Backtesting Value-at-Risk Models

Kansantaloustiede

Maisterin tutkinnon tutkielma

Olli Nieppola

2009

Kansantaloustieteen laitos
HELSINGIN KAUPPAKORKEAKOULU
HELSINKI SCHOOL OF ECONOMICS

HELSINKI SCHOOL OF ECONOMICS
Department of Economics

BACKTESTING VALUE-AT-RISK MODELS

Master's Thesis in Economics
Olli Nieppola
Spring Term 2009

Approved by the Head of the Economics Department ___/___ 200___ and awarded the grade _____

Author: Olli Nieppola

Department: Economics

Major Subject: Economics

Title: Backtesting Value-at-Risk Models

Abstract:

Value-at-Risk has become one of the most popular risk measurement techniques in finance. However, VaR models are useful only if they predict future risks accurately. In order to evaluate the quality of the VaR estimates, the models should always be backtested with appropriate methods. Backtesting is a statistical procedure where actual profits and losses are systematically compared to corresponding VaR estimates.

The main contribution of this thesis consists of empirical studies. The empirical part of the thesis is carried out in close cooperation with a Finnish institutional investor. The primary objective of the study is to examine the accuracy of a VaR model that is being used to calculate VaR figures in the company's investment management unit. As a secondary objective the empirical research tries to figure out which backtests are the most reliable, and which tests are suitable for forthcoming model validation processes in the company.

The performance of the VaR model is measured by applying several different tests of unconditional coverage and conditional coverage. Three different portfolios (equities, bonds and equity options) with daily VaR estimates for one year time period are used in the backtesting process.

The results of the backtests provide some indication of potential problems within the system. Severe underestimation of risk is discovered, especially for equities and equity options. However, the turbulent market environment causes problems in the evaluation of the backtesting outcomes since VaR models are known to be accurate only under normal market conditions.

Keywords: Value-at-Risk, VaR, backtesting, risk management

Number of Pages: 78

Contents

1. Introduction

1.1 Background

During the past decade, Value-at-Risk (commonly known as VaR) has become one of the most popular risk measurement techniques in finance. VaR is a method which aims to capture the market risk of a portfolio of assets. Put formally, VaR measures the maximum loss in value of a portfolio over a predetermined time period for a given confidence interval.

Despite the wide use and common acceptance of VaR as a risk management tool, the method has frequently been criticized for being incapable to produce reliable risk estimates. When implementing VaR systems, there will always be numerous simplifications and assumptions involved. Moreover, every VaR model attempts to forecast future asset prices using historical market data which does not necessarily reflect the market environment in the future.

Thus, VaR models are useful only if they predict future risks accurately. In order to verify that the results acquired from VaR calculations are consistent and reliable, the models should always be backtested with appropriate statistical methods. Backtesting is a procedure where actual profits and losses are compared to projected VaR estimates. Jorion (2001) refers to these tests aptly as 'reality checks'. If the VaR estimates are not accurate, the models should be reexamined for incorrect assumptions, wrong parameters or inaccurate modeling.

A variety of different testing methods have been proposed for backtesting purposes. Basic tests, such as Kupiec's (1995) POF-test, examine the frequency of losses in excess of VaR. This so called *failure rate* should be in line with the selected confidence level. For instance, if daily VaR estimates are computed at 99% confidence for one year (250 trading days), we would expect on average 2.5 VaR violations, or exceptions, to occur during this period. In the POF-test we would then examine whether the observed amount of exceptions is reasonable compared to the

expected amount. The Basel Committee (1996) has set up a regulatory backtesting framework in order to monitor the frequency of exceptions but, due to the simplicity of the test, there is hardly a reason to use it in internal model validation processes when there are more powerful approaches available.

In addition to the acceptable amount of exceptions, another equally important aspect is to make sure that the observations exceeding VaR levels are serially independent, i.e. spread evenly over time. A good model is capable of avoiding exception clustering by reacting quickly to changes in instrument volatilities and correlations. These types of tests that take into account the independence of exceptions have been suggested, for instance, by Christoffersen (1998) and Haas (2001).

Backtesting is, or at least it should be, an integral part of VaR reporting in today's risk management. Without proper model validation one can never be sure that the VaR system yields accurate risk estimates. The topic is especially important in the current market environment where volatile market prices tend to make investors and more interested in portfolio risk figures as losses accumulate. On the other hand, VaR is known to have severe problems in estimating losses at times of turbulent markets. As a matter of fact, by definition, VaR measures the expected loss only under *normal* market conditions (e.g. Jorion, 2001). This limitation is one of the major drawbacks of VaR and it makes the backtesting procedures very interesting and challenging, as will be shown later in the thesis.

1.2 Objective

The main contribution of this thesis consists of empirical studies. However, in order to provide an exhaustive description about the backtesting process in the empirical part, I will first discuss VaR in general and the theory of different backtesting methods. The purpose of the theoretical part of the thesis is to familiarize the reader with some of the most common backtests by presenting the basic procedures for conducting the tests. The basics of VaR calculation and different approaches to VaR are discussed only briefly, to the extent that the fundamental ideas behind VaR are presented.

Emphasis is laid on the shortcomings of VaR methods while keeping in mind that the potential flaws give motivation for backtesting.

The empirical research is conducted in close cooperation with a large Finnish institutional investor (to which I will refer as the *Company* from here on) who has lately acquired a new VaR calculation system. The software has not yet been backtested with appropriate statistical methods, so the need to validate the model is evident. *The primary goal of this thesis is therefore to examine the accuracy of the software by applying several backtests, analyze the different reasons affecting the outcomes of the tests and to draw conclusions on the results*. In short, the idea behind the backtesting process is to use three investment portfolios for which daily VaR estimates at three confidence levels, namely 90%, 95% and 99%, are calculated for a one year time period. These VaR figures are then compared to actual daily portfolio returns and analyzed with several frequency and independence tests.

All of the backtests presented in the theoretical part cannot be applied in practice due to the nature and certain data requirements of the tests, but the conducted backtests do provide the necessary evidence in order to draw some meaningful conclusions. In addition, the study is limited due to some other technical restrictions. The backtests are applied using historical performance and position data from December 2007 to November 2008. The number of observations is thus limited to 250 trading days. Even though many backtests require a minimum of one year of data, preferably even longer, we are still able to obtain some statistically significant results with the right choice of relatively low confidence levels, such as 90% and 95%.

Despite the fact that the primary purpose of the thesis is to evaluate the quality of the Company's VaR model, one additional aspect is to compare the backtesting methods in such a manner that a solid view on the reliability of the different tests can be formed. *Thus, as a secondary objective the empirical research tries to figure out which tests are the most accurate and powerful, and most importantly, which tests are suitable for forthcoming model validation processes in the Company*.

Methodological issues will not be covered in great detail in this paper, meaning that the reader is assumed to be familiar with statistical decision theory and related

mathematics to some extent. Moreover, thorough proofs for presented functions and formulae are not relevant from the perspective of this thesis, so they are left outside of the scope of this study.

1.3 Structure of the Study

The thesis consists of five chapters of which the first one is the introduction. The second chapter describes the basic idea behind VaR and gives some background and history on the subject. The chapter also discusses the criticism presented against VaR in general.

The third chapter concentrates on the backtesting procedures. Several backtests are presented in detail, but the discussion is by no means exhaustive since it is impossible in this context to go through the variety of different methods and their applications. The aim is rather to focus on the most common backtests, and especially on those that will be applied in practice later in the study.

The fourth chapter forms the empirical part of the thesis and, as such, can be considered to be the core of this study. Some of the tests presented in the preceding chapter are applied to actual VaR calculations. The results are discussed in detail and the factors affecting the outcome are analyzed thoroughly.

The fifth chapter concludes and reviews the most significant results of both theoretical and empirical parts. In addition, some ideas regarding future backtesting processes in the Company are presented.

2. Value at Risk

2.1 History in Brief

Over the past few decades, risk management has evolved to a point where it is considered to be a distinct sub-field in the theory of finance. The growth of risk management industry traces back to the increased volatility of financial markets in 1970's. The breakdown of Bretton Woods system of fixed exchange rates and the rapid advance of new theory, such as adoption of Black-Scholes model, were among the important events that contributed to this 'risk management revolution'. Another factor is simply the fact that trading activity increased significantly. (Linsmeier & Pearson, 1996, Dowd, 1998) For instance, the average number of shares traded per day grew from 3.5 million in 1970 to 40 million in 1990 (Dowd, 1998). At least equally impressive was the growth of the dollar value of outstanding derivatives positions; from $1.1 trillion in 1986 to $72 trillion in 1999 (Jorion, 2001). These elements combined with the unpredictable events in 1990's, such as financial disasters in Barings Bank, Orange County, Daiwa and Metallgesellschaft, highlighted the need for improved internal risk management tools. (Dowd, 1998, Jorion, 2001)

The mathematical roots of VaR were developed already in the context of portfolio theory by Harry Markowitz and others in 1950's. Financial institutions began to construct their own risk management models in 1970's and 1980's, but it was not until the pioneering work from J.P. Morgan and their publication of RiskMetrics system[1] in 1994 that made VaR the industry-wide standard. (Dowd, 1998, Jorion, 2001) During this process, also regulators became interested in VaR. The Basel Capital Accord of 1996 played a significant role as it allowed banks to use their internal VaR models to compute their regulatory capital requirements. (Linsmeier & Pearson, 1996) Since then, VaR has been one of the most used measures of market

[1] RiskMetrics was originally an Internet-based service with the aim to promote VaR as a risk management method. The service provided free data for computing market risk. Later, RiskMetrics became an independent consulting and software firm. (www.riskmetrics.com)

risk and it is likely to gain more acceptance in the near future as the methods are improved further.

2.2 VaR Basics

Firms face many different kinds of risks, including market risks, credit risks, liquidity risks, operational risks and legal risks. VaR was originally developed to measure market risk, which is caused by movements in the level or volatility of asset prices.[2] (Jorion, 2001) According to Dowd (1998), market risks can be subdivided into four classes: interest rate risks, equity price risks, exchange rate risks and commodity price risks. Linsmeier and Pearson (1996) give the following formal definition for VaR:

> *"Using a probability of x percent and a holding period of t days, an entity's value at risk is the loss that is expected to be exceeded with a probability of only x percent during the next t-day period."*

The basic idea behind VaR is straightforward since it gives a simple quantitative measure of portfolio's downside risk. VaR has two important and appealing characteristics. First, it provides a common consistent measure of risk for different positions and instrument types. Second, it takes into account the correlation between different risk factors. This property is absolutely essential whenever computing risk figures for a portfolio of more than one instrument. (Dowd, 1998)

Assuming that asset returns are normally distributed, VaR may be illustrated graphically as in **Figure 1**. In mathematical terms, VaR is calculated as follows:

$$VaR_{\alpha} = \alpha * \sigma * W$$

[2] Even though the original purpose of VaR was to gauge market risk, it was soon realized that VaR methodology may be applied to measure also other types of risks, e.g. liquidity risks and credit risks. (Dowd, 1998)

Here α reflects the selected confidence level,[3] σ the standard deviation of the portfolio returns and W the initial portfolio value. (Jorion, 2001) As an example, consider a situation where initial portfolio value is €100 million and the portfolio returns have an annual volatility of 20%. Calculating a 10-day VaR at 99% confidence level for this portfolio gives us the following result:

$$VaR_{99\%} = -2.33 * 20\% * \sqrt{\left(\frac{10}{250}\right)} * €100M \approx -€9.3M$$

The square root in this function represents the 10-day time horizon assuming 250 trading days in a year. As can be seen, VaR computation is very straightforward if normality is assumed to prevail. However, this assumption has some severe drawbacks which will be discussed shortly.

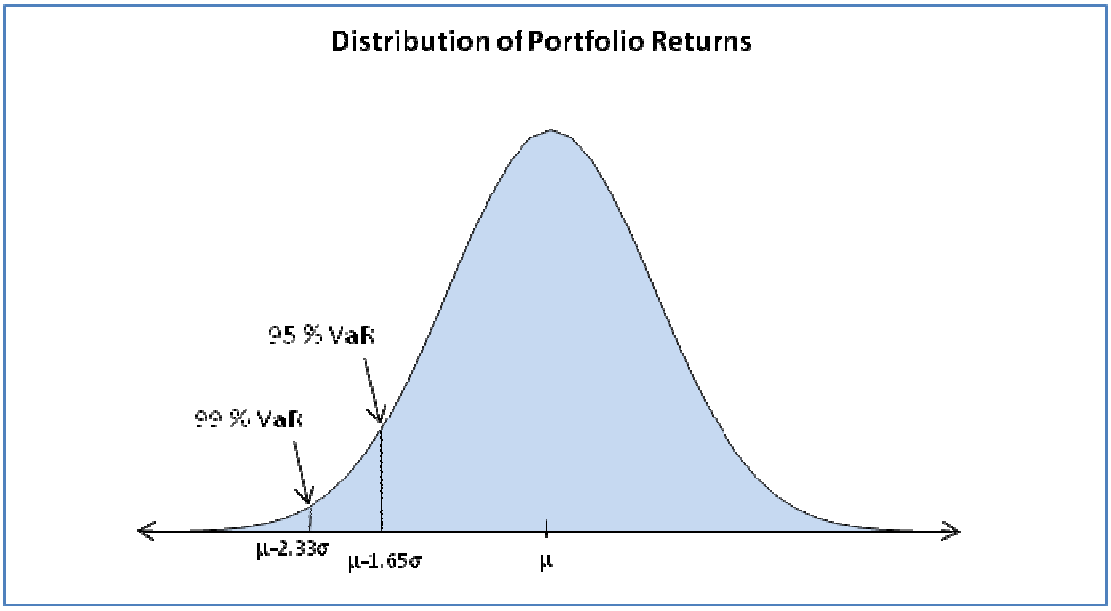


Figure 1: VaR for normal distribution. The graph illustrates Value at Risk for two different confidence levels when portfolio returns are normally distributed.

[3] For instance, α is equal to -2.33 for 99% confidence level and -1.65 for 95% confidence level. These values can be read off from standard normal distribution tables.

When interpreting VaR figures, it is essential to keep in mind the time horizon and the confidence level since without them, VaR numbers are meaningless. Those investors who have actively traded portfolios, such as financial firms, typically use 1-day time horizon, whereas institutional investors and non-financial corporations prefer longer horizons. (Linsmeier & Pearson, 1996) Dowd (1998) suggests that firms should select the holding period according to the length of time it takes to liquidate the portfolio. On the other hand, one must also take into account the properties of the calculation method. For instance, if methods with normal approximations are used, then a relatively short time horizon should be applied.

The choice of confidence level depends on the purpose at hand. If the objective is, as in this paper, to validate a VaR model, then high confidence levels should be avoided in order to be able to observe enough VaR violations. When assessing capital requirements, the confidence level depends on the risk aversion of senior management; risk averse managers choose higher confidence levels. One additional aspect is to consider the possibility of comparing VaR levels with estimates from other sources. (Dowd, 1998)

2.3 Different Approaches to VaR

VaR calculation methods are usually divided into parametric and non-parametric models. Parametric models are based on statistical parameters of the risk factor distribution, whereas non-parametric models are simulation or historical models (Ammann & Reich, 2001).

In this section I will briefly present the basics of the three most common VaR calculation methods; variance-covariance approach, historical simulation and Monte Carlo simulation. The following discussion is meant to be mainly descriptive as the focus is on strengths and weaknesses of each method. Thorough mathematical presentations are beyond the scope of this short review. For more comprehensive approaches regarding different VaR methods please see, for instance, Dowd (1998) or Jorion (2001).

2.3.1 Variance-covariance Approach

Variance-covariance approach is a parametric method. It is based on the assumption that changes in market parameters and portfolio value are normally distributed. (Wiener, 1999) The assumption of normality is the most basic and straightforward approach and is therefore ideal for simple portfolios consisting of only linear instruments (Dowd, 1998).[4]

When implementing variance-covariance approach, the first step is to 'map' individual investments into a set of simple and standardized market instruments.[5] Each instrument is then stated as a set of positions in these standardized market instruments. For example, ten-year coupon bond can be broken down into ten zero-coupon bonds. After the standard market instruments have been identified, the variances and covariances of these instruments have to be estimated. The statistics are usually obtained by looking at the historical data. The final step is then to calculate VaR figures for the portfolio by using the estimated variances and covariances (i.e. covariance matrix) and the weights on the standardized positions. (Damodaran, 2007)

The advantage of variance-covariance approach is its simplicity. VaR computation is relatively easy if normality is assumed to prevail, as standard mathematical properties of normal distribution can be utilized to calculate VaR levels. In addition, normality allows easy translatability between different confidence levels and holding periods.[6] (Dowd, 1998)

---

[4] Linearity means that portfolio returns are linear functions of risk variables. Returns of linear instruments are therefore assumed to be normally distributed. Nonlinear instruments, such as options, do not have this property of normality. (Dowd, 1998)

[5] The reason for the mapping procedure is that as the number of instruments increase, the variance-covariance matrix becomes too large to handle in practice. Instead of calculating variances and covariances for potentially thousands of individual assets, one may estimate these statistics only for the general market factors which will then be used as risk factors for the assets. (Damodaran, 2007)

[6] VaR can be adjusted for different time horizons by rescaling by the ratio of the square root of the two holding periods:

$VaR_{t_2} = \sqrt{\frac{t_2}{t_1}} VaR_{t_1}$.

