the independence of exceptions over time. This goal is reached well. The test performs better than the likelihood ratio tests for independence, since it can capture dependence over the whole look back period instead of a single period. The test also suffers from the inability to reject inaccurate models, but a power test by Christoffersen and Pelletier shows that it is not as bad as the likelihood ratio test for independence. We do not include the duration method, since a later study improved this test.

Improved Duration

The scores for this method are almost equal to the ones of the duration test. But, the use of a discrete probability distribution instead of the continuous distribution increases the power of the test. So, we select the

improved duration method. As stated before, we should test each of the three properties of a VaR model. This test has the highest power of all independence tests, so we include it.

7.3 Implementation

One of the requirements of the tool states that we should present the test results for each back testing method with a zone classification similar to the Basel II back test. The first subsection describes how this classification is implemented for improved duration test. The second subsection explains why the 10-day VaR is excluded from the improved duration analysis.

7.3.1 Zone Classification Improved Duration

For the improved duration test we use a similar zone classification as for the Basel II back test. But the improved duration test considers exception dependency, so the interpretation is a little different. If a result falls in the yellow zone for the Basel II back test, we can say that this result, at a confidence level of 95%, is produced by an inaccurate model that generates too many exceptions. For the duration test, a yellow zone result indicates that, at a confidence level of 95%, the result is produced by an inaccurate model that generates clustered exceptions. The red zone in the improved duration test has the same interpretation, but this uses a confidence level of 99%. Appendix D.6 provides a more detailed explanation of the zone classification design for the improved duration test.

8 EXCEPTION SIZE

If we only look at the number of exceptions or how they are clustered, we do not test an important property of an accurate VaR model. Consider again the previous example: if the number of exceptions is 3 and they are equally spread over the 250 days, the model seems to be accurate according to both the frequency and the independence tests. But if these losses are extremely large and far beyond the VaR, the consequences for the bank might be severe. How well a VaR model handles the size of exceptions can be tested by addressing the exception size. This type of test only considers the size of the exceptions and does not take into account the exception frequency or time dependence.

This chapter gives an overview of the exception frequency tests that we discuss in this project. The first section describes the details about the methods we encountered during the literature research. The second section contains the selection of the methods that we implement in the tool. Finally, the third section gives an overview of the results of the initial test run.

8.1 Test Descriptions

In this section we describe the exception size tests we investigated in the literature research. For each back test we first give a general description of the test. The goal is one of the requirements that we use in the test selection. Next, we indicate what the underlying distribution of the exceptions is. After that we indicate what the test measurements (e.g. hypotheses) are. Power is the requirement that indicates how large the test's statistical errors are. Large errors mean low power of the test. Finally, we describe what the influence of the length of the look back period on the test results is.

8.1.1 Capital and Shortfall Test

Description

A previous report within Rabobank International contains the capital and shortfall test (Mesters, Jonkergouw and Ernshaus, 2001 2001). This test defines the amount of money that is the difference between the exception and the VaR as the shortfall. The other term that is used in this test is regulatory capital. This was described in subsection 4.1.1 as the amount of capital that a bank has to hold for covering its market risk in the form of exceptions.

Goal

This test checks the assumption that a good VaR model combines a low average shortfall with a low average capital requirement. A low average shortfall indicates that the size of the exceptions is not very large. If the capital requirement is small, the average VaR is small. If we can combine these two things the VaR model works well.

Exception Distribution

There are no assumptions on distributions. The only properties of the P&L's which the designers test are the size of the shortfall and the capital requirement and these do not need any distribution assumptions.

Test measurements

The designers measure the size of the shortfall and regulatory capital by drawing conclusions from statistics like the mean and standard deviation. They do not formulate hypotheses.

Power

The capital and shortfall test on its own is not convenient to judge the accuracy of a VaR model. If a VaR model has a very low capital requirement, this is not necessarily an indication of a good model. The capital requirement is based on the average VaR (see section 4.1.1). So a low capital requirement is a result of a low VaR. But a low VaR might be breached often such that the model generates many exceptions, which obviously is not a good model.

A low average shortfall can be caused by either a large or a small number of exceptions. The Capital and Shortfall test has no ability to check the exception frequency, so it cannot provide good judgement on a VaR model's accuracy.

8.2 Selection

This section describes which of the exception dependency tests we include in the implementation of the new tool.

8.2.1 Score sheet

The following table contains the scores that we assigned to each of the requirements for the tests.

Back testing method	Goals	Power	Size	Feasibility	Selected
Size tests					
Capital and Shortfall	+	n/a	-	+ / -	yes

Table 8.1 – Score sheet tests

8.2.2 Argumentation

Capital and Shortfall

We include the capital and shortfall method. We cannot use it to draw strong conclusions on the accuracy of the VaR model, but it provides additional insight in the size property of a VaR model. If the size of the shortfalls is big, this can be an indication of a problem with the VaR model or the suitability of the VaR. Next to that, it is the only test of exception size.