Also translatability between confidence levels is simple, for example from 99% to 95%:

$VaR_{0.99} = \left(\frac{2.33}{1.65}\right) VaR_{0.95}$. (Dowd, 1998)

Despite the ease of implementation of this method, the assumption of normality also causes problems. Most financial assets are known to have 'fat tailed' return distributions, meaning that in reality extreme outcomes are more probable than normal distribution would suggest. As a result, VaR estimates may be understated. (Jorion, 2001) Problems grow even bigger when the portfolio includes instruments, such as options, whose returns are nonlinear functions of risk variables. One solution to this issue is to take first order approximation to the returns of these instruments and then use the linear approximation to compute VaR. This method is called *delta-normal* approach. However, the shortcoming of delta-normal method is that it only works if there is limited non-linearity in the portfolio. (Dowd, 1998) Britten-Jones and Scheafer (1999) have proposed *quadratic Value at Risk* methods, also known as *delta-gamma* models, which go even further as they use a second order approximation rather than a first order one. The improvement over delta-normal method is obvious, but at the same time some of the simplicity of the basic variance-covariance approach is lost (Damodaran, 2007).

2.3.2 Historical Simulation

When it comes to non-parametric methods, historical simulation is probably the easiest approach to implement (Wiener, 1999). The idea is simply to use only historical market data in calculation of VaR for the current portfolio.

The first step of historical simulation is to identify the instruments in the portfolio and to obtain time series for these instruments over some defined historical period. One then uses the weights in the current portfolio to simulate hypothetical returns that would have realized assuming that the current portfolio had been held over the observation period. VaR estimates can then be read off from histogram of the portfolio returns. The assumption underlying this method is that the distribution of historical returns acts as a good proxy of the returns faced over the next holding period. (Dowd, 1998)

Historical simulation has some undeniable advantages due to its simplicity. It does not make any assumptions about the statistical distributions nor does it require estimation of volatilities and correlations. Basically everything that is needed is the time series of portfolio returns. Most importantly, historical simulation can account for fat tails of the return distributions. The method also applies virtually to any type of instrument and uses full valuations.[7] (Jorion, 2001)

However, historical simulation is also flawed in many respects. A common problem with this method is that there is not enough data available. This complication arises when new instruments that have been in the market for a short time are introduced to the portfolio. Despite the fact that this could be a critique of any of the three approaches, it is most prominent in historical simulation method since VaR is calculated entirely on the basis of historical price data. (Damodaran, 2007)

A more serious shortcoming is that historical simulation effectively assumes that the history will repeat itself. Even though this assumption is often reasonable, it may lead to severely distorted VaR estimates in some cases. (Dowd, 1998) For example, there may be potential risks that are not captured by the historical data set, such as times of very high volatility which may lead to extreme tail losses.

In addition to the abovementioned disadvantages, the users of historical simulation face a challenging trade-off when choosing the time period for the historical market data. It is important to have a long run of data in order to have reliable estimates about the tails of the distribution. This is particularly necessary if high confidence levels are used. On the other hand, using long estimation period leads to a situation where old market data is emphasized too much compared to new information. As a consequence, VaR estimates react slowly to recent changes in market prices, causing estimates to become distorted. One more related problem is that every historical observation is given a weight of one if it is included in the time horizon and zero if it falls out of the

[7] Full valuation means that the instruments in the portfolio are valued properly without any simplifications or approximations. An alternative for this is local valuation where the portfolio is valued only at the initial position and local derivatives are used to infer possible movements. (Jorion, 2001)

horizon. This has an unpleasant effect on VaR estimates when big market jumps fall out of the data set. (Dowd, 1998, Wiener, 1999) A convenient solution to these issues is to use weighted historical simulation which gives lower weights on observations that lie further in the past (Dowd, 1998).

2.3.3 Monte Carlo Simulation

Monte Carlo simulation is another non-parametric method. It is the most popular approach when there is a need for a sophisticated and powerful VaR system, but it is also by far the most challenging one to implement. (Dowd, 1998)

The Monte Carlo simulation process can be described in two steps. First, stochastic processes for financial variables are specified and correlations and volatilities are estimated on the basis of market or historical data. Second, price paths for all financial variables are simulated (thousands of times). These price realizations are then compiled to a joint distribution of returns, from which VaR estimates can be calculated. (Jorion, 2001)

The strength of Monte Carlo simulation is that no assumptions about normality of returns have to be made. Even though parameters are estimated from historical data, one can easily bring subjective judgments and other information to improve forecasted simulation distributions. The method is also capable of covering nonlinear instruments, such as options. (Damodaran, 2007) In addition to these advantages, Jorion (2001) reminds that Monte Carlo simulation generates the entire distribution and therefore it can be used, for instance, to calculate losses in excess of VaR.

The most significant problem with Monte Carlo approach is its computational time. The method requires a lot of resources, especially with large portfolios.[8] As a

[8] Monte Carlo simulation converges to the true value of VaR as $\frac{1}{\sqrt{N}}$, where N is the number of simulations. In order to increase the accuracy of the model by 10 times, one must run 100 times more simulations. (Wiener, 1999) Thus, Monte Carlo is subject to *sampling variation*, which is caused by

consequence, the implementation may turn out to be expensive. (Jorion, 2001) Nevertheless, Monte Carlo will most likely increase its popularity in the future as the costs of computer hardware continuously decrease.

A potential weakness is also *model risk*, which arises due to wrong assumptions about the pricing models and underlying stochastic processes. If these are not specified properly, VaR estimates will be distorted. (Jorion, 2001) Moreover, Dowd (1998) points out that complicated procedures associated with this method require special expertise. Senior management may therefore have hard time keeping abreast of how VaR figures are calculated when Monte Carlo is being used.

2.3.4 Comparing the Methods

Linsmeier and Pearson (1996) suggest that the three methods differ roughly in four dimensions: 1) the ability to capture the risk of options and other nonlinear instruments, 2) the ease of implementation and ease of explanation to senior management, 3) flexibility in incorporating alternative assumptions and 4) reliability of the results. The choice of method ought to be made according to the importance of each of these dimensions and by looking at the task at hand.

Nonlinearity of instruments causes problems for users of variance-covariance approach. This means that when portfolio includes derivative positions, simulation methods should be preferred over (delta-normal) variance-covariance models. (Linsmeier and Pearson, 1996) However, Dowd (1998) argues that if one had a simple portfolio that includes only linear instruments, there would be no point in using Monte Carlo approach since variance-covariance method should yield the same results cheaper and with less effort.

---

limited number of simulation rounds. For example, VaR for portfolio of linear instruments is easily calculated by using variance-covariance approach. Monte Carlo simulation based on the same variance-covariance matrix yields only an approximation and is therefore biased. Accuracy increases only when the simulation rounds are added. (Jorion, 2001)

Variance-covariance and historical simulation methods are known to be straightforward to implement. On the contrary, Monte Carlo has by far the most complicated implementation procedure. This problem is closely related to the issue of ease of explanation to senior management. While Monte Carlo may be difficult to infer, historical simulation is intuitively easy to understand. Variance-covariance approach falls somewhere in between of these two methods. (Linsmeier and Pearson, 1996)

Flexibility of a VaR model is an advantage whenever historical estimates of standard deviations and correlations do not represent the corresponding parameters adequately in the future. In Monte Carlo simulation and variance-covariance approach it is easy to bring in subjective views to the calculation. Historical simulation, on the other hand, does poorly here since the risk estimates are directly derived from historical returns. (Linsmeier and Pearson, 1996)

The reliability of the results is probably the most important issue when comparing the different methods. This is also the dimension that is the most interesting in this context as the focus is on backtesting in the forthcoming chapters. Several studies have been conducted to compare the accuracy of the three approaches. Ammann and Reich (2001) studied the accuracy of linear approximation models (delta-normal approach) versus Monte Carlo simulation. They showed that linear approximation gives fairly accurate VaR estimates, but only to the extent where there is a very limited amount of nonlinear derivatives in the portfolio. Their studies also verified that Monte Carlo simulation yields superior results to linear models when confidence levels and time horizons are increased. Hendricks (1996) examined portfolios with linear instruments using delta normal and historical simulation methods. He found out that delta normal variance-covariance method tends to underestimate VaR, especially under high confidence levels. Historical simulation, on the other hand, performs well also under higher confidence levels. This observation is to be expected as variance-covariance method assumes normality and most assets have fat tailed return distributions.

2.4 Criticism

The previous sections discussed the common shortcomings of the different VaR methods. Let us now turn the focus towards the general criticism that has been raised against VaR as a risk management tool.

The concept of VaR is very simple but this is also one of the main sources of critique. VaR reduces all the information down to a single number, meaning the loss of potentially important information. For instance, VaR gives no information on the extent of the losses that might occur beyond the VaR estimate. As a result, VaR estimates may lead to incorrect interpretations of prevailing risks. One thing that is particularly important to realize is that portfolios with the same VaR do not necessarily carry the same risk. (Tsai, 2004) Longin (2001) suggests a method called *Conditional VaR* to deal with this problem. Conditional VaR measures the expected value of the loss in those cases where VaR estimate has been exceeded.

VaR has also been criticized for its narrow focus. In its conventional form it is unable to account for any other risks than market risk (Damodaran, 2007). However, VaR has been extended to cover other types of risks. For instance, Monte Carlo simulation can handle credit risks to some extent (Jorion, 2001). VaR has also problems in estimating risk figures accurately for longer time horizons as the results quickly deteriorate when moving e.g. from monthly to annual measures. (Damodaran, 2007) Further criticism has been presented by Kritzman and Rich (2002) who point out that VaR considers only the loss at the end of the estimation period, but at the same time many investors look at risk very differently. They are exposed to losses also during the holding period but this risk is not captured by normal VaR models. To take into account for this, the authors suggest a method called *continuous Value at Risk*.

Many economists argue that history is not a good predictor of the future events. Still, all VaR methods rely on historical data, at least to some extent. (Damodaran, 2007) In addition, every VaR model is based on some kinds of assumptions which are not necessarily valid in any circumstances. Due to these factors, VaR is not a foolproof method. Tsai (2004) emphasizes that VaR estimates should therefore always be

accompanied by other risk management techniques, such as stress testing, sensitivity analysis and scenario analysis in order to obtain a wider view of surrounding risks.

3. Backtesting Methods

> *"VaR is only as good as its backtest. When someone shows me a VaR number, I don't ask how it is computed, I ask to see the backtest."*
> (Brown, 2008, p.20)

In the last chapter different VaR calculation methods were discussed. The numerous shortcomings of these methods and VaR in general are the most significant reason why the accuracy of the risk estimates should be questioned. Therefore, VaR models are useful only if they predict future risks accurately. In order to evaluate the quality of the estimates, the models should always be backtested with appropriate methods.

Backtesting is a statistical procedure where actual profits and losses are systematically compared to corresponding VaR estimates. For example, if the confidence level used for calculating daily VaR is 99%, we expect an exception to occur once in every 100 days on average. In the backtesting process we could statistically examine whether the frequency of exceptions over some specified time interval is in line with the selected confidence level. These types of tests are known as tests of *unconditional coverage*. They are straightforward tests to implement since they do not take into account for when the exceptions occur. (Jorion, 2001)

In theory, however, a good VaR model not only produces the 'correct' amount of exceptions but also exceptions that are evenly spread over time i.e. are independent of each other. Clustering of exceptions indicates that the model does not accurately capture the changes in market volatility and correlations. Tests of *conditional*

coverage therefore examine also conditioning, or time variation, in the data. (Jorion, 2001)

This chapter aims to provide an insight into different methods for backtesting a VaR model. Keeping in mind that the aim of this thesis is in the empirical study, the focus is on those backtests that will be applied later in the empirical part. The tests include Basel Committee's (1996) traffic light approach, Kupiec's (1995) proportion of failures-test, Christoffersen's (1998) interval forecast test and the mixed Kupiec-test by Haas (2001). Some other methods are shortly presented as well, but thorough discussion on them is beyond the scope of this study.

3.1 Unconditional Coverage

The most common test of a VaR model is to count the number of VaR exceptions, i.e. days (or holding periods of other length) when portfolio losses exceed VaR estimates. If the number of exceptions is less than the selected confidence level would indicate, the system overestimates risk. On the contrary, too many exceptions signal underestimation of risk. Naturally, it is rarely the case that we observe the exact amount of exceptions suggested by the confidence level. It therefore comes down to statistical analysis to study whether the number of exceptions is reasonable or not, i.e. will the model be accepted or rejected.

Denoting the number of exceptions as x and the total number of observations as T, we may define the *failure rate* as x/T. In an ideal situation, this rate would reflect the selected confidence level. For instance, if a confidence level of 99 % is used, we have a null hypothesis that the frequency of tail losses is equal to $p = (1 - c) = 1 - 0.99 = 1\%$. Assuming that the model is accurate, the observed failure rate x/T should act as an unbiased measure of p, and thus converge to 1% as sample size is increased. (Jorion, 2001)

Each trading outcome either produces a VaR violation exception or not. This sequence of 'successes and failures' is commonly known as Bernoulli trial.[9] The number of exceptions x follows a binomial probability distribution:

$$f(x) = \binom{T}{x} p^x (1-p)^{T-x}$$

As the number of observations increase, the binomial distribution can be approximated with a normal distribution:

$$z = \frac{x - pT}{\sqrt{p(1-p)T}} \approx N(0,1) \ ,$$

where pT is the expected number of exceptions and $p(1-p)T$ the variance of exceptions. (Jorion, 2001)

By utilizing this binomial distribution we can examine the accuracy of the VaR model. However, when conducting a statistical backtest that either accepts or rejects a null hypothesis (of the model being 'good'), there is a tradeoff between two types of errors. Type 1 error refers to the possibility of rejecting a correct model and type 2 error to the possibility of not rejecting an incorrect model. A statistically powerful test would efficiently minimize both of these probabilities. (Jorion, 2001)

Figure 2 displays these two types of errors. Consider an example where daily VaR is computed at 99% percent confidence level for 250 trading days. Assuming that the model is correct (that is, the actual coverage of the model is 99%), the expected number days when losses exceed VaR estimates is $250 * 0.01 = 2.5$. One may set the cut-off level for rejecting a model, for instance, to 5 exceptions. In this case, the probability of committing a type 1 error is 10.8%. On the other hand, if the model has an incorrect coverage of 97%, the expected number of exceptions is $250 * 0.03 = 7.5$.

[9] Bernoulli trial is an experiment where a certain action is repeated many times. Each time the process has two possible outcomes, either success or failure. The probabilities of the outcomes are the same in every trial, i.e. the repeated actions must be independent of each other.

There is now a 12.8% probability of committing a type 2 error, that is, accepting an inaccurate model. **Appendix 1** displays the same probabilities for several model coverages and for different cut-off levels.

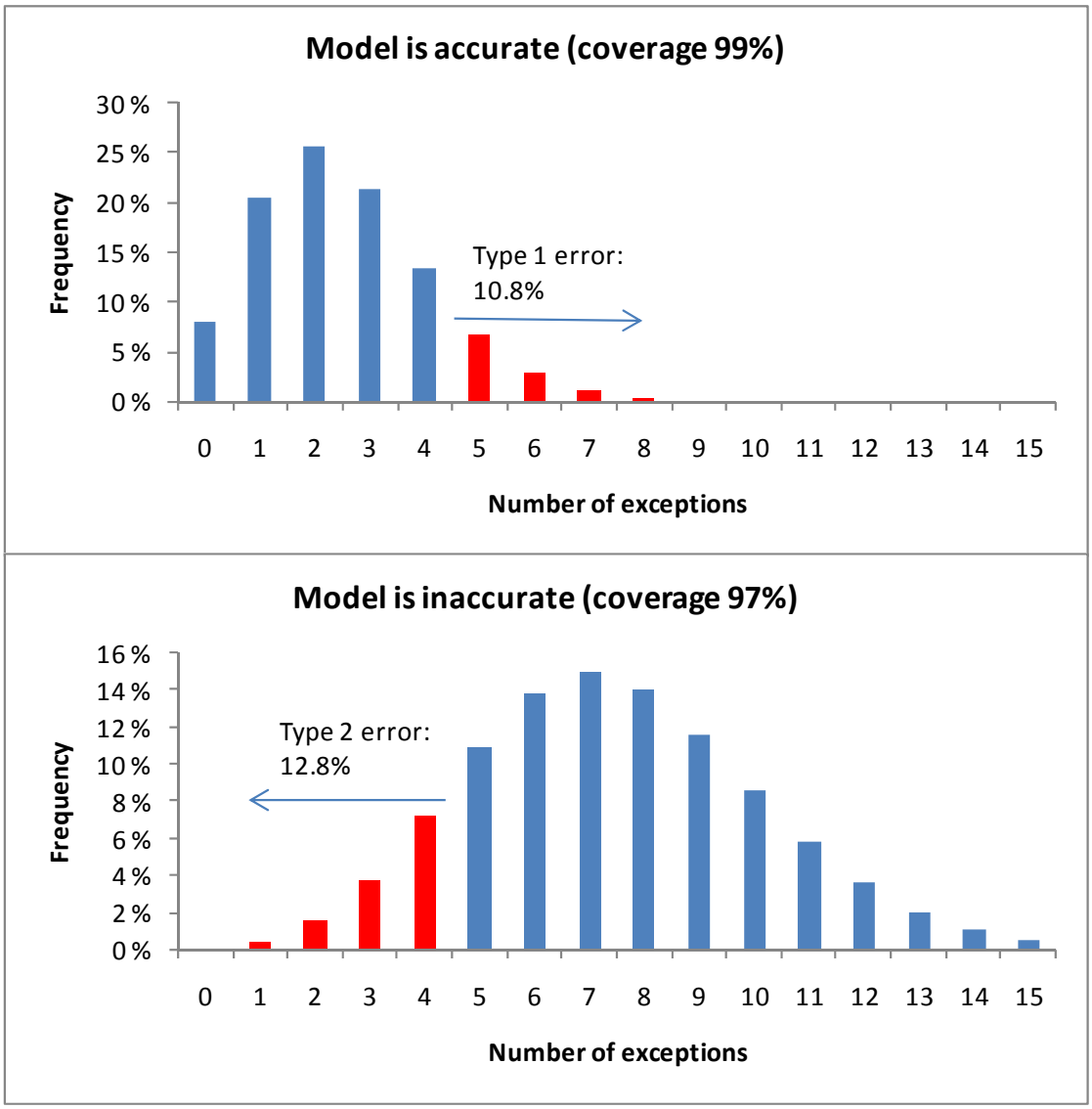


Figure 2: Error types (Jorion, 2001): The upper graph describes an accurate model, where $\boldsymbol{p = 1\%}$. The probability of committing a type 1 error (rejecting a correct model), is 10.8%. The lower graph presents an inaccurate model, where $\boldsymbol{p = 3\%}$. The probability for accepting an inaccurate model, i.e. committing a type 2 error is 12.8%.