8.3 Implementation

We modify the test slightly. The previous study calculated the capital requirement using the rules of Basel II. All Mesters says about this is: "*a low average capital requirement and a low average shortfall will likely indicate a good model*". It is much easier to simply use the average daily VaR instead of the capital requirement. The VaR and the regulatory capital size have a linear relation and for this reason we draw conclusions only upon the average VaR.

8.4 Test results

The tool executes five back test methods. In the test analysis we only take into account the results of four methods. Previously we identified that the shortfall back test is not able to judge the model's performance like the other methods. The test is useful to provide additional insight into the exceptions size. The goal of the tests we performed here is to judge the model's overall performance. Since the VaR and shortfall back test cannot support the decision making in this judgement, we do not include the results of this test in the analysis.

9 OTHER BACK TESTING METHODS

This chapter contains two tests that cannot be categorised under exception frequency, clustering or size.

9.1 Test Descriptions

In this section we describe the two-category[3] tests that we investigated in the literature research. For each back test we first give a general description of the test. The goal is one of the requirements that we use in the test selection. Next, we indicate what the underlying distribution of the exceptions is. After that we indicate what the test measurements (e.g. hypotheses) are. Power is the requirement that indicates how large the test's statistical errors are. Large errors mean low power of the test. Finally, we describe what the influence of the length of the look back period on the test results is.

9.1.1 Conditional Coverage Test

Description

In his article on the likelihood ratio test for independence, Christoffersen proposes to combine Kupiec's proportion of failure test with his own likelihood ratio test for independence. This combination tests the so-called conditional coverage of the underlying VaR model, which is a combination of exception frequency and clustering

Goal

The conditional coverage test checks both the frequency and independency of exceptions.

Exception Distribution

Christoffersen assumes that the exceptions have a binomial distribution and he uses the result vector $I_t(\alpha)$ again.

Test measurements

The test combines the null hypothesis of the frequency test and alternative hypothesis of the independency test:

$$H_0: p = \tilde{p} = \frac{x}{n} \tag{9.1}$$

$$H_1: \pi_{01} \neq \pi_{11} \tag{9.2}$$

This test has the following test statistic with a χ^2distribution with 2 degrees of freedom:

$$LR_{cc} = -2\log\left(\frac{L(p; I_1, I_2, \dots, I_T)}{L(\widehat{\Pi_1}; I_1, I_2, \dots, I_T)}\right) \tag{9.3}$$

Power

The tests performed by Berkowitz and O'Brien in their article show that the power of this test is good, but only if either both of the underlying tests reject the model or one of the tests rejects the model because the test statistic is much bigger than the critical value.

9.1.2 Bootstrap Method

Description

Dowd presents a general method that can increase insight in the power of statistical tests like the ones mentioned in the previous subsections (Dowd, 2002).

Goal

This test's goal is to extend a test sample such that additional information about distribution parameters can be extracted from the larger data set.

Exception Distribution

From the original sample of the performed test, this method draws a large number of new samples with the same length. From these new samples the bootstrap test makes up distributions for the parameters of the original test like the probability of a certain number of rejections and the null hypothesis.

Test measurements

The test does not have real test measurements. But it draws confidence intervals and expected sizes of the parameters from the new distribution.

Power

Besides simply accepting or rejecting the model, we can draw conclusions with respect to the probability of an accurate or inaccurate model. This is because we obtain such a large dataset with acceptances or rejections. The power of the test remains the same as for the original sample, but it can give more insight into the accuracy of the VaR model.

9.2 Selection

9.2.1 Score sheet

The following table contains the scores that we assigned to each of the requirements for the tests.

Back testing method	Goals	Power	Size	Feasibility	Selected
Other methods					
Conditional Coverage	+ / -	+ / -	-	-	no
Bootstrap	n/a	n/a	+	-	no

Table 9.1 – Score sheet tests

9.2.2 Argumentation

Conditional Coverage

The goal of this method is to test both the number of exceptions and the independence of exceptions over time to determine VaR model accuracy. These goals are reached reasonably well. Both properties are tested, but the test does not perform well when the VaR model violates only one of the two properties.

We need a large sample size for this test, since both underlying tests need a large sample to have good power.

Overall, the feasibility of this test is low, since the properties it tests are already tested separately. So we do not select this test.

Bootstrap

We do not include the bootstrap method in the implementation. The major reason lies in the informational burden that is necessary for implementing the test. For every back test that we perform we would have to create many new samples out of the initial sample to determine the tests parameters. Since we include seven trading books of Rabobank International into the test, this is not a realistic option, due to time limitations.