3.1.1 Kupiec Tests

POF-Test

The most widely known test based of failure rates has been suggested by Kupiec (1995). Kupiec's test, also known as the POF-test (proportion of failures), measures whether the number of exceptions is consistent with the confidence level. Under null hypothesis of the model being 'correct', the number of exceptions follows the binomial distribution discussed in the previous section. Hence, the only information required to implement a POF-test is the number of observations (*T*), number of exceptions (*x*) and the confidence level (*c*). (Dowd, 2006)

The null hypothesis for the POF-test is

$$H_0: p = \hat{p} = \frac{x}{T}$$

The idea is to find out whether the observed failure rate $\hat{p}$ is significantly different from p, the failure rate suggested by the confidence level. According to Kupiec (1995), the POF-test is best conducted as a likelihood-ratio (LR) test.[10] The test statistic takes the form

$$LR_{POF} = -2ln\left(\frac{(1-p)^{T-x}p^x}{\left[1-\left(\frac{x}{T}\right)\right]^{T-x}\left(\frac{x}{T}\right)^x}\right)$$

Under the null hypothesis that the model is correct, LR_{POF} is asymptotically χ^2 (chi-squared) distributed with one degree of freedom. If the value of the LR_{POF} -statistic

[10] Likelihood-ratio test is a statistical test that calculates the ratio between the maximum probabilities of a result under two alternative hypotheses. The maximum probability of the observed result under null hypothesis is defined in the numerator and the maximum probability of the observed result under the alternative hypothesis is defined in the denominator. The decision is then based on the value of this ratio. The smaller the ratio is, the larger the LR-statistic will be. If the value becomes too large compared to the critical value of χ^2 distribution, the null hypothesis is rejected. According to statistical decision theory, likelihood-ratio test is the most powerful test in its class (Jorion, 2001).

exceeds the critical value of the χ^2 distribution (see **Appendix 2** for the critical values), the null hypothesis will be rejected and the model is deemed as inaccurate.

According to Dowd (2006), the confidence level[11] (i.e. the critical value) for any test should be selected to balance between type 1 and type 2 errors. It is common to choose some arbitrary confidence level, such as 95%, and apply this level in all tests. A level of this magnitude implies that the model will be rejected only if the evidence against it is fairly strong.

Probability Level p	VaR Confidence Level	Nonrejection Region for Number of Failures N		
		T = 255 days	T = 510 days	T = 1000 days
0.01	99 %	N < 7	1 < N < 11	4 < N < 17
0.025	97.5 %	2 < N < 12	6 < N < 21	15 < N < 36
0.05	95 %	6 < N < 21	16 < N < 36	37 < N < 65
0.075	92.5 %	11 < N < 28	27 < N < 51	59 < N < 92
0.1	90 %	16 < N < 36	38 < N < 65	81 < N < 120

Table 1: Nonrejection regions for POF-test under different confidence levels and sample sizes (Kupiec, 1995)

Table 1 displays 95% confidence regions for the POF-test. The figures show how the power of the test increases as the sample size gets larger. For instance, at 95% confidence level with 255 observations the interval $\frac{x}{T}$ for accepting the model is

$$\left[\frac{6}{255} = 0.024;\ \frac{21}{255} = 0.082\right]$$

With 1000 observations the corresponding interval is much smaller:

$$\left[\frac{37}{1000} = 0.037;\ \frac{65}{1000} = 0.065\right]$$

[11] Note that the confidence level of the backtest is not in any way related to the confidence level used in the actual VaR calculation.

Thus, with more data we are able to reject an incorrect model more easily. (Jorion, 2001)

Kupiec's POF-test is hampered by two shortcomings. First, the test is statistically weak with sample sizes consistent with current regulatory framework (one year). This lack of power has already been recognized by Kupiec himself. Secondly, POF-test considers only the frequency of losses and not the time when they occur. As a result, it may fail to reject a model that produces clustered exceptions. Thus, model backtesting should not rely solely on tests of unconditional coverage. (Campbell, 2005)

TUFF-Test

Kupiec (1995) has also suggested another type of backtest, namely the TUFF-test (time until first failure). This test measures the time (v) it takes for the first exception to occur and it is based on similar assumptions as the POF-test. The test statistic is a likelihood-ratio:

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

Here again, LR_{TUFF} is distributed as a χ^2 with one degree of freedom. If the test statistic falls below the critical value the model is accepted, and if not, the model is rejected. The problem with the TUFF-test is that the test has low power in identifying bad VaR models. For example, if we calculate daily VaR estimates at 99% confidence level and observe an exception already on day 7, the model is still not rejected. (Dowd, 1998)

Due to the severe lack of power, there is hardly any reason to use TUFF-test in model backtesting especially when there are more powerful methods available. Later in the study I will present empirical evidence to show that the test actually generates quite

misleading results compared other methods. As Dowd (1998) puts it, the TUFF-test is best used only as a preliminary to the POF-test when there is no larger set of data available. The test also provides a useful framework for testing independence of exceptions in the mixed Kupiec-test by Haas (2001).

3.1.2 Regulatory Framework

Banks with substantial trading activity are required to set aside a certain amount of capital to cover potential portfolio losses. The size of this market risk capital is defined by the bank's VaR estimates. The current regulatory framework requires that banks compute VaR for a 10-day horizon using a confidence level of 99 % (Basel Committee, 2006). Under this framework, it is obvious that a strict backtesting mechanism is required to prevent banks understating their risk estimates. This is why backtesting played a significant role in Basel Committee's decision allowing banks to use their internal VaR models for capital requirements calculation. (Jorion, 2001)

The regulatory backtesting process is carried out by comparing the last 250 daily 99% VaR estimates with corresponding daily trading outcomes.[12] The accuracy of the model is then evaluated by counting the number of exceptions during this period. (Basel Committee, 1996)

The size of the risk capital requirement rises as portfolio risk increases. In addition, the risk capital requirement depends on the outcome of the model backtest (Campbell, 2005) [13]:

---

[12] To align the official backtesting framework with the computation of market risk capital requirement, the Basel Committee has decided that the 99 % confidence should also be used in backtesting, although the Committee recognizes the fact that lower levels would be more suitable in model validation. On the other hand, the Committee insists that using the 10-day holding period in backtesting is not a meaningful exercise, and therefore a period of one day should be used instead. (Basel Committee, 1996)

[13] Market risk capital requirement is formally defined as:
$MRC_t = \max\left[VaR_t(0.01),\ S_t\frac{1}{60}\sum_{i=0}^{59}VaR_{t-i}(0.01)\right] + c$

$$S_t = \begin{cases} 3 & if\ x \leq 4 \quad green \\ 3 + 0.2(x-4) & if\ 5 \leq x \leq 9 \quad yellow \\ 4 & if\ 10 \leq x \quad red \end{cases}$$

Here S_t is the scaling factor of market risk capital requirement and x the number of exceptions over 250 trading days. Basle Committee (1996) classifies backtesting outcomes into three categories: green, yellow and red zones. These categories, which are presented in **Table 2**, are chosen to balance between type 1 and type 2 errors.

Zone	Number of exceptions	Increase in scaling factor	Cumulative probability
Green Zone	0	0.00	8.11 %
	1	0.00	28.58 %
	2	0.00	54.32 %
	3	0.00	75.81 %
	4	0.00	89.22 %
Yellow Zone	5	0.40	95.88 %
	6	0.50	98.63 %
	7	0.65	99.60 %
	8	0.75	99.89 %
	9	0.85	99.97 %
Red Zone	10 or more	1.00	99.99 %

Table 2: Traffic light approach (Basel Committee, 1996): Cumulative probability is the probability of obtaining a given number or fewer exceptions when the model is correct (i.e. true coverage is 99%) The boundaries are based on a sample of 250 observations. For other sample sizes, the yellow zone begins at the point where cumulative probability exceeds 95%, and the red zone begins at cumulative probability of 99.99%

Assuming that the model is correct, the expected number of exceptions is 2.5. If there are zero to four exceptions observed, the model falls into *green zone* and is defined to be accurate as the probability of accepting an inaccurate model is quite low. (Basel Committee, 1996)

The capital requirement is either the current VaR estimate or a multiple (S_t) of the bank's average VaR over the last 60 trading days plus an additional amount of capital (c) set by portfolio's underlying credit risk. (Campbell, 2005)

Yellow zone consists of exceptions from five to nine. These outcomes could be produced by both accurate and inaccurate models with relatively high probability, even though they are more likely for inaccurate models. Backtesting results in the yellow zone generally cause an increase in the multiplication factor, depending on the number of exceptions. However, these increases are not purely automatic since yellow zone does not necessarily imply an inaccurate model. Thus, if the bank is able to demonstrate that the VaR model is 'fundamentally sound' and suffers, for example, from bad luck, supervisors may consider revising their requirements. Basel Committee (1996) therefore classifies the reasons for backtesting failures into following categories:

- *Basic integrity of the model*: The system is unable to capture the risk of the positions or there is a problem in calculating volatilities and correlations.
- *Model's accuracy could be improved*: Risk of some instruments is not measured with sufficient precision.
- *Bad luck or markets moved in fashion unanticipated by the model*: For instance, volatilities or correlations turned out to be significantly different than what was predicted.
- *Intra-day trading*: There is a change in positions after the VaR estimates were computed.

Red zone generally indicates a clear problem with the VaR model. As can be seen from **Table 2**, there is only a very small probability that an accurate model would generate 10 or more exceptions from a sample of 250 observations. As a result, red zone should usually leads to an automatic rejection of the VaR model. (Basel Committee, 1996)

Haas (2001) reminds that the Basel traffic light approach cannot be used to evaluate the goodness of a VaR model because it does not, for instance, take into account the independence of exceptions. The framework has also problems in distinguishing good models from bad ones. These shortcomings were already recognized by the Basel Committee (1996) itself but the Committee has justified the framework as follows:

> *"However, the Committee does not view this limitation as a decisive objection to the use of backtesting. Rather, conditioning supervisory standards on a clear framework, though limited and imperfective, is seen as preferable to a purely judgmental, standard or one with no incentive features whatsoever."*

Due to the severe drawbacks of the Basel framework, the method is probably best used as a preliminary test for VaR accuracy. In any kind credible model validation process the traffic light approach is simply inadequate, and more advanced tests should also be applied.

3.2 Conditional Coverage

The Basel framework and unconditional coverage tests, such as the POF-test, focus only on the number of exceptions. In theory, however, we would expect these exceptions to be evenly spread over time. Good VaR models are capable of reacting to changing volatility and correlations in a way that exceptions occur independently of each other, whereas bad models tend to produce a sequence of consecutive exceptions. (Finger, 2005)

Clustering of exceptions is something that VaR users want to be able to detect since large losses occurring in rapid succession are more likely to lead to disastrous events than individual exceptions taking place every now and then. (Christoffersen & Pelletier, 2003) Tests of *conditional coverage* try to deal with this problem by not only examining the frequency of VaR violations but also the time when they occur. In this section I will present two conditional coverage tests: Christoffersen's (1996) interval forecast test and the mixed Kupiec-test by Haas (2001).

3.2.1 Christoffersen's Interval Forecast Test

Probably the most widely known test of conditional coverage has been proposed by Christoffersen (1998). He uses the same log-likelihood testing framework as Kupiec, but extends the test to include also a separate statistic for independence of exceptions. In addition to the correct rate of coverage, his test examines whether the probability of an exception on any day depends on the outcome of the previous day. The testing procedure described below is explained, for example, in Jorion (2001), Campbell (2005), Dowd (2006) and in greater detail in Christoffersen (1998).

The test is carried out by first defining an indicator variable that gets a value of 1 if VaR is exceeded and value of 0 if VaR is not exceeded:

$$I_t = \begin{cases} 1 & if\ violation\ occurs \\ 0 & if\ no\ violation\ occurs \end{cases}$$

Then define n_{ij} as the number of days when condition j occurred assuming that condition i occurred on the previous day. To illustrate, the outcome can be displayed in a 2 x 2 contingency table:

	$I_{t-1} = 0$	$I_{t-1} = 1$	
$I_t = 0$	n_{00}	n_{10}	$n_{00} + n_{10}$
$I_t = 1$	n_{01}	n_{11}	$n_{01} + n_{11}$
	$n_{00} + n_{01}$	$n_{10} + n_{11}$	N

In addition, let π_i represent the probability of observing an exception conditional on state i on the previous day:

$$\pi_0 = \frac{n_{01}}{n_{00} + n_{01}}, \qquad \pi_1 = \frac{n_{11}}{n_{10} + n_{11}} \qquad and \qquad \pi = \frac{n_{01} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}}$$

If the model is accurate, then an exception today should not depend on whether or not an exception occurred on the previous day. In other words, under the null hypothesis the probabilities π_0 and π_1 should be the equal. The relevant test statistic for independence of exceptions is a likelihood-ratio:

$$LR_{ind} = -2ln\left(\frac{(1-\pi)^{n_{00}+n_{10}}\pi^{n_{01}+n_{11}}}{(1-\pi_0)^{n_{00}}\pi_0{}^{n_{01}}(1-\pi_1)^{n_{10}}\pi_1{}^{n_{11}}}\right)$$

By combining this independence statistic with Kupiec's POF-test we obtain a joint test that examines both properties of a good VaR model, the correct failure rate and independence of exceptions, i.e. conditional coverage:

$$LR_{cc} = LR_{POF} + LR_{ind}$$

LR_{cc} is also χ^2 (chi-squared) distributed, but in this case with two degrees of freedom since there are two separate LR-statistics in the test. If the value of the LR_{cc}-statistic is lower than the critical value of χ^2 distribution (**Appendix 2**), the model passes the test. Higher values lead to rejection of the model.

Christoffersen's framework allows examining whether the reason for not passing the test is caused by inaccurate coverage, clustered exceptions or even both. This evaluation can be done simply by calculating each statistic, LR_{POF} and LR_{ind}, separately and using χ^2 distribution with one degree of freedom as the critical value for both statistics. Campbell (2005) reminds that in some cases it is possible that the model passes the joint test while still failing either the independence test or the coverage test. Therefore it is advisable to run the separate tests even when the joint test yields a positive result.

3.2.2 Mixed Kupiec-Test

Christoffersen's interval forecast test is a useful backtest in studying independence of VaR violations but unfortunately it is unable to capture dependence in all forms because it considers only the dependence of observations between two successive days. It is possible that likelihood of VaR violation today does not depend whether a violation occurred yesterday but whether the violation occurred, for instance, a week ago. (Campbell, 2005) Indeed, the empirical part of this study provides some evidence

that the Christoffersen's test is perhaps inadequate method in capturing dependence between exceptions.