Next to that, the method generates data is generated from nothing. This is a limitation, because this will enlarge errors in the original sample.

10 TEST CONCLUSIONS SUMMARY

<<deleted, confidential>>

11 ALTERNATIVE VAR MODELS

<<deleted, confidential>>

12 IMPLICATIONS FUTURE RESEARCH

In the optimal situation a VaR model would predict the actual and hypothetical P&L's perfectly. For internal use Rabobank International could use 1- day VaR as an indicator for risk that reliably shows the market risk Rabobank is exposed to. Next to that the 10-day VaR would be as small as possible to reduce the regulatory capital amount. There is not one single model available that has these properties. But the back test results give an indication of the shortcomings of the current model and the alternative model investigation showed some interesting room for improvement. The following sections present implications for future research that might improve the current VaR model.

12.1 Exception clustering

<<deleted, confidential>>

LITERATURE

Basle Committee on Banking Supervision (1988). "International Convergence of Capital Measurement and Capital Standards", Bank for International Settlements, Basle.

Basle Committee on Banking Supervision (1996). "Amendment to the Capital Accord to Incorporate Market Risks", Bank for International Settlements, Basle.

Basle Committee on Banking Supervision (2006). "International Convergence of Capital Measurement and Capital Standards: A Revised Framework Comprehensive Version", Bank for International Settlements, Basle.

Blanco, C. and Oks, M. (2004). "Backtesting VaR models. Quantitative and Qualitative Tests". *Risk Desk* IV(I).

Campbell, S. D. (2005). "A review of backtesting and backtesting procedures". *Finance and Economics Discussion Series*: 1-23.

Christoffersen, P. and Pelletier, D. (2004). "Backtesting Value-at-Risk: A Duration-Based Approach". *Journal of Financial Econometrics* 2(1): 84-108.

Christoffersen, P. F. (1998). "Evaluating Interval Forecasts". *International Economic Review* 39(4): 841-862.

Crnkovic, C. and Drachman, J. (1997). "Quality Control". *Risk* Value-At-Risk and Backtesting Techniques.

Dallavecchia, E. (2008). "A VAR, VAR better thing?" *Risk* 21(2).

de la Pena, V. H., Rivera, R., et al. (2007). "Quality Control of Risk Measures: Backtesting VaR Models". *Journal of Risk* 9(2): 39-54.

De Nederlandsche Bank (2006). "Supervisory Regulation on Solvency Requirements for Market Risk", DNB N.V., Amsterdam.

Dowd, K. (2002). "A Bootstrap Back-test". *Risk* October.

Dufour, J.-M. (2005). "Monte Carlo Tests With Nuisance Parameters:A General Approach to Finite-SampleIinference and Nonstandard Asymptotics". *Journal of Econometrics* 133: 443-477.

Ermshaus, S. W. L. (2001). "Parsimonious Value-at-Risk Models". Tilburg University, Tilburg

Gloria Mundi (2008)."GloriaMundi.org Directory of Resources". Retrieved 03-03-2008, http://www.gloriamundi.org/directory2.asp?SubCatLev1ID=Backtesting.

Haas, M. (2001). "New Methods in Backtesting". *Financial Engineering Research Center Working Paper*.

Haas, M. (2006). "Improved Duration-Based Backtesting of Value-at-Risk". *Journal of Risk* 8(2): 17-38.

Hull, J. C. (2007). "Risk Management and Financial Institutions" 1st ed., Prentice Hall, New York.

Investopedia (2008)."Investopedia.com: Financial Dictionary". Retrieved 19-02-2008, http://www.investopedia.com/dictionary/.

ISI Web of Knowledge (2008)."ISI Web of Knowledge [v.4.1] - Web of Science Home". Retrieved 19-02-2008, http://apps.isiknowledge.com/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch&SID=S18fHKJM9AkadeNo3NO&preferencesSaved=&highlighted_tab=WOS.

ISI Web of Knowledge (2008)."Product Specs - ISI Web of Knowledge ". Retrieved 19-02-2008, http://isiwebofknowledge.com/currentuser_wokhome/cu_productspecs/.

Jorion, P. (2001). "Value at Risk: The New Benchmark for Managing Financial Risk" Second ed., McGraw-Hill, New York.

Jorion, P. (2007). "Financial Risk Manager Handbook" Fourth ed., John Wiley & Sons, Inc., New Jersey.

Kupiec, P. H. (1995). "Techniques for Verifying the Accuracy of Risk Management Models". *Journal of Derivatives* 3: 73-84.

Lopez, J. A. (1999). "Methods for evaluating value-at-risk estimates". *Federal Reserve Bank of San Francisco Economic Review*: 3-17.

Mesters, M., Jonkergouw, E., et al. (2001). "Evaluating Approaches to Forecasting Value at Risk: Basle and Beyond". Rabobank International, Utrecht

Scopus (2008)."Scopus - Basic Search". Retrieved 29-02-2008, http://www.scopus.com/scopus/home.url.

Scopus (2008)."Scopus Info - Scopus Overview - What is it?" Retrieved 19-02-2008, http://info.scopus.com/overview/what/.

Appendix A. KEYWORDS SEARCH

The following list contains the keywords that we used in the literature study to find articles on back testing methods. We used two search engines: ISI Web of Knowledge and Scopus.

ISI Web of Knowledge

backtesting
evaluating "value at risk"
testing "value at risk"
"internal models approach"
"capital requirements" Basel
"capital requirements" market risk
"capital requirements" VaR
"market risk" backtest
"market risk" VaR

Scopus

backtesting
evaluating "Value at Risk"
testing "Value at Risk"
"capital requirements" AND "market risk"
"capital requirements" AND "VaR"
"market risk" AND back test
"market risk" AND VaR

Appendix B. POWER POF TEST

This appendix contains tables with detailed information of the power determination of the proportion of failure test.

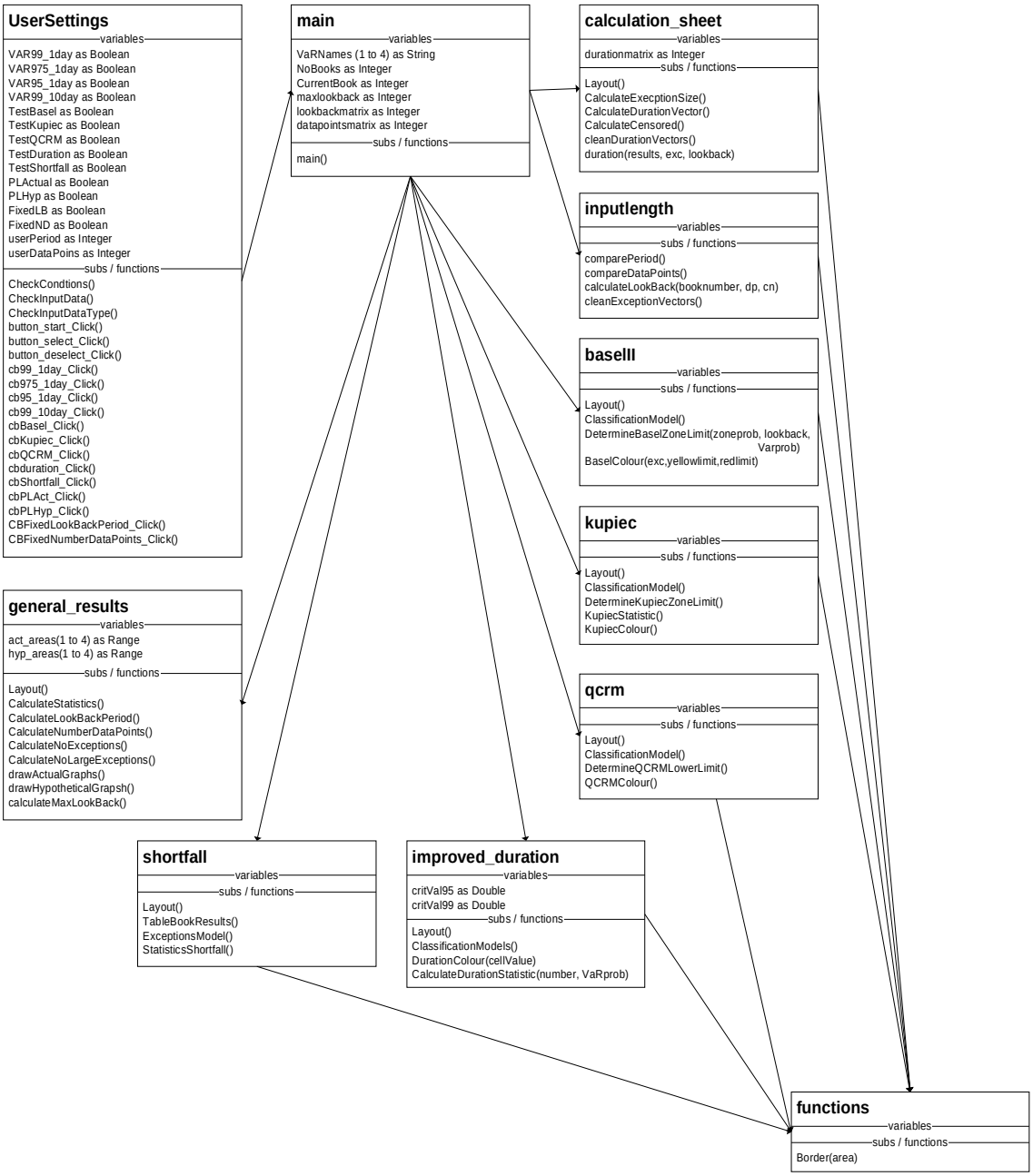
coverage model	99%
sample size (days)	250
expected p	1%
# expected exceptions	2.5
critical value (95%)	3.841459
critical value (99%)	6.634897