Haas (2001) argues that the interval forecast test by Christoffersen is too weak to produce feasible results. He therefore introduces an improved test for independence and coverage, using the ideas by Christoffersen and Kupiec. Haas proposes a mixed Kupiec-test which measures the time between exceptions instead of observing only whether an exception today depends on the outcome of the previous day. Thus, the test is potentially able to capture more general forms of dependence.[14]

According to Haas (2001), the Kupiec's TUFF-test, which measures the time until the first exception, can be utilized to gauge the time between two exceptions. The test statistic for each exception takes the form

$$LR_i = -2ln\left(\frac{p(1-p)^{v_i-1}}{\left(\frac{1}{v_i}\right)\left(1-\frac{1}{v_i}\right)^{v_i-1}}\right)$$

where v_i is the time between exceptions i and $i-1$. For the first exception the test statistic is computed as a normal TUFF-test. Having calculated the LR-statistics for each exception, we receive a test for independence where the null hypothesis is that the exceptions are independent from each other. With n exceptions, the test statistic for independence is

[14] A similar test based on duration between exceptions has been proposed by Christoffersen and Pelletier (2004). The authors provide evidence that their test has power against more general forms of dependence but at the same time the test does not require any additional information compared to Christoffersen's interval forecast test. The basic insight is that if exceptions are completely independent of each other, then the upcoming VaR violations should be independent of the time that has elapsed since the last exception (Campbell, 2005). To measure the duration between two exceptions, Christoffersen and Pelletier (2004) define the *no-hit duration* as $D_i = t_i - t_{i-1}$. A correct model with coverage rate $p = (1-c)$ should have an expected conditional duration of $1/p$ days and the no-hit duration should have no memory. (Christoffersen & Pelletier, 2004)

$$LR_{ind} = \sum_{i=2}^{n}\left[-2ln\left(\frac{p(1-p)^{v_i-1}}{\left(\frac{1}{v_i}\right)\left(1-\frac{1}{v_i}\right)^{v_i-1}}\right)\right] - 2ln\left(\frac{p(1-p)^{v-1}}{\left(\frac{1}{v}\right)\left(1-\frac{1}{v}\right)^{v-1}}\right)$$

which is a χ^2 distributed with n degrees of freedom. Similarly to the Christoffersen's framework, the independence test can be combined with the POF-test to obtain a mixed test for independence and coverage, namely the mixed Kupiec-test:

$$LR_{mix} = LR_{POF} + LR_{ind}$$

The LR_{mix} -statistic is χ^2 distributed with $n + 1$ degrees of freedom. Just like with other likelihood-ratio tests, the statistic is compared to the critical values of χ^2 distribution. If the test statistic is lower, the model is accepted, and if not, the model is rejected.

3.3 Other Approaches

3.3.1 Backtesting Based on Loss Function

Information contained in the basic backtesting frameworks is somewhat limited. Instead of only observing whether VaR estimate is exceeded or not, one might be interested, for example, in the magnitude of the exceedance. (Campbell, 2005)

Lopez (1998, 1999) suggests a method to examine this aspect of VaR estimates. The idea is to gauge the performance of VaR models by how well they minimize a loss function that represents the evaluator's concerns. Unlike most other backtesting methods, loss function approach is not based on hypothesis-testing framework. Dowd (2006) argues that this makes loss functions attractive for backtesting with relatively small amount of observations.

The general form of the loss functions is such that an exception is given a higher score than nonexception. For example, the loss function may take the following quadratic form:

$$L\left(VaR_t(\alpha), x_{t,t+1}\right) = \begin{cases} 1 + \left(x_{t,t+1} - VaR_t\right)^2 & if \;\; x_{t,t+1} \leq -VaR_t(\alpha) \\ 0 & if \;\; x_{t,t+1} > -VaR_t(\alpha) \end{cases}$$

where $x_{t,t+1}$ is the realized return and VaR_t the corresponding VaR estimate. The numerical score of the model is calculated by plugging the data into this loss function. The score increases with the magnitude of the loss. A backtest based on this approach would then be conducted by calculating the sample average loss (with T observations):

$$\hat{L} = \frac{1}{T}\sum_{t=1}^{T} L\left(VaR_t(\alpha), x_{t,t+1}\right)$$

In order to determine whether the average loss $\hat{L}$ is too large compared to "what it should be", one needs to have some kind of a benchmark value. In practice, this means that the backtest makes an assumption about the stochastic behavior and distribution of the returns. Once the distribution has been determined, an empirical distribution can be generated by simulating portfolio returns. The benchmark value can then be obtained from this distribution. If the sample average loss $\hat{L}$ is larger than the benchmark value, the model should be rejected. (Campbell, 2005)

The loss function approach is flexible. The method can be tailored to address specific concerns of the evaluator since the loss function may take different forms. On the other hand, the loss function approach relies on the correct assumption about the return distribution. In case the distribution is defined incorrectly, the results of the backtest become distorted. Observing a score that exceeds the benchmark value could imply either an inaccurate VaR model or wrong assumptions about the stochastic behavior of profits and losses. (Campbell, 2005)

Lopez (1999) recognizes some of the problems associated with the loss function based backtesting. Due to the nature of this backtest, the method cannot be used to statistically classify a model accurate or inaccurate. Rather, it should be used to monitor the relative accuracy of the model over time and to compare different models with each other. If the purpose is to validate a VaR model, the loss function backtest should be accompanied with hypothesis-testing methods.

3.3.2 Backtests on Multiple VaR Levels

All the backtests discussed so far focus solely on VaR estimates at one single confidence level. However, there is no particular reason to examine only one VaR level since properties of unconditional coverage and independence of exceptions should hold for any confidence level. Several backtests have been proposed to test the entire forecast distribution.

One test of this type has been presented by Crnkovic and Drachman (1997). Their insight is that if the model is accurate, then 1% VaR should be violated 1% of the time, 5% VaR should be violated 5% of the time, and 10% VaR should be violated 10% of the time and so on. In addition, a VaR violation at any confidence level should be independent on violations at any other level. (Campbell, 2005)

The test is set up as follows. Each day we forecast a probability density function for portfolio returns. On the following day when the portfolio return is known we determine the percentile of the forecasted distribution in which the actual return falls. Assuming that the model is calibrated correctly, we expect that each percentile occurs with the same probability and they are independent of each other. These hypotheses can then be tested with several statistics. (Crouchy et. al., 2000)

The advantage of this kind of test is that it provides additional power in identifying inaccurate models (Campbell, 2005). However, the problem with the approach is its data intensity. According to Crnkovic and Drachman (1997), their test requires observations of at least four years in order to obtain reliable estimates. In practice,

there is rarely a chance to use such an extensive set of data, which makes it challenging to use this method in actual backtesting processes.

3.4 Conclusions

Backtesting provides invaluable feedback about the accuracy of the models to risk managers and the users of VaR. This chapter has presented some of the most popular approaches to VaR model validation. A good VaR model satisfies two equally important properties. First, it produces the 'correct' amount of exceptions indicated by the confidence level. Second, the exceptions are independent of each other. The simplest tests focus only on the number of exceptions, whereas more advanced methods take into account the dependence between exceptions. Many of the tests are based on one single confidence level but more recent methods are capable of testing the whole distribution, providing the test with more power.

The most common test is the Kupiec's POF-test, which measures the number of exceptions over some specified time interval. If statistically too many or too few exceptions are observed, the model is rejected. The regulatory framework by Basel Committee is based on the same assumptions as the POF-test. Independence of exceptions can be examined with Christoffersen's interval forecast test. However, as the empirical research will show, a better alternative is to use the mixed Kupiec-test by Haas since it is capable of capturing more general forms of dependence.

Hypothesis-based backtesting always involves balancing between two types of errors: rejecting an accurate model versus accepting an inaccurate model. A statistically powerful test efficiently minimizes both of these probabilities. In order to increase the power of the test, one may choose a relatively low confidence level in VaR calculation so that enough exceptions are observed. For example, in the empirical part of this thesis I will apply lower VaR levels of 90% and 95% to enhance the power of the tests. One should also use as large set of data as possible. However, in practice there rarely is a sufficient amount of observations available.

In the model validation process, careful attention should be paid to the selection of the backtests. A too narrow perspective, such as relying only on the unconditional coverage tests, could potentially lead to a situation where we accept a model that does produce the 'correct' amount of exceptions but is generally unable to react to changes in correlations and volatilities, yielding some closely bunched exceptions. According to Haas (2001), one backtest is never enough and a good result in some test should always be confirmed with another type of test. This argument will be kept in mind in the empirical section.

4. Empirical Backtesting

The empirical part of the thesis is carried out in close cooperation with a large Finnish institutional investor, the Company. The objective of the study is to examine the accuracy of a VaR model that is currently being used to calculate VaR figures in the Company's investment management unit.

The backtesting procedures are conducted by comparing daily profits and losses with daily VaR estimates using a time period of one year, i.e. 250 trading days. The performance of the software is measured by applying the Basel framework and tests by Kupiec (1995), Christofferssen (1998) and Haas (2001). Due to some technical limitations which will be discussed later, it is not possible to apply all the tests presented in the previous chapter. Nevertheless, the backtesting process here is thorough enough for the Company's purposes and provides a satisfactory view on the accuracy of the VaR software at this point. Ideas for further backtesting are presented in the concluding chapter of the thesis.

The purpose of this chapter is not only to present the backtesting process and results in detail, but also to analyze the outcome and the factors that may have affected the

outcome. First, I will give describe the case at hand and the setup for the backtesting process. After that, each of the backtests will be separately conducted with some numerical examples. Finally, the results are interpreted at portfolio level and the concluding section evaluates the performance of the model from a more general perspective.

4.1 VaR Calculation and Backtesting Process

4.1.1 Background

VaR computation process has lately been reorganized in the Company. New VaR calculation software was purchased from an outside vendor in early 2008. The idea behind this transaction was to acquire a system that could calculate VaR estimates for every instrument, including derivatives, in the company's portfolio. This objective has been achieved and the program is currently being used as the primary tool for VaR reporting and stress testing.

The software is based on full Monte Carlo valuation, meaning that no approximations are used in market value calculations. The main factor that we may expect to have an adverse effect on the reliability of the results is model risk. A good presentation regarding the sources of model risk can be found, for example, in Dowd (2006). These issues will be discussed also later in this paper as I analyze the backtesting results.

Even though during the short preliminary testing period the VaR estimates seemed to be in line with benchmarks (i.e. with calculations acquired from other sources), the need for systematical and controlled backtesting was evident in order to make sure that the results are valid and consistent in every respect. Before this study, no proper testing of any kind had been performed on the software. The software itself has an elementary backtesting module installed, but it cannot be considered to be of any use in credible model validation, since it is unable to compare the VaR estimates with actual portfolio performance.

4.1.2 Portfolio Setup and Performance Data

The Company has investment activity in all kinds of financial instruments, including equities, funds, fixed income securities, real estate, commodities and derivatives. Even though VaR reporting will be conducted on the whole portfolio in the near future, it is not possible to include all of the Company's positions into the backtesting process, for several reasons.

First, many instruments do not have consistent daily price quotations available. Daily pricing is absolutely essential since if one calculated performance figures for instruments with price updating rarely, e.g. once a month, there would be long times of zero returns and then once a month high jumps in profits and losses. These figures would not be in line with corresponding VaR estimates which are calculated on the basis of some daily valued market risk factor. As a result, backtesting results would become severely distorted and in practice useless. The lack of daily pricing rules out, for instance, credit bonds, private equities, funds and many types of derivatives from our analysis.

Second, there are also other technical obstacles, such as manual updating of the instrument properties (input data for computing VaR), associated with some instrument types. This issue concerns most notably floating rate bonds, which have to be excluded as well.

Third, because of time restrictions it is simply impossible in this context to produce the position data for the whole portfolio for 250 days. It takes approximately 20-30 minutes to calculate daily VaR estimates for the whole portfolio. By using a smaller portfolio in the backtesting process, computational time can easily be reduced to 5-10 minutes.

Because of these limitations, serious attention has to be paid to the selection of appropriate testing portfolios. Therefore, we have to construct a top portfolio solely for the purposes of this study. The portfolio consists of three subportfolios: an equity portfolio, a fixed income portfolio and a derivative portfolio, including Finnish quoted equities, government bonds, and equity options, respectively. The portfolios in the

study represent actual positions of the Company. With these portfolios we are able to examine the model's ability to capture interest rate risk, equity risk and the risk of nonlinear instruments. This kind of diversified portfolio structure enables us to effectively identify potential problems in different asset and risk classes.

Altogether there are about 30 to 60 instruments included in each VaR calculation, depending on the day of reporting. Development of portfolio market values over time is displayed in **Figure 3**. The most significant issue to pay attention to is the change in bond portfolio market value over the one year time horizon. This is caused by major transactions as the Company altered its fixed income allocation by selling government bonds and purchasing credit bonds. Ideally, we would also like to include credit bonds in the backtesting process but unfortunately, due to the reasons already discussed, this is not possible. The decrease in bond portfolio market value should not have an effect on the backtesting results because position data is calculated on a daily basis. This is rather merely an issue that is good to keep in mind when evaluating the different graphs presented in this paper.

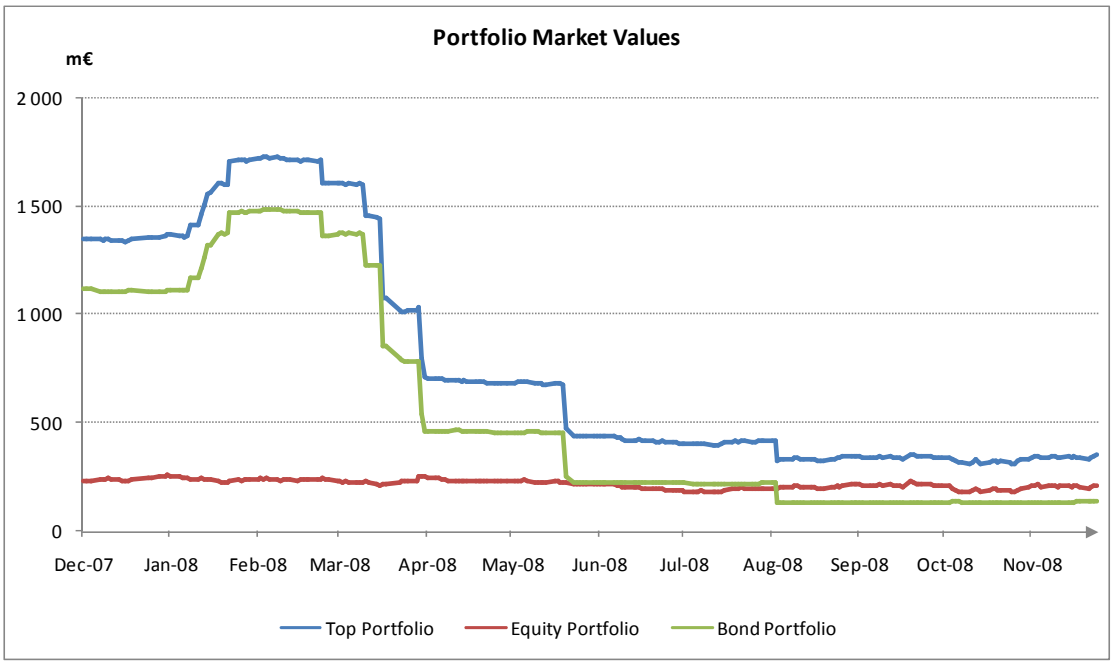


Figure 3: Portfolio market values over time: Derivative portfolio is excluded from this graph since the positions are relatively small with portfolio market value ranging from -2m€ to 6m€ over time.

Daily performance data for the positions is obtained from the investment management systems which are directly connected to market data providers, such as Bloomberg and Reuters. Daily return distributions for each portfolio are presented in **Appendix 3**.

One issue regarding the performance data has to be recognized at this point. VaR methods are generally unable to capture intra-day trading. In other words, the portfolio is assumed to remain stable during the holding period, which is one day in this case. If there are significant trades made during this period, portfolio returns become "contaminated" and VaR estimates are not directly comparable to profits and losses. (Jorion, 2001) Basel Committee (1996) therefore suggests that banks should develop "uncontaminated" backtests to deal with this issue. In practice, this would mean using *hypothetical* changes is portfolio value that would occur if portfolio was assumed to remain the same.

Intra-day trading is not a significant problem in this empirical study. The portfolios under examination tend to remain relatively stable within the one day period and therefore it is unnecessary to calculate hypothetical returns. Moreover, the procedure of calculating hypothetical returns would be technically too cumbersome to handle in this context.

4.1.3 VaR Calculation

As was discussed earlier, the choice of parameters in VaR calculations is not arbitrary whenever backtesting is involved. To construct a solid view on the validity of the model, relatively low confidence levels should be used. According to Jorion (2001), a confidence level of 95% suits well for backtesting purposes. With this approach, it is possible to observe enough VaR violations within the one year time period. However, since the software in this case yields VaR estimates under different confidence levels without any additional simulation, I am able to use levels of 90%, 95% and 99% and test each one individually. Having more than one level of confidence in the backtesting process makes the testing more effective.

The software allows choosing between two estimation methods: exponentially weighted moving average (EWMA) and ordinary least squares (OLS). In this case, I choose to use EWMA which is a more useful estimation method in forecasting financial phenomena as it puts more weight to recent market developments. In addition, the user can define the weight that will be used in the estimation process. **Figure 4** displays three different weights for EWMA estimation.

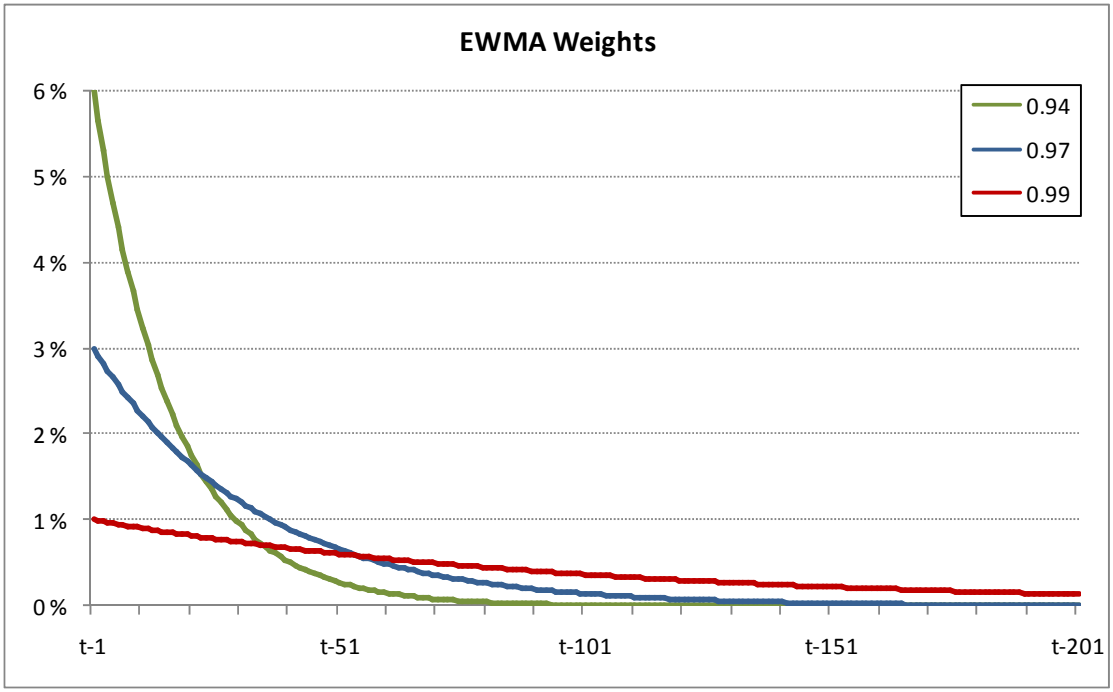


Figure 4: EWMA weights

In this study I will set the EWMA decay factor at 0.94, according to RiskMetrics (1996) recommendation for daily data (Jorion, 2001). Compared to using a decay factor of 0.97 (which probably will be normally used in the Company as the VaR calculations are based on longer time horizons), this method has relatively more emphasis on recent developments of market prices. It is important to realize that the choice of this parameter has a significant effect on the outcome of the estimation. For example, using a decay factor 0.94 leads to a situation where the last observation (t-1) is given a 6% weight and an observation one month ago (t-21) only 1.74% weight. As can be seen from the above figure, observations over 2 or 3 months ago have very little effect on the outcome of the estimation. In practice, if the market experiences

sudden jumps in volatility, VaR estimates react faster to these changes when using a lower decay factor. The downside of a low decay factor is that the short period does not necessarily capture all the potential events that should be included in the estimation process. There are no standards determining the 'correct' decay factor, as the recommendation by RiskMetrics was also found out by empirically testing different factors. It may even be that some other choice of decay factor would perform better in the current market environment.

The VaR estimates are obtained by computing daily VaR levels for a time period of one year, ranging from December 3th 2007 to November 26th 2008. The number of trading days (observations) totals 250, which is enough to produce some statistically significant backtests and is as well in line with the Basel backtesting framework.[15] Simulation rounds are set to 10 000, which is the most that can be used under these circumstances but still should be enough to obtain fairly accurate estimates.

4.1.4 Backtesting Process

Figure 5 below illustrates the backtesting process. After calculating daily profits and losses and simulating VaR estimates, it is time to perform the actual backtests.

Throughout the backtesting process daily trading outcomes are compared to daily VaR estimates**.** Let $x_{t,t+1}$ denote profit or loss of the portfolio over one day time interval. Corresponding VaR estimate is then defined as VaR_t, which is calculated at the beginning of the period, i.e. using the closing prices of day t. For example, the first VaR estimate is calculated with the closing prices of 3rd of December. This estimate is

[15] The time period used in this study includes the last twelve months of data (at the time of the calculation). In an ideal situation we would prefer to have a significantly larger data set. The VaR estimates are calculated using a historical data of one year, with larger weights on recent market developments. Since the database of the software does not include historical data beyond 2007, we do not have a chance to calculate VaR estimates for positions in early 2007, for instance. Technically it would be possible to feed the historical data into the system but in this context the amount of additional work is so extensive that I do not consider it to be worth the effort.

then compared to the trading outcome (profit or loss) that is realized at the end of 4th of December.

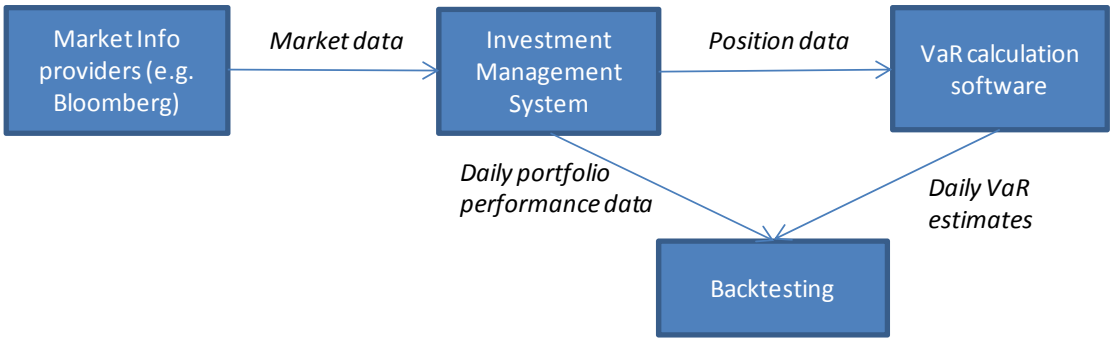


Figure 5: VaR calculation and backtesting process

Table 3 presents the results of consolidating the performance data with corresponding VaR estimates. The column that mainly draws the attention here is the observed number of exceptions. On the basis of this data only we can already perform several different backtests.

	Confidence Level	Number of Observations	Expected Number of Exceptions	Observed Number of Exceptions	Time Until First Exception
Top Portfolio					
	99 %	250	2.5	10	70
	95 %	250	12.5	25	23
	90 %	250	25	36	23
Equity Portfolio					
	99 %	250	2.5	10	9
	95 %	250	12.5	33	1
	90 %	250	25	50	1
Bond Portfolio					
	99 %	250	2.5	7	33
	95 %	250	12.5	18	3
	90 %	250	25	30	3
Equity Option Portfolio					
	99 %	236	2.36	12	33
	95 %	236	11.8	20	2
	90 %	236	23.6	29	2

Table 3: Backtesting data

4.2 Backtests

As was discussed earlier, single backtests can never be enough to evaluate goodness of a VaR model. If one test yields a decent outcome, the result should always be confirmed with another test. (Haas, 2001) Following this argument, I will apply Basel framework and Kupiec's POF-test to examine the frequency of exceptions, as well as Christoffersen's interval forecast test and mixed Kupiec-test to study the independence of exceptions. These tests represent a fairly traditional approach to backtesting since they can be applied virtually in every case where VaR figures are computed. Using these tests requires only the number of total observations, number of VaR violations and the time when the violations occur.

Percentile test by Crnkovic and Drachman (1997) will not applied because the framework is based on testing the whole distribution, and this data is not available in this case. Also the loss function approach by Lopez (1999) will not be used. The reason for this is that the method requires strong assumptions about the stochastic behavior and distribution of profits and losses, and this is something I do not want to do in this case. Moreover, Lopez's approach is more suited to comparing different VaR models, rather than to statistically classify a model either good or bad.

4.2.1 Frequency of Exceptions

Basel Traffic Light Approach

The Basel backtesting framework applies only to banks. It should be noted that the Company is not involved in banking business and therefore is not obliged to conduct any kind of official backtesting for regulatory purposes. Nevertheless, the Basel framework provides a useful exercise as a preliminary test before moving towards statistical hypothesis-based backtests.

Basel backtesting framework uses 99% confidence level and a period of 250 trading days. With these settings daily returns are expected to exceed VaR estimates 2.5 times

on average. According to Basel Committee (1996), an accurate model would fall into green zone with 0 to 4 exceptions. Yellow zone, which consists of 5 to 9 exceptions, indicates a potential problem with the model. If 10 or more exceptions are observed, the model falls into the red zone, and this should generally lead to an automatic assumption that the model is false.

From purely statistical point of view, model validation should be conducted with lower confidence levels (e.g. Jorion, 2001). Therefore we ought to find out the cut-off points for other confidence levels as well. Recalling that the yellow zone begins at the cumulative probability of 95% and the red zone begins at 99.99% (see **Table 2**), we can utilize the binomial distribution to calculate the cut-off points for confidence levels 95% and 90% with 250 observations:

Zone	90 %	95 %	99 %
Green Zone	0 - 32	0 - 17	0 - 4
Yellow Zone	33 - 43	18 - 26	5 - 9
Red Zone	44 or more	27 or more	10 or more

Since the equity option portfolio includes only 236 observations, the corresponding values are slightly different:

Zone	90 %	95 %	99 %
Green Zone	0 - 30	0 - 17	0 - 4
Yellow Zone	31 - 41	18 - 25	5 - 9
Red Zone	42 or more	26 or more	10 or more

It should be recognized that outcomes close to zero at lower confidence levels also indicate a problem within the model even though the green zone represents an accurate model. For example, if we observed zero exceptions at 90% level over 250 days, we would define the model to be overly conservative and in fact quite useless. However, since regulators are only interested in identifying models that systematically *under*estimate risk, these outcomes, even if clearly false, are acceptable from regulators' point of view.

The results of fitting our data into the three classes of Basel traffic light approach are displayed in **Table 4**. The chart suggests severe underestimation of risk in majority of

cases, most notably in equities. The bond portfolio performs best as it is the only portfolio avoiding the red zone.

	Confidence Level	Number of observations	Number of Exceptions	Test Outcome
Top Portfolio				
	99 %	250	10	Red Zone
	95 %	250	25	Yellow Zone
	90 %	250	36	Yellow Zone
Equity portfolio				
	99 %	250	10	Red Zone
	95 %	250	33	Red Zone
	90 %	250	50	Red Zone
Bond portfolio				
	99 %	250	7	Yellow Zone
	95 %	250	18	Yellow Zone
	90 %	250	30	Green Zone
Equity option portfolio				
	99 %	236	12	Red Zone
	95 %	236	20	Yellow Zone
	90 %	236	29	Green Zone

Table 4: Basel traffic light test results

At 99% confidence level the model produces the most worrying results. For example, for the top portfolio the VaR model generated 10 exceptions out of 250 observations. As we already know, there is only a very small probability (less than 0.01%) that an accurate model with a correct coverage of 99% would produce as much as 10 or more exceptions. The results at the 99% confidence level therefore raise a concern whether the model is able to estimate extreme tail losses with enough precision.

Despite these quite alarming results, not too hasty conclusions should be drawn based on this test only. More comprehensive analysis is required in order to judge the quality of the model.

Kupiec Tests

Kupiec's POF-test is used in this case to examine whether the amount of exceptions is too large in statistical terms, as was suggested by the Basel traffic light approach. Although the number of observations is limited to one year, the POF-test should yield some significant results, especially with lower confidence levels.

The test statistics for each portfolio and confidence level are calculated by plugging the data (number of observations, number of exceptions and confidence level) into the test statistic function:

$$LR_{POF} = -2ln\left(\frac{(1-p)^{T-x}p^{x}}{\left[1-\left(\frac{x}{T}\right)\right]^{T-x}\left(\frac{x}{T}\right)^{x}}\right)$$

Throughout the backtesting process I will use 95% percentile of the χ^2 distribution (**Appendix 2**) as the critical value for all the likelihood-ratio tests. This means that reasonably strong evidence is required in order to reject the model.

As an example, consider the top portfolio for which we observed 36 exceptions at 90% confidence level over 250 trading days. The Basel traffic light approach indicated a result in the yellow zone. The corresponding LR-statistic is calculated as

$$LR_{POF} = -2ln\left(\frac{(1-0.10)^{250-36}0.10^{36}}{\left[1-\left(\frac{36}{250}\right)\right]^{250-36}\left(\frac{36}{250}\right)^{36}}\right) \approx 4.80$$

Compared to the critical value of 3.84, the test statistic is slightly larger and the model is rejected. By calculating the statistics for the other portfolios and confidence levels with similar fashion, we obtain results in **Table 5** below.

	Confidence Level	Kupiec's POF-Test Test statistic LR_{POF}	Critical Value $\chi^2(1)$	Test Outcome
Top Portfolio				
	99 %	12.96	3.84	Reject
	95 %	10.33	3.84	Reject
	90 %	4.80	3.84	Reject
Equity portfolio				
	99 %	12.96	3.84	Reject
	95 %	24.89	3.84	Reject
	90 %	22.20	3.84	Reject
Bond portfolio				
	99 %	5.50	3.84	Reject
	95 %	2.26	3.84	Accept
	90 %	1.05	3.84	Accept
Equity option portfolio				
	99 %	20.15	3.84	Reject
	95 %	5.01	3.84	Reject
	90 %	1.29	3.84	Accept

Table 5: POF-Test Results

For all confidence levels the test more or less confirms the results obtained from the traffic light approach. This is the obvious outcome, since the Basel traffic light framework is directly derived from the failure rate test. All portfolios perform poorly, and in some cases the critical value is exceeded with a very large margin. Equity portfolio performs the worst, whereas the fixed income portfolio is the only one to avoid the most severe underestimation of risk. Despite the fact that POF-test has been criticized for having low statistical power in distinguishing bad models from good ones, the results can be considered to be fairly reliable with one year of data and lower confidence levels of 95% and 90%.

As an additional backtest for failure rates I also conducted Kupiec's TUFF-test, for which the results are presented in **Appendix 4**. Since the statistical significance of this test is very limited, no conclusions regarding the quality of a VaR model should be drawn from it. On the contrary, the test generates quite misleading outcomes compared to POF-test and the Basel traffic light approach.

4.2.2 Independence of Exceptions

Christoffersen's Independence Test

Failure rate tests suggested that the VaR model understates risk, especially for equity and equity option portfolios. To examine whether the exceptions are spread evenly over time or are they occurring in clusters, I conduct Christoffersen's interval forecast test. However, at this point I will calculate only the LR_{ind} -statistics in order to focus on the independence property:

$$LR_{ind} = -2ln\left(\frac{(1-\pi)^{n_{00}+n_{10}}\pi^{n_{01}+n_{11}}}{(1-\pi_0)^{n_{00}}\pi_0{}^{n_{01}}(1-\pi_1)^{n_{10}}\pi_1{}^{n_{11}}}\right)$$

As an example, consider again the top portfolio at 90% confidence level. The contingency table can be presented as follows:

	$I_{t-1} = 0$	$I_{t-1} = 1$	
$I_t = 0$	186	28	214
$I_t = 1$	28	8	36
	214	36	250

In addition, we need to solve the probabilities π_0, π_1 and π:

$$\pi_0 = \frac{n_{01}}{n_{00}+n_{01}} = \frac{28}{186+28} = 13.08\%$$

$$\pi_1 = \frac{n_{11}}{n_{10}+n_{11}} = \frac{8}{28+8} = 22.22\%$$

$$\pi = \frac{n_{01}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}} = \frac{28+8}{186+28+28+8} = 14.40\%$$

Plugging this data into the likelihood-ratio statistic we obtain the test value:

$$LR_{ind} = -2ln\left(\frac{(1-0.144)^{186+28} * 0.144^{28+8}}{(1-0.1308)^{186} * 0.1308^{28} * (1-0.2222)^{28} * 0.2222^{8}}\right) \approx 1.88$$

The critical value is the 95% percentile of the χ^2 distribution with one degree of freedom, 3.84. As the test statistic value remains below the critical value, the model is accepted.

Table 6 shows the input data for calculating the LR_{ind}-statistics for each portfolio and confidence level.