# exceptions	realised p	value test statistic	accept / reject (95%)	accept / reject (99%)
1	0.004	1.176491135	accept	accept
2	0.008	0.108435216	accept	accept
3	0.012	0.094940123	accept	accept
4	0.016	0.769138364	accept	accept
5	0.020	1.956809788	accept	accept
6	0.024	3.555354771	accept	accept
7	0.028	5.496990448	reject	accept
8	0.032	7.733550724	reject	reject
9	0.036	10.22903063	reject	reject
10	0.040	12.95549106	reject	reject
11	0.044	15.89061952	reject	reject

Appendix C. MODEL OVERVIEW

C.1. Model Diagram

This diagram shows the different modules that we implemented in the model.



C.2. Model Calculation Steps

The following subsections provide a stepwise approach that shows which calculations the tool makes when it performs a test run.

Introduction Steps

When a user has configured the tool and starts a new test run the tool needs to perform the following calculation steps. These calculations prepare the data that the tool needs for the individual tests and the general results.

1) Test if an input data file is opened and make sure it has the correct format.
2) Determine if the user selected a fixed look back period or a fixed number of data points.
3) Calculate the exception vector (also contains exception size).
4) Calculate the duration vector D_i (see subsection 7.1.3) and censored / uncensored values C (see subsection 7.1.3) for the improved duration test.
5) Count the number of observations in the exception vector and store this value as the actual look back period.
6) Generate the sheets and layout that will contain the test results later.

General Results

1) Calculate statistics selected VaR measures.
2) Calculate statistics selected P&L's.
3) Calculate the look back period that is used.
4) Calculate the number of data points this look back period contains (leave out blank cells).
5) Calculate the number of exceptions for each VaR measure and type of P&L.
6) Calculate the number of exceptions larger than €50,000 for each VaR measure and type of P&L.
7) Draw graph(s) containing the P&L's and VaR measures. For each VaR measure and type of P&L separate graphs are drawn.

Basel II Test

The tool needs to perform the following calculation steps for the Basel II back test:

1) Copy the number of data points from the general results sheet.
2) Determine zone classification in green, yellow and red zones given the number of data points.
3) Classify realised number of exceptions into one of the zones.
4) Colour the cell containing the realised exceptions to show the test result

Kupiec Proportion of Failure Test

The tool needs to perform the following calculation steps for the Kupiec proportion of failure back test:

1) Determine the critical value of the test statistic for the 90 and 98% confidence levels which have been selected as cut off levels for the zone classification.
2) Copy the number of data points from the general results sheet.
3) Determine the zone classification (dark blue, light blue, green, yellow and red) based on the number of data points.
4) Copy the number of realised exceptions from the general results sheet.
5) Colour the cell containing the realised exceptions to show the test results.

QCRM Test

The tool needs to perform the following calculation steps for the quality control of risk measure back test:

1) Copy the number of data points from the general results sheet.
2) Determine the zone classification (green, yellow and red) based on the number of data points. For this calculation step we implemented numerical optimisation. Appendix E shows the theoretical derivations that we made for this.

3) Copy the number of realised exceptions from the general results sheet.
4) Colour the cell containing the number of realised exceptions to show the test results.

Improved Duration Test

The tool needs to perform the following calculation steps for the improved duration based back test:

1) Determine the critical value of the test statistic for the 95 and 99% confidence levels which have been selected as cut off levels for the zone classification.
2) Copy the number of exceptions and number of data points from the general results sheet.
3) Calculate the test statistics for each combination of VaR and P&L. For this the duration vector and censored / uncensored values are used. For this calculation step we implemented numerical optimisation in MATLAB. The theoretical derivations that we made for this are shown in Appendix E.
4) Colour the test statistics cells to show the test results.

Shortfall Test

The tool needs to perform the following calculation steps for the Shortfall back test:

1) Copy the number of exceptions from the general results sheet.
2) Compute shortfall mean and standard deviation, extracted from the exception vector.

Appendix D. IMPLEMENTATION

This appendix discusses some choices that were made in the implementation of the tool. The first section describes the programming language we used. The second section discusses the fixed format of the input data. In the third section we describe what we do when empty values occur in the input data. Section four describes how equal look back periods can have different number of data points. The final two sections provide detailed insight in the zone classification we designed for the Kupiec and improved duration back test.

D.1. Programming Language

In order to choose a convenient programming language it is important to have an overview of the issues and circumstances that can influence this decision:

- The program that Rabobank International currently uses for the back test framework is Microsoft Access®. This program summarises the exceptions that occurred for all trading books and group level and automatically generates part of the documentation for DNB. The tool consists of many tables and queries and no documentation is available.
- The input data that the tool should use is extracted from a Microsoft Excel® tool by experts from the Global Market Risk department.
- The QCRM and improved duration back test both need numerical optimisation in its computations. The most convenient software for this is a mathematical package like MATLAB®.
- The user settings for the new tool need to be available in an input screen.
- The new tool needs to provide an output overview of the test results in graphs and figures.
- Personal experience is high for Excel and Java, moderate for MATLAB and low for Access and VBA.

We implement the tool in Microsoft Excel® in the underlying programming language VBA and partially in MATLAB®. This is the most convenient choice given the issues and circumstances:

- Microsoft Access would have been a logical choice since the current framework is programmed in this environment. But since the current tool does not have any documentation, we would require a lot of time to build in additional functionality in this program. Next to that, Microsoft Access can provide a reasonable user interface, but this requires more programming effort than Microsoft Excel. In Excel we can represent the results immediately in spreadsheets. Finally, the use of the new back testing tool will be different than the current tool. Rabobank International uses the current tool quarterly for the reporting to DNB. The new tool can be used to provide additional insight to DNB on the performance on the VaR model. Experts can use the tool internally to judge the performance of individual trading books. So, there is no added value to combine the tools in one program, since experts will use the tools at different occasions.
- Excel workbooks provide the input data. The choice for Excel and VBA is very convenient in this case. The input data is directly available.
- The tool performs the numerical optimisations for the QCRM and duration back test in MATLAB. This is the most convenient choice, since this package is very suitable for performing these calculations. Next to that, the toolbox ExcelLink of MATLAB provides functionality to call MATLAB procedures from VBA code. A drawback of this procedure is the fact that the user of the tool will need MATLAB on his / her computer to run the tool.
- Excel provides the option to use UserForms to implement screens with checkboxes, text boxes, messages, buttons, etc. This is very convenient for obtaining the user settings and providing messages to the user.

But, this option is also available in Access. So there is no difference between Access and Excel concerning this issue.

- Excel provides functionality to present the results in many ways. There are a lot of options in graphs and tables that we can use to provide the test results in a nice overview. Access also this functionality, but the options are not as extensive as in Excel.
- In terms of personal experience Excel has a big advantage over Access. Although our personal experience with VBA is low, using macro's in Excel provides a very suitable way to get familiar with VBA

D.2. Fixed Format Input Data

In order for the tool to work efficient without asking many actions from the user, it is necessary to use a fixed format for the input data file. If the user makes sure the input data is in the correct format, all that has to be done is to open the input data file and the tool, select the right user settings and run the test. Next to that this makes the implementation of the tool easier, since we can presume that input data always has the same format. We provide the precise requirements for the input data workbook in the technical documentation.

D.3. Missing Values Input Data

The input data is not always complete. The VaR and P&L vectors levels contain empty units. If we would use these cells in the computation, this would generate errors in the test results. If, for example, the input data misses a VaR value, Excel will interpret the VaR has a value of zero. In that case, any negative P&L value will lead to an exception. Rabobank International's risk framework handles missing data by leaving out trading days for which the data is not present. So, for the back testing tool we use the same criterion. If, for any combination of VaR and P&L, we miss at least one of the two, we leave this particular trading day out of the analysis.

D.4. Difference look back periods

Not all trading books have an equally long history, due to the start of new books over time. Next, due to the previous decision to leave out trading days with missing values, even differences in the number of data points to be tested may exist within a trading book. The following table shows an example of the difference between look back period and number of data points.

Date	P&L vector 1	P&L vector 2
31/03/2008	-500,000	-203,592
28/03/2008	234,354	-90,453
27/03/2008	890,342	
26/03/2008	23,404	295,567
25/03/2008	-506,321	12,523
24/03/2008	151,005	163,598
21/03/2008	34,235	-124,830
20/03/2008	98,563	
19/03/2008	-324,091	235,322
18/03/2008	-10,352	143,867
look back period	10	10
# data points	10	8

Table 1 – look back period versus number of data points

From Table 1 we can conclude that a similar look back period can contain a different amount of data points. To make the tool user friendly, the user can choose if he / she wants to compare the books for a fixed number of data points or for a fixed look back period.