		Backtesting Data for the Independence Test							
		Number of Exceptions	n_{00}	n_{01}	n_{10}	n_{11}	π_0	π_1	π
Top Portfolio									
	99 %	10	230	10	10	0	4.2 %	0.0 %	4.0 %
	95 %	25	204	21	21	4	9.3 %	16.0 %	10.0 %
	90 %	36	186	28	28	8	13.1 %	22.2 %	14.4 %
Equity portfolio									
	99 %	10	230	10	10	0	4.2 %	0.0 %	4.0 %
	95 %	33	188	29	29	4	13.4 %	12.1 %	13.2 %
	90 %	50	161	39	39	11	19.5 %	22.0 %	20.0 %
Bond portfolio									
	99 %	7	236	7	7	0	2.9 %	0.0 %	2.8 %
	95 %	18	215	17	17	1	7.3 %	5.6 %	7.2 %
	90 %	30	195	25	25	5	11.4 %	16.7 %	12.0 %
Equity option portfolio									
	99 %	12	227	11	11	1	4.6 %	8.3 %	5.1 %
	95 %	20	213	17	17	3	7.4 %	15.0 %	8.5 %
	90 %	29	199	22	22	7	10.0 %	24.1 %	12.3 %

Table 6: Data for the independence test: n_{ij} is the number of days where state *j* occurred conditional on state *i* occurring on the previous day (0 = no exception, 1 = exception). Thus, n_{11} presents the number of consecutive exceptions. π_i is the probability of an exception assuming a state *i* on the previous day, and π is the probability of an exception regardless of the previous day's state. The probabilities are calculated from the observed data.

Results are displayed in **Table 7**. Apart from the equity portfolio at 90% confidence level, no dependence between exceptions according to Christoffersen's test occurs, at least not in statistically significant terms.

	Independence Test (Christoffersen's Test)		
	Test Statistic LR_{ind}	Critical Value $\chi^2(1)$	Test Outcome
Top Portfolio			
99 %	0.83	3.84	Accept
95 %	0.98	3.84	Accept
90 %	1.88	3.84	Accept
Equity portfolio			
99 %	0.83	3.84	Accept
95 %	0.04	3.84	Accept
90 %	0.15	3.84	Accept
Bond portfolio			
99 %	0.40	3.84	Accept
95 %	0.08	3.84	Accept
90 %	0.65	3.84	Accept
Equity option portfolio			
99 %	0.33	3.84	Accept
95 %	1.27	3.84	Accept
90 %	4.25	3.84	Reject

Table 7: Christoffersen's independence test results

Despite the good outcome of the independence test, no instant conclusions should be drawn from it. Nevertheless, it is fair to say that for the most part the model seems to avoid at least the most severe type of independence, namely multiple exceptions occurring on consecutive days.

Independence Test of the Mixed Kupiec-Test

The problem with the Christoffersen's independence test is that it considers only two successive observations. As was already previously discussed, the test has low power in capturing dependence between exceptions since it effectively ignores all other forms of dependence. To overcome this shortcoming, I will use the independence test suggested by Haas (2001). The test statistic for each exception is:

$$LR_i = -2ln\left(\frac{p(1-p)^{v_i-1}}{\left(\frac{1}{v_i}\right)\left(1-\frac{1}{v_i}\right)^{v_i-1}}\right)$$

For the sake of simplicity, let us use the top portfolio at 99% confidence level as an example, instead of the 90% level in previous tests. At 99% confidence level we observed 10 exceptions for the top portfolio with the following durations measured between the exceptions:

	Number of the exception									
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Time between the exceptions	70	21	23	15	14	31	4	13	21	7

Inserting this data into the function above, we can calculate the independence statistics for every exception:

	Number of the exception									
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
LR-statistic	0.11	1.57	1.43	2.14	2.27	0.98	4.77	2.40	1.57	3.59

Summing up the *LR*-statistics we obtain:

$$LR_{ind} = \sum_{i=2}^{n}\left[-2ln\left(\frac{p(1-p)^{v_i-1}}{\left(\frac{1}{v_i}\right)\left(1-\frac{1}{v_i}\right)^{v_i-1}}\right)\right] - 2ln\left(\frac{p(1-p)^{v-1}}{\left(\frac{1}{v}\right)\left(1-\frac{1}{v}\right)^{v-1}}\right) \approx 20.83$$

The test statistic is distributed as χ^2 with n degrees of freedom equal to the number of exceptions, 10. The critical value at 95% percentile is 18.31. Because LR_{ind} exceeds the critical value, the model is rejected. This indicates that the independence property is not satisfied.

Test results for other portfolios and confidence levels are displayed in **Table 8** below. Critical values are not the same in every case because the value is determined according to the number of exceptions in each case. The most important thing to notice here is that the test outcomes differ significantly from the Christoffersen's

independence test. While the Christoffersen's test accepted all but one of the test statistics, this test suggests that only the fixed income portfolio produces exceptions that are independent of each other.

		Independence Test (Mixed Kupiec-Test)		
		Test statistic LR_{ind}	Critical Value $\chi^2(x)$	Test Outcome
Top Portfolio				
	99 %	20.83	18.31	Reject
	95 %	46.54	37.65	Reject
	90 %	59.96	51.00	Reject
Equity portfolio				
	99 %	25.30	18.31	Reject
	95 %	67.98	47.40	Reject
	90 %	89.03	67.50	Reject
Bond portfolio				
	99 %	10.76	14.07	Accept
	95 %	19.92	28.87	Accept
	90 %	36.48	43.77	Accept
Equity option portfolio				
	99 %	43.79	21.03	Reject
	95 %	42.32	31.41	Reject
	90 %	51.43	42.56	Reject

Table 8: Results of the independence test of mixed Kupiec-test

4.2.3 Joint Tests of Unconditional Coverage and Independence

Christoffersen's Interval Forecast Test

Now that we have conducted the tests for coverage and independence separately, the Kupiec's POF-test and Christoffersen's independence test can be combined into a joint test of conditional coverage. The test statistic can be derived directly from the results of the previous backtests as follows:

$$LR_{cc} = LR_{POF} + LR_{ind}$$

Now the LR-statistic has a critical value with two degrees of freedom, 5.99. Again, consider the top portfolio at 90% confidence level. The result from the POF-test was 4.80, and the result from the Christoffersen's independence test was 1.88. Summing up these two statistics we obtain the test value for the conditional coverage test:

$$LR_{cc} = 4.80 + 1.88 \approx 6.69$$

which slightly exceeds the critical value of 5.99, resulting in rejection of the model.

Table 9 presents the joint test results. Since we already know that the POF-test produced results where critical values were exceeded significantly, the results from the joint test are not surprising. Also in this case the fixed income portfolio performs best as it passes the test at all confidence levels.

		Joint Test of Unconditional Coverage and Independence		
		Test statistic LR_{cc}	Critical Value $\chi^2(2)$	Test Outcome
Top Portfolio				
	99 %	13.79	5.99	Reject
	95 %	11.30	5.99	Reject
	90 %	6.69	5.99	Reject
Equity portfolio				
	99 %	13.79	5.99	Reject
	95 %	24.93	5.99	Reject
	90 %	22.35	5.99	Reject
Bond portfolio				
	99 %	5.90	5.99	Accept
	95 %	2.34	5.99	Accept
	90 %	1.70	5.99	Accept
Equity option portfolio				
	99 %	20.48	5.99	Reject
	95 %	6.28	5.99	Reject
	90 %	5.54	5.99	Accept

Table 9: Joint test results

Mixed Kupiec-Test

Mixed Kupiec-test can be considered to be more informative and reliable than the joint test by Christoffersen, since the mixed test is capable of capturing more general forms of dependence between exceptions instead of just two consecutive days.

Similarly to the Christoffersen's interval forecast test above, the mixed Kupiec-test can also be conducted in a straightforward fashion since we already have the results of the POF-test and the independence test (of mixed Kupiec-test):

$$LR_{mix} = LR_{POF} + LR_{ind}$$

For the example case of top portfolio at 90% confidence level the LR_{cc} –statistic is calculated as:

$$LR_{mix} = 4.80 + 59.96 \approx 64.76$$

Now the critical value is the 95% percentile of χ^2 distribution with $n+1$ degrees of freedom, where n is the number of exceptions. Thus, the critical value of 52.19 in this case is obtained from χ^2 distribution with 37 degrees of freedom. Due to the clear violation of the critical value, the model is again rejected. This outcome was to be expected, as the failure resulted also from the Christoffersen's test which is statistically weaker than the mixed Kupiec-test.

All the results of the mixed Kupiec-test are displayed in **Table 10**. Apart from the fixed income portfolio at lower levels of confidence, the model is rejected. The results are worrying by anyone's standards.

		Mixed Kupiec-Test		
		Test statistic LR_{mix}	Critical Value $\chi^2(1)$	Test Outcome
Top Portfolio				
	99 %	33.79	19.68	Reject
	95 %	56.87	38.89	Reject
	90 %	64.76	52.19	Reject
Equity portfolio				
	99 %	38.26	19.68	Reject
	95 %	92.88	48.60	Reject
	90 %	111.23	68.67	Reject
Bond portfolio				
	99 %	16.25	15.51	Reject
	95 %	22.18	30.14	Accept
	90 %	37.53	44.99	Accept
Equity option portfolio				
	99 %	63.94	22.36	Reject
	95 %	47.32	32.67	Reject
	90 %	52.72	43.77	Reject

Table 10: Mixed Kupiec-test results

4.3 Evaluation of Backtesting Results

4.3.1 Equity Portfolio

In 2008, equity markets all over the world performed poorly. The market was characterized by high volatility, especially in autumn 2008 when macroeconomic events seemed to affect stock prices more than company-specific news. In bear markets it is not unusual that correlations between equities increase. For example, Longin and Solnik (2001) and Campbell et al. (2002), among others, have shown evidence of significant increased correlation in equity returns during bear markets. This kind of sudden increase in correlations makes it very challenging to forecast future portfolio performance since all VaR models rely on historical market data in one way or another.

The economic downturn also affected the test portfolio which consisted of quoted Finnish equities. Over the 250-day period, the portfolio return was -36%. This performance is quite well in line with overall market performance, as can be seen from **Figure 6**.

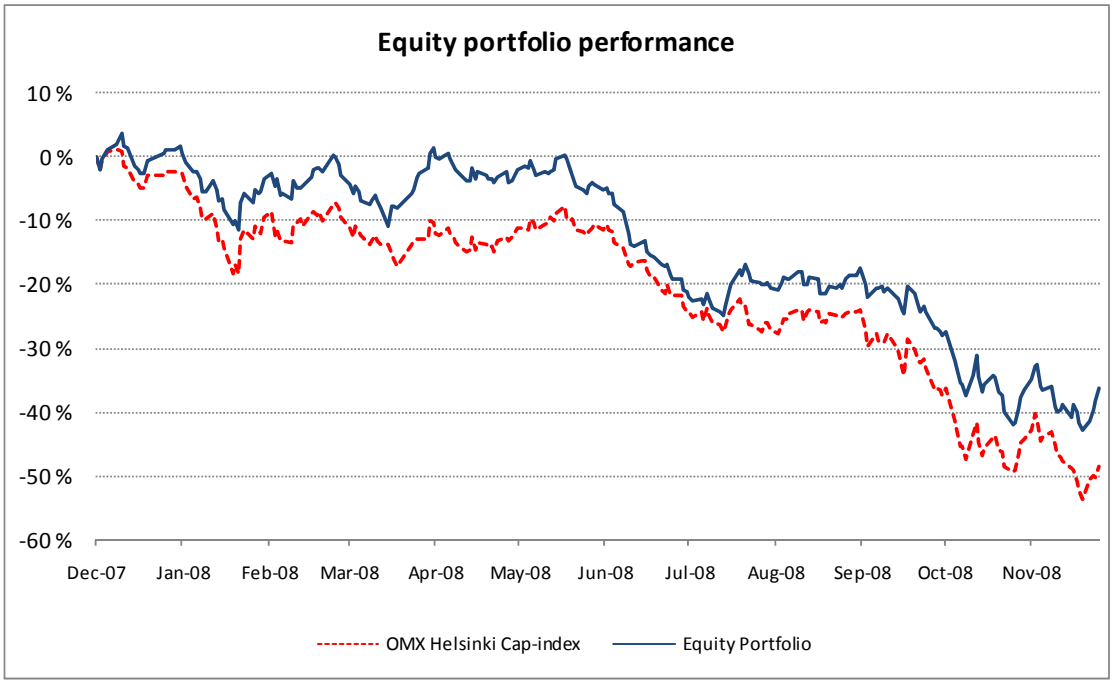
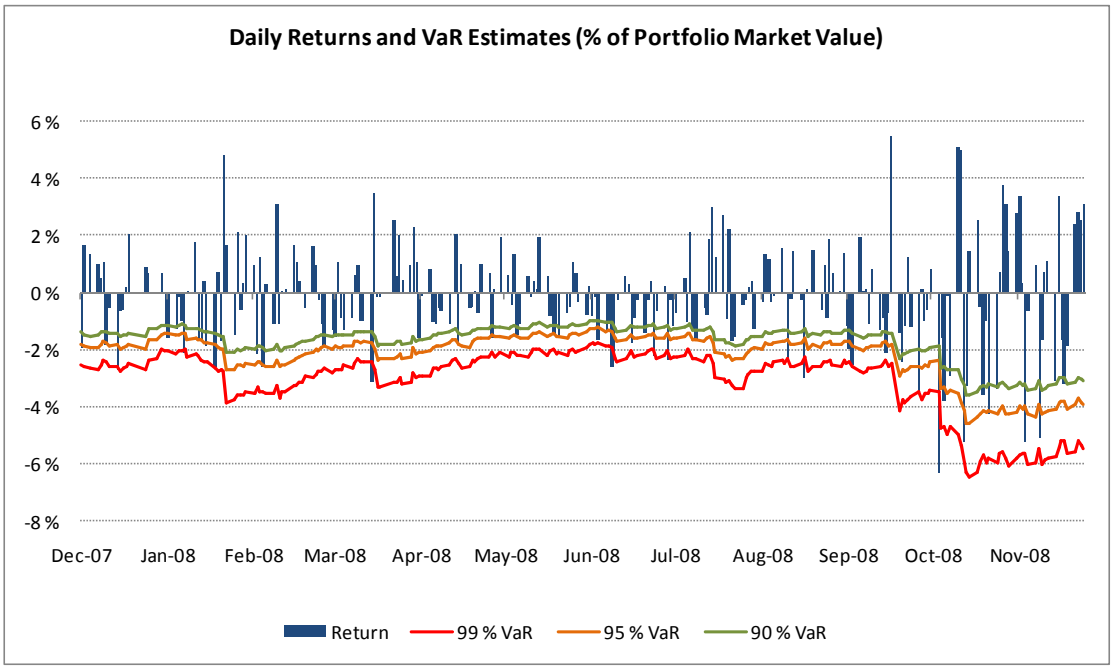


Figure 6: Equity portfolio performance

Figure 7 presents a summary of the backtesting results and a graph where daily returns are displayed with VaR estimates at three different confidence levels over the one year time period. It is advisable to present backtesting results in this manner since graphical illustration provides a very good overall view of the results already at the first look. For instance, we can see that the unusually high volatility, and perhaps increased correlation, is reflected in portfolio returns especially during the last two or three months of the observation period.



Confidence Level	Exceptions / Observations	Frequency Tests			Independence Tests		Joint Tests	
		Traffic Light	TUFF-test	POF-test	Christof-fersen	Mixed Kupiec	Christof-fersen	Mixed Kupiec
99 %	10 / 250	Red Zone	Accept	Reject	Accept	Reject	Reject	Reject
95 %	33 / 250	Red Zone	Reject	Reject	Accept	Reject	Reject	Reject
90 %	50 / 250	Red Zone	Reject	Reject	Accept	Reject	Reject	Reject

Figure 7: Backtesting results for equity portfolio

The backtesting results are very poor for the equity portfolio. VaR estimates at all of the tested confidence levels are violated many times more than what was expected. For instance, at 95% confidence level we observed 33 exceptions, which is nearly three times more than the expected value of 12.5. As a result of this severe systematic underestimation of equity risk, POF-test indicated a rejection at all confidence levels.

Christoffersen's test for independence of exceptions produced decent results. However, since the test does not capture dependence in all forms, we cannot conclude that the exceptions are totally independent. Looking at the results from the other independence test, namely the mixed Kupiec-test, we notice that the model is rejected at all confidence levels. The exceptions thus exhibit some kind of dependence, even though we can see in **Figure 7** that rising volatility of portfolio returns caused also significant increases in VaR estimates. Perhaps laying even more emphasis on recent

market data, i.e. choosing a lower EWMA weight, could result in better outcome in independence tests.