D.5. Kupiec Zone Classification

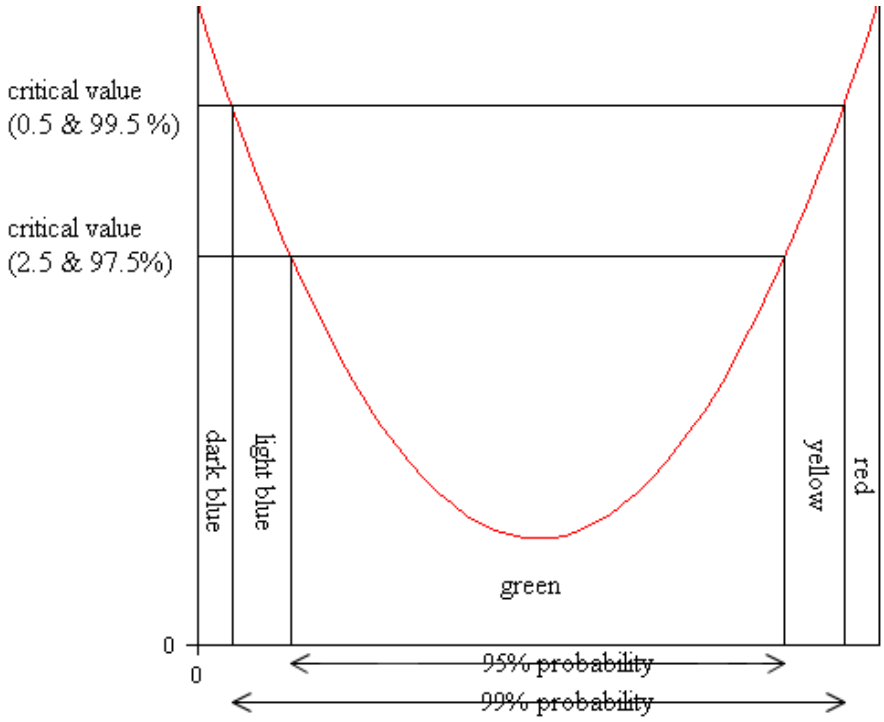


Figure 1 – Kupiec zone classification

The graph represents the value of the test statistic. The horizontal axis represents the realised number of exceptions. The lowest point of the graph is where the realised number of exceptions is equal to the expected number of exceptions. The graph shows that the more different the realised number of exceptions is from the expected number, the higher the value of the test statistic becomes.

D.6. Improved Duration Classification

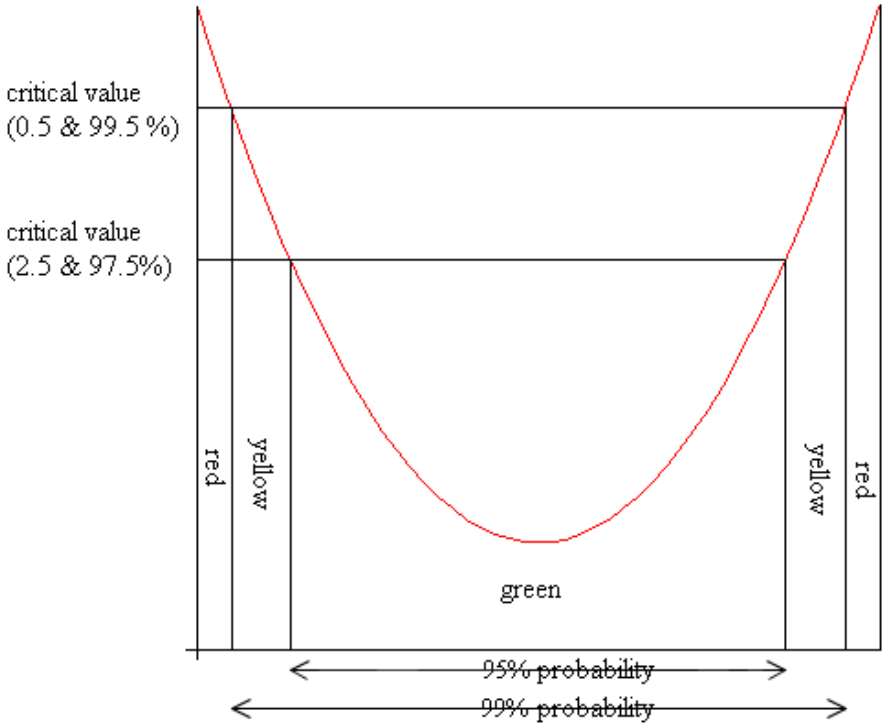


Figure 2 – Improved duration zone classification

The graph represents the value of the test statistic. The horizontal axis represents the degree of independence of the realised durations. The lowest point of the graph is where the realised degree of independence is such that the tested model generates completely independent exceptions. A higher value of the statistic indicates less independent exceptions.