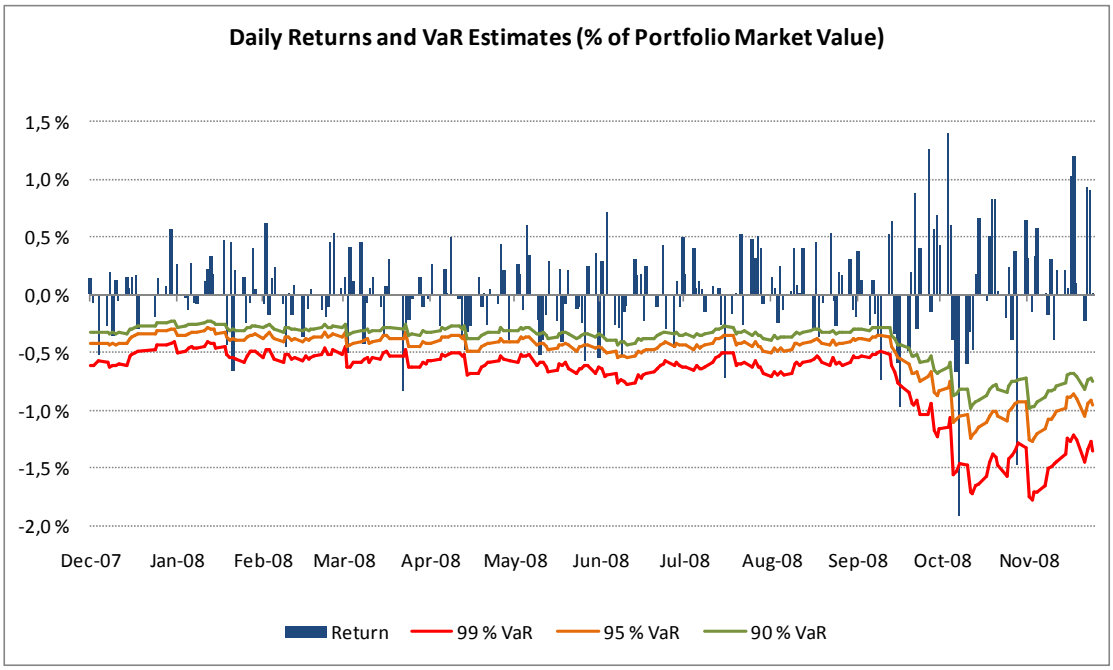
We may conclude that the model does not estimate equity risks with satisfactory precision. As such, it is obvious that the model is rejected. A totally different topic of discussion then is whether we can judge the model to be inadequate or should we accept the fact that the turbulent equity markets of 2008 is simply outside of what any VaR model is able to capture. I will return to this issue later in this chapter.

4.3.2 Fixed Income Portfolio

The fixed income portfolio consists of government bonds, which are usually considered to be very secure investments. In practice, there is no credit risk in government bonds and the only type of risk arises from changes in interest rates. As a matter of fact, despite the fairly low return expectation, government bonds were the best performers in the Company's investment portfolio in 2008. The return of the test portfolio was 9% during the observation period.

Due to significant selling transactions, the portfolio's market value dropped from 1.1 billion to 0.14 billion over the one year observation period. This intra-day trading could have potentially distorted the VaR estimates, but in this case notable problems did not occur.

Out of the three test portfolios used in the study, the fixed income portfolio performs best in terms of backtesting results. The portfolio passes all other tests except the frequency test at 99% confidence level. Seven exceptions are more than expected and the same outcome is also suggested by the Basel traffic light approach. For other confidence levels there are a few exceptions more than expected but still the backtests confirm these outcomes to be acceptable.



Confidence Level	Exceptions / Observations	Frequency Tests			Independence Tests		Joint Tests	
		Traffic Light	TUFF-test	POF-test	Christof-fersen	Mixed Kupiec	Christof-fersen	Mixed Kupiec
99 %	7 / 250	Yellow Zone	Accept	Reject	Accept	Accept	Accept	Reject
95 %	18 / 250	Yellow Zone	Accept	Accept	Accept	Accept	Accept	Accept
90 %	30 / 250	Green Zone	Accept	Accept	Accept	Accept	Accept	Accept

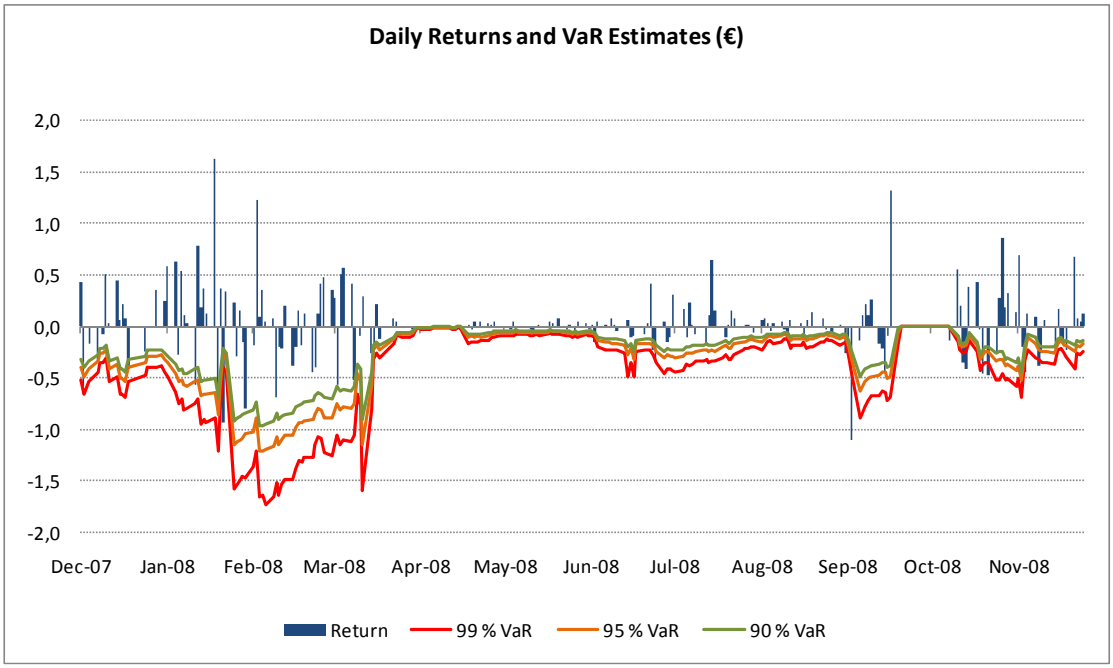
Figure 8: Backtesting results for fixed income portfolio

Portfolio returns and VaR levels remained stable until September when the market started to experience higher volatility and the portfolio lost some of its diversification effect. The higher portfolio volatility in autumn is fairly well captured by the VaR model, as can be seen from **Figure 8**. Some successive VaR violations occurred during this period of rising volatility. In addition, 4 out of 7 exceptions under 99% confidence level took place in just 40 days, which raises some concern about the models ability to estimate interest rate risk. However, both tests of independence produced positive results and suggested acceptance of the model.

As a conclusion, we can argue that the VaR model captures interest rate risk fairly accurately, at least at lower confidence levels. Additional testing with a new set of data or another type of fixed income securities, such as credit bonds or derivatives, still deserves consideration.

4.3.3 Equity Option Portfolio

The VaR model in this case should be able to capture the market risk of any derivative instrument, as long as the properties and pricing models of the instruments are defined correctly in the system. Our test portfolio consists of equity options. Under the period of 250 days there are 14 days when the portfolio is empty, and no VaR figures are computed. Thus, the data set reduces to 236 observations, which should still be enough to obtain statistically significant results, especially with lower confidence levels.



Confidence Level	Exceptions / Observations	Frequency Tests			Independence Tests		Joint Tests	
		Traffic Light	TUFF-test	POF-test	Christof-fersen	Mixed Kupiec	Christof-fersen	Mixed Kupiec
99 %	12 / 236	Red Zone	Accept	Reject	Accept	Reject	Reject	Reject
95 %	20 / 236	Yellow Zone	Accept	Reject	Accept	Reject	Reject	Reject
90 %	29 / 236	Green Zone	Accept	Accept	Reject	Reject	Accept	Reject

Figure 9: Backtesting results for equity option portfolio

Failure rate test yields quite interesting results. At 90% confidence level the number of exceptions totals 29, which is an acceptable amount as we would expect on average 24 exceptions. At 95% level with 20 exceptions the model is only slightly rejected.

But on the contrary, 99% level produces 12 exceptions, which is approximately five times the expected amount.

The equity option portfolio is also the only one of the three subportfolios to fail the Christoffersen's test for independence. At 90% confidence level the test indicates rejection of the model. On the other hand, the independence statistics of mixed Kupiec-test suggest rejection of the model at all confidence levels. Joint tests also produce negative results. It is therefore easy to say that a problem exists in the model.

The problem in the model is at least partly caused by the calculation of volatility. The algorithm is unable to compute volatility for some instruments and in these cases the software uses a standard approximation that does not necessarily represent the true volatility. An undeniable evidence of this problem is reflected in the last one and a half months of the observation period, where 6 exceptions at 99% confidence level are observed in just 20 days. This is a result that an accurate model would generate only with an extremely low probability, and it is a clear sign of exception clustering, even though at 99% level the Christoffersen's independence test failed to capture this correlation. Since these closely bunched exceptions did not go unnoticed in the mixed Kupiec-test we can also without hesitation argue that the Christoffersen independence test is occasionally misleading and unreliable.

Strange behavior of VaR levels is also occurring at other times, causing additional VaR violations. This is direct evidence of model risk which is specifically associated with Monte Carlo-based VaR systems.

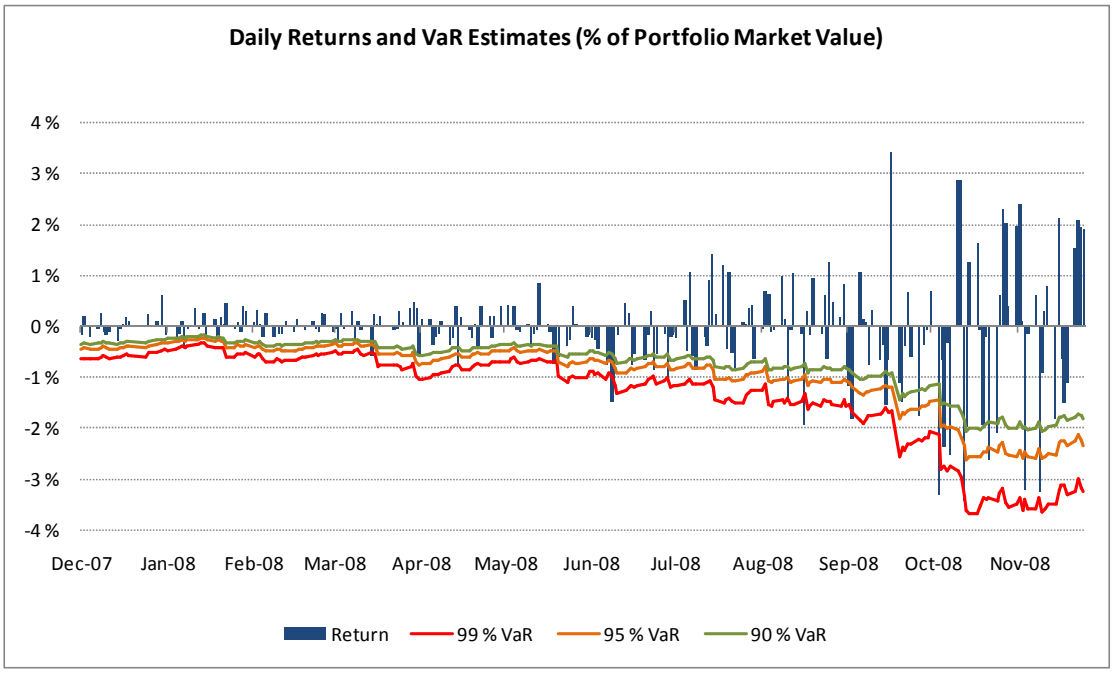
Without the problem of volatility calculation, there is a reason to assume that the model would yield fairly reliable results. Nevertheless, additional backtesting of derivatives, also other than equity options, is strongly recommended in order to identify potential problems. Before this study, the derivative VaR estimates were only compared against static results obtained from an external VaR model. The evidence presented here proves that this kind of model validation is simply inadequate and occasionally even fallacious.

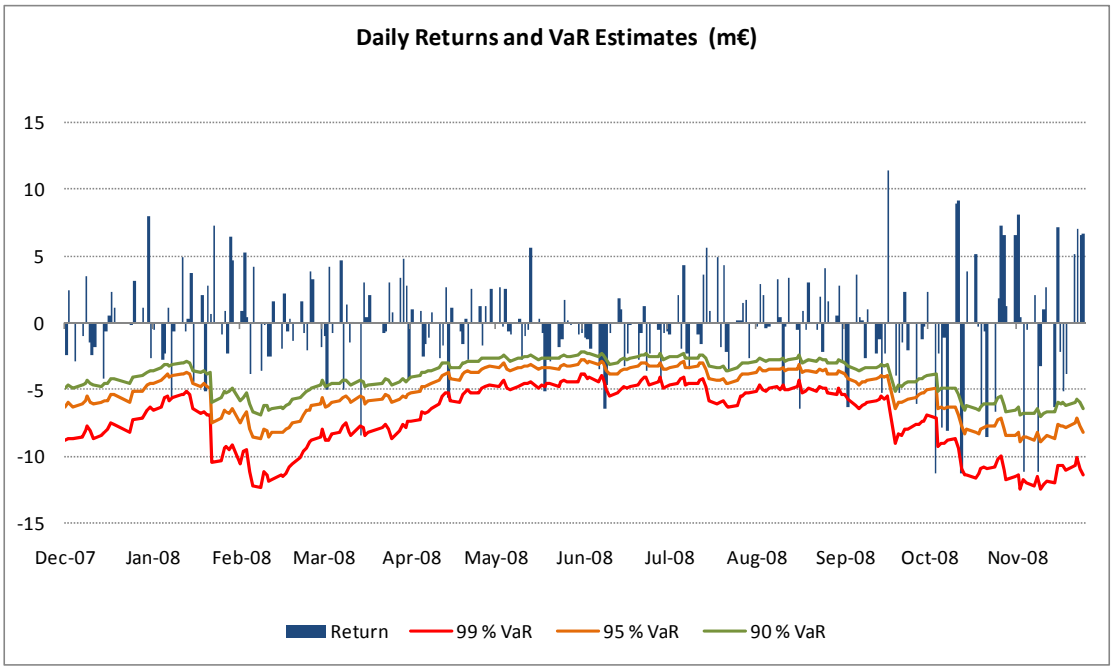
4.3.4 Top Portfolio

Combining the subportfolios under one top portfolio is a useful exercise because it gives some indication about the accuracy of the model when different asset classes are included in the VaR estimates.

Figure 10 displays the VaR levels with daily returns. For the sake of simplicity, the data in this case is presented in two graphs because the allocation and market value of the portfolio changes dramatically during the observation period (due to drop in fixed income portfolio market value). The upper graph shows the returns and the VaR levels as percentages, and the lower one shows the same information in euros.

VaR estimates clearly adapt to changes in portfolio allocation and market volatility, but the model seems to systematically understate risks. This result is to be expected since the magnitude of equity risk underestimation is so large. The positive results of fixed income portfolio are not enough to compensate the bad equity risk estimation. As a consequence, all joint tests reject the model.





Confidence Level	Exceptions / Observations	Frequency Tests			Independence Tests		Joint Tests	
		Traffic Light	TUFF-test	POF-test	Christof-fersen	Mixed Kupiec	Christof-fersen	Mixed Kupiec
99 %	10 / 250	Red Zone	Accept	Reject	Accept	Reject	Reject	Reject
95 %	25 / 250	Yellow Zone	Accept	Reject	Accept	Reject	Reject	Reject
90 %	36 / 250	Yellow Zone	Accept	Reject	Accept	Reject	Reject	Reject

Figure 10: Backtesting results for top portfolio

4.4 Discussion

Backtesting results for all of the test portfolios are presented as a summary sheet in **Appendix 5**. Out of the several backtests conducted in this study, we should focus our attention on those tests that are considered to be the most reliable. Therefore, we ought to be cautious when interpreting backtesting results, for instance, from the TUFF-test by Kupiec.[16] Similarly, one should recognize the shortcomings of Christoffersen's independence test. The most informative and reliable test in this context is the mixed Kupiec-test by Haas.

[16] The purpose of using the TUFF-test in this backtesting process is not to validate our VaR model, but rather to provide evidence of the fact that the test may produce very misleading results compared to other test procedures.

The evidence from the backtests is undeniable; most statistical failure rate tests indicate rejection of the model. The model appears to underestimate risk, especially for equity and equity option portfolios. The fixed income portfolio performs much better in this respect even though at the higher confidence level of 99% there is some signal of underestimation as well.

For the most part, the model avoids the most severe type of dependence between exceptions, namely VaR violations occurring on successive days. However, the mixed Kupiec-test, which should tell us if the model exhibits more general type of dependence, yields worse results. Apart from the fixed income portfolio, the model is rejected at all confidence levels. Hence, because of bad results from both independence and coverage tests, it comes as no surprise that most joint tests of conditional coverage reject the model for equities and equity options.

Despite these somewhat alarming results we should take a minute and consider the big picture. As it is commonly recognized, VaR has been developed to measure portfolio market risk under *normal* market conditions. VaR is known to be fairly accurate during normal market conditions but even a good model may perform poorly if the market suddenly experiences times of high volatility or changes in asset correlations. As Basel Committee (1996) points out, if the market is subjected to a major regime shift, volatilities and correlations may shift substantially. When the economy experiences major macroeconomic shocks, usual correlations may even break down, causing a dramatic change in potential portfolio losses. No VaR model will be immune from this problem since all models rely on past data in predicting future market movements.

The time period under observation was exceptional, at least compared to the previous few years. Equity prices were affected by macroeconomic events more than usually, especially during the autumn of 2008. Also fixed income securities experienced abnormally high volatility but apparently not to the extent that the model would have had major difficulties in capturing interest rate risk. Taking into account these circumstances, it is reasonable to argue that one of the most fundamental assumptions underlying any VaR model, namely the *normal market conditions* (whatever the

definition for ‘normal’ is), does not hold for equity markets during the observation period from December 2007 to November 2008.

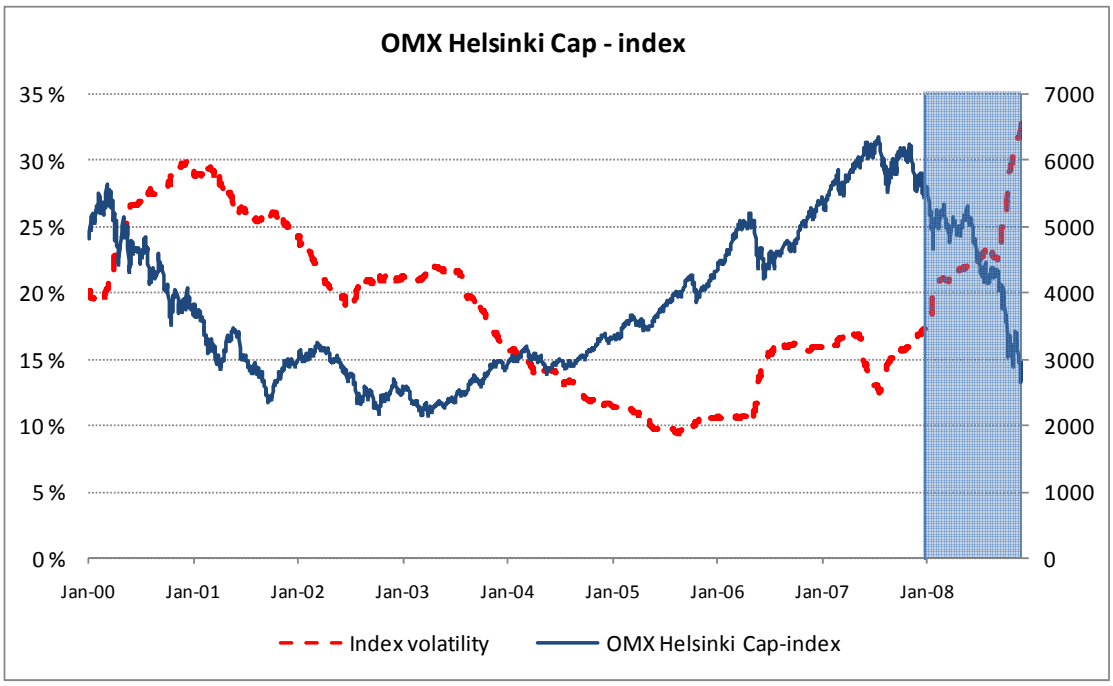


Figure 11: OMX Helsinki Cap-index volatility: The annual volatility is calculated as a rolling volatility from daily index returns. This type of calculation is not directly equivalent to the algorithm of the VaR model since it does not give weights to the historical observations but it still provides a useful illustration of the development of volatility.

To investigate the reasons for the excess amount of observed exceptions a bit closer, consider the OMX Helsinki Cap-index which reflects the performance of our equity test portfolio quite accurately. The annual one year rolling volatility in **Figure 11** gives some indication about the underlying problem. During the one year observation period (on blue background), we can see remarkable increase in market volatility, from 17% in December 2007 to 33% in November 2008. At some points, especially in autumn 2008, this increase is very rapid. In a situation like this it is evident that the VaR estimates, which are simulated on the basis of historical market data, are systematically too low because the ‘true’ volatility at each point of time is higher than the volatility estimated by the model.

The backtesting results of the equity portfolio are direct evidence of this problem. We therefore in one way confirm the claim that VaR works only under normal market

conditions. Note that this is not to say that the model would perform any better in other circumstances, since we do not have evidence to prove this. However, there is a justifiable reason to believe that additional backtests with a new set of data, i.e. in a more stable market environment, would produce better results.

Having concluded that our VaR model clearly does not capture equity risk under exceptional market conditions, there are some general issues that deserve attention. The ongoing financial crisis raises a question whether risk management methods, such as VaR, have failed to describe the prevailing risks adequately. One may argue that at least to some extent risk management models are not as sound as they should be. Some studies regarding the flaws of VaR models have been conducted, for example by Beder (1995), who applied eight different calculation methods to three different portfolios, and found out that VaR estimates differed significantly between the methods. Berkowicz and O'Brien (2002) investigated VaR models of six financial institutions and concluded that the models were too conservative while being inaccurate in capturing changes in volatility. These results indicate that real life VaR models are often inaccurate. Unfortunately, empirical evidence concerning the performance of VaR models under exceptional market conditions is somewhat limited at the moment. It is very likely that the accuracy of VaR models will be a topic of very critical discussion in the near future.

The purpose of this thesis is not to question the validity of VaR as a risk measurement method but it evidently becomes a topic of great interest as we look at the backtesting results presented here and consider the current financial environment. According to Einhorn (2008), the financial crisis has shown that extreme losses have been much more likely than the backtested models predicted. He does not present any direct evidence to back his statement but it would be more or less hair-splitting to claim anything else. Einhorn (2008) goes on to argue:

> *"This (Value-at-Risk) is like an airbag that works all the time except when you have a car accident."* (Einhorn, 2008, p.12)

Of course, this statement is quite extreme but in some sense there is a legitimate point in it. One could indeed, perhaps provocatively, ask what is the purpose of having a

risk management system that performs well only when there is no real danger of extreme events even though that is exactly what VaR should be designed to measure? Specifically, in a situation such as the empirical case of this paper where there *should* be a fairly sophisticated VaR system in place, we would expect backtesting results to be at least decent. Since this clearly is not the case, we have arrived to a situation where we should decide whether to reject our VaR model or just recognize the fact that sudden turbulent market movements are simply beyond of what any VaR system is intended to capture?

The easiest explanation would be to rely on the quoted claim by Einhorn above. However, since we do not have any evidence of the model's performance under normal market conditions and I do not want to draw this kind of conclusion purely on the basis of this data, I recommend further backtesting to be performed.

5. Conclusions

> *"In short, we ought to be able to identify most bad VaR models, but the worrying issue is whether we can find any good ones."* (Dowd, 2006, p.37)

VaR has become one of the most popular methods in measuring market risks. Every VaR model uses historical market data to forecast future portfolio performance. In addition, the models rely on approximations and assumptions that do not necessarily hold in every situation. Since the methods are far from perfect, there is a good reason to question the accuracy of estimated VaR levels.

The theoretical part of this thesis discussed different approaches to computing VaR, and evaluated specifically the shortcomings of these models. Some of the most

common backtests that are being used to measure the accuracy of VaR models were presented. Early tests, such as Kupiec's POF-test or Christoffersen's interval forecast test are statistically weak, at least with insufficient number of observations and high confidence levels. Tests that have been developed more recently are much more powerful and they take into account the dependence between exceptions. As Dowd (2006) points out, the state of the art in backtesting is improving all the time, and the current tests should already be relatively powerful in identifying bad models.

The empirical part of the thesis studied the accuracy of a new Monte Carlo-based VaR model acquired by a Finnish institutional investor. Using three different confidence levels and a data set of one year, I conducted several backtests in order to evaluate the performance of the model. The tests included the Basel traffic light approach, Kupiec's POF-test, Christoffersen's interval forecast test and the mixed Kupiec-test by Haas. The test portfolio consisted of three subportfolios; equities, bonds and equity options.

The outcomes of the backtests provided some indication of potential problems within the system. The results from unconditional coverage tests suggested underestimation of risk, especially for equities and equity options. In addition, a potential flaw regarding the volatility calculation for equity options was discovered. Christoffersen's independence test indicated more positive results. An exception yesterday did not seem to have an effect on whether an exception occurred today or not. However, the mixed Kupiec-test which captures also more general forms of dependence produced a different outcome, suggesting that the exceptions are not totally independent of each other.

The backtesting results raise concerns about the model's ability to estimate equity risk in satisfactory precision. However, the turbulent market of 2008 and especially the rising volatility during the autumn inevitably cause problems in estimating parameters that should describe future market movements. Since all VaR models rely on historical market data, this issue not only concerns the case at hand but VaR systems in general. Abnormal market behavior is simply beyond of what any VaR model is intended to capture. Hence, with this amount of data I do not find it necessary to reject the VaR model even though the evidence against it is very strong. On the contrary,

now that we have discovered that the model is somewhat inaccurate during turbulent markets, we should at least confirm that the model works under normal market conditions.

As a byproduct of the empirical investigation, I provided some evidence that few of the testing frameworks may produce false results and are therefore unable to distinguish good VaR models from bad ones. Specifically, the backtests indicated that the Christoffersen's framework is incapable of capturing exception dependence, at least with sample size of only one year. In addition, the TUFF-test by Kupiec produced very misleading results compared to the POF-test.

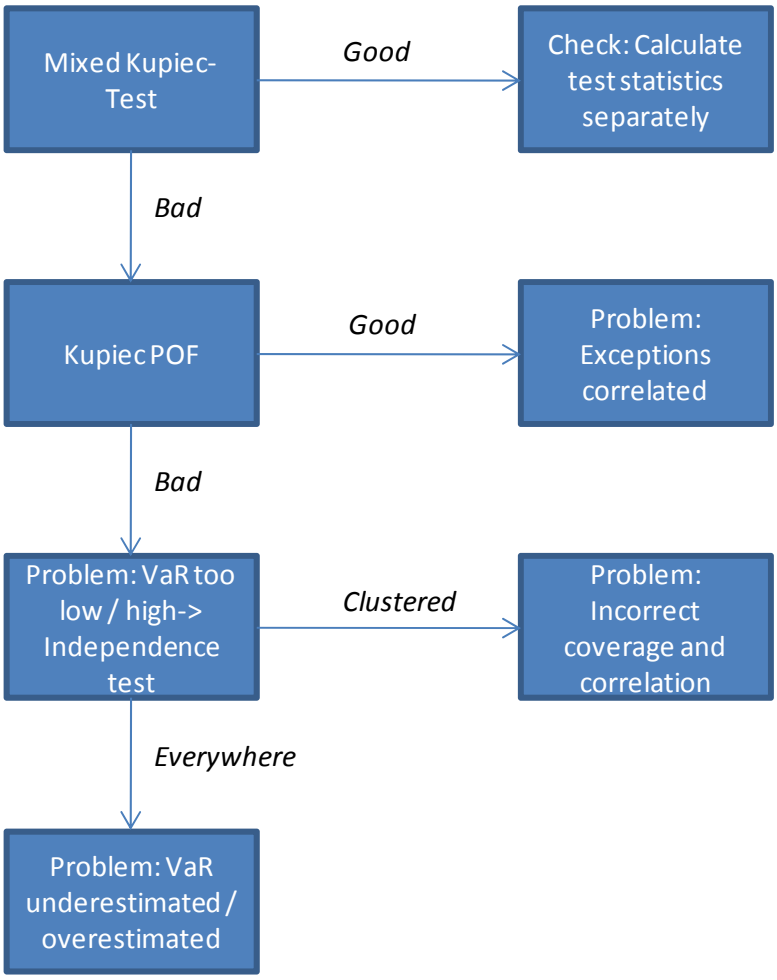
In order to find out whether the model systematically understates risk, addition testing is required. Once the market volatility has settled down, it would be a very useful exercise to reexamine the model's performance with a different and perhaps longer observation period. The problem is, as it is indeed with any kind of backtesting, that we have to wait a long time in order to acquire a new set of data. A minimum of one year, preferably even longer, time horizon is required. Alternatively, one could utilize historical data from 2007 and beyond but there are some technical difficulties associated with this approach.

It ought to be recognized that parameter choices can make a significant difference in the outcome of the results. Therefore, it remains to be tested whether different choice of parameters would yield better results. For instance, using a different decay factor is a potential idea for future testing. In this study, I used EWMA weight of 0.94, which is recommended by RiskMetrics for daily data. However, the unusual circumstances of 2008 could perhaps require a slightly different approach. Laying even more emphasis on recent developments, i.e. using a lower decay factor, would make the model more sensitive to changes in volatilities and correlations.

This empirical study tested only equities, government bonds and equity options. In addition to these instrument classes, we should at some point consider testing other types of instruments as well, such as commodity and interest rate derivatives, floating rate notes and perhaps credit bonds, if possible. However, there are practical difficulties associated also with some of these cases.

Systematic backtesting should be a part of regular VaR reporting in order to constantly monitor the performance of the model. However, the problem is that the inflexibility and slowness of data processing makes it challenging to conduct any regular daily-based backtesting in the Company. Nevertheless, the issue of future backtesting, whether it will be continuous or random testing, ought to be thoroughly considered.

Even though I used several different backtests in this thesis, the purpose is not to apply the wide scale of tests in forthcoming testing. Rather, the focus should be on the most efficient tests and, of course, more than only one single test. When it comes to potential future backtesting processes in the Company, Haas (2001) provides an appealing strategy for optimal backtesting. Following loosely his ideas, one alternative is to use the process below in model validation:



The testing process starts with mixed Kupiec-test. A positive result should be confirmed with separate coverage and independence tests since we know that joint tests may not always detect the violation of these properties alone. Also in the case where the mixed Kupiec-test rejects the model, we should investigate whether the failure is due to incorrect coverage, dependence between exceptions, or both. These statistical tests should be incorporated with visual presentations, such as in this paper.

Whatever the framework for future backtesting will be, the most important lesson to learn from this paper is to understand the weaknesses of VaR calculation. As the empirical research proves, VaR figures should never be considered to be 100 percent accurate, no matter how sophisticated the systems are. However, if the users of VaR know the flaws associated with VaR, the method can be a very useful tool in risk management, especially because there are no serious contenders that could be used as alternatives for VaR.

References

Ammann, M. & Reich, C. (2001), *Value-at-Risk for Nonlinear Financial Instruments – Linear Approximation or Full Monte-Carlo?* University of Basel, WWZ/Department of Finance, Working Paper No 8/01.

Basle Committee of Banking Supervision (1996), *Supervisory Framework For The Use of "Backtesting" in Conjunction With The Internal Models Approach to Market Risk Capital Requirements.* Available at www.bis.org.

Basle Committee of Banking Supervision (2006), *International Convergence of Capital Measurement and Capital Standards – A Revised Framework, Comprehensive Version.* Available at www.bis.org.

Beder, T. (1995) *VAR: Seductive but Dangerous,* Financial Analysts Journal, September / October 1995.

Berkowitz, J. & O'Brien, J. (2002), *How Accurate are Value-at-Risk Models at Commercial Banks?* Journal of Finance, Vol. 5, 2002.

Brown, A. (2008), *Private Profits and Socialized Risk – Counterpoint: Capital Inadequacy,* Global Association of Risk Professionals, June/July 08 issue.

Campbell, R., Koedijk, K. & Kofman, P. (2002), *Increased Correlation in Bear Markets*, Financial Analysts Journal, Jan/Feb 2002, 58, 1.

Campbell, S. (2005), *A Review of Backtesting and Backtesting Procedure,* Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington D.C.

Christofferssen, P. (1998), *Evaluating Interval Forecasts*. International Economic Review, 39, 841-862.

Christofferssen, P. & Pelletier, P. (2004), *Backtesting Value-at-Risk: A Duration-Based Approach.* Journal of Empirical Finance, 2, 2004, 84-108.

Crnkovic, C. & Drachman, J. (1997), *Quality Control* in *VaR: Understanding and Applying Value-at-Risk,* Risk 9, 139-143.

Crouhy, M., Galai, D. & Robert, M. (2000), *Risk Management,* McGraw-Hill Professional.

Damodaran, A. (2007), *Strategic Risk Taking: A Framework for Risk Management*, Pearson Education, New Jersey.

Dowd, K. (1998), *Beyond Value at Risk, The New Science of Risk Management*, John Wiley & Sons, England.

Dowd, K. (2006), *Retrospective Assessment of Value-at-Risk.* Risk Management: A Modern Perspective, pp. 183-202, San Diego, Elsevier.

Einhorn, D. (2008), *Private Profits and Socialized Risk,* Global Association of Risk Professionals, June/July 08 issue.

Finger, C. (2005), *Back to Backtesting,* Research Monthly, May 2005, RiskMetrics Group.

Haas, M. (2001), *New Methods in Backtesting,* Financial Engineering, Research Center Caesar, Bonn.

Hendricks, D. (1996), *Evaluation of Value-at-Risk Models Using Historical Data*, Economic Policy Review, April 1996.

Jorion, P. (2001), *Value at Risk, The New Benchmark for Managing Financial Risk, 2nd Edition*, McGraw-Hill, United States.

Kritzman, M. & Rich D. (2002), *The Mismeasurement of Risk,* Financial Analysts Journal, Vol. 58, No. 3, May/June 2002.

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

Linsmeier, J. & Pearson, N.D. (1996), *Risk Measurement: An Introduction to Value at Risk*, Working Paper 96-04, University of Illinois at Urbana-Champaign.

Longin, F. (2001), *Beyond the VaR*, Journal of Derivatives, Vol. 8, Iss. 4; p. 36, Summer