Appendix E. THEORETICAL BACKGROUND

E.1. QCRM lower limit determination

According to de la Pena et. al. (de la Pena, Rivera and Ruiz-Mata, 2007), the lower limit of a zone can be determined by finding the smallest value of p for which the following formula holds:

$$1 - F_X(x) = P\left(\hat{P} \geq \frac{x}{n} | p_L(x, \alpha)\right) \leq \alpha \quad (1)$$

But this equation contains an error. The cause of this error is another formula de la Pena uses in his article:

$$1 - \alpha = P_{p_1}(S_n \leq s(p_1)) \quad (2)$$

The probability function that we mentioned here should be set equal not to the low quantile of α, but to the high quantile $1 - \alpha$:

$$1 - \alpha = P_{p_1}(S_n \leq s(p_1)) \quad (3)$$

We can formulate this in words as that the probability under p_1 that the number of exceptions is smaller than a certain threshold is equal to $1 - \alpha$. If in equation 1 the correct threshold is used (α), the sign within the probability should hence not be $\geq$, but $>$. So the correct formula that is to be used is:

$$1 - F_X(x) = P\left(\hat{P} > \frac{x}{n} | \, p_L(x, \alpha)\right) \leq \alpha \quad (4)$$

We make the following derivations to obtain to an equation that is computable in MATLAB.

$$F_X(x) = P\left(\hat{P} \leq \frac{x}{n} | p_L(x, \alpha)\right) > 1 - \alpha \quad (5)$$

$$F_X(x) = P\left(\frac{s}{n} \leq \frac{x}{n} | p_L(x, \alpha)\right) > 1 - \alpha \quad (6)$$

Here s = number of realised exceptions.

$$F_X(x) = P(s \leq x | x) > 1 - \alpha \quad (7)$$

$$F_X(x) = \sum_{i=0}^{x-1} P(s = i) > 1 - \alpha \quad (8)$$

$$F_X(x) = \sum_{i=0}^{x-1} \binom{n}{i} p^i (1-p)^{n-i} > \alpha \quad (9)$$

In the original formula from the article (equation 1), we had to obtain the smallest value of p. Due to the derivations that we make here, it is possible to determine immediately what the lower limits of the zones are. We do this by setting p equal to the VaR models' probability of generating an exception at any given day. Next, we want to find the smallest number of exceptions s for which equation holds. This is implemented in MATLAB.

E.2. Duration test statistic computation

The duration test has the following the log-likelihood function:

$$\ln L(D;\Theta) = C_1 \ln S(D_1) + (1 - C_1)\ln f(D_1) + \sum_{i=2}^{N(T)-1} \ln\big(f(D_i)\big) + C_{N(T)} \ln S\big(D_{N(T)}\big) + \big(1 - C_{N(T)}\big)\ln f\big(D_{N(T)}\big) \tag{10}$$

Where the following formulas represent the probability density function, survivor function, hazard function and censored values respectively:

$$f_{DW}(d; a, b) = \exp\big(-a^b(d-1)^b\big) - \exp\big(-a^b d^b\big), \qquad a, b > 0, d \in \mathbb{N} \tag{11}$$

$$S_{DW}(d) = \exp\big(-a^b(d-1)^b\big) \tag{12}$$

$$\lambda_{DW}(d) = 1 - \exp\left(-a^b\big(d^b - (d-1)^b\big)\right) \tag{13}$$

$$C_i = \begin{cases} 1, & if\ D_i\ is\ censored \\ 0, & if\ D_i\ is\ uncensored \end{cases} \tag{14}$$

The maximum likelihood ratio test has the following null hypothesis, as formulated by (Haas, 2006):

$$\begin{aligned} H_{0,IND}&: b = 1 \\ H_{1,IND}&: b \neq 1 \end{aligned} \tag{15}$$

We obtain the maximum likelihood estimate for the parameter b from this null hypothesis by maximising the log-likelihood function over this parameter. We use the value we obtained for MLE $\hat{b}$ to compute the test statistic. The following equation is used to compute the test statistic:

$$LR_{ID}(d) = -\frac{\exp\big(-a^b(d-1)^b\big) - \exp\big(-a^b d^b\big)}{\exp\big(-\hat{a}^{\hat{b}}(d-1)^{\hat{b}}\big) - \exp\big(-\hat{a}^{\hat{b}} d^{\hat{b}}\big)} \tag{16}$$

We implemented the maximisation of the log-likelihood function and the computation of the test statistic in MATLAB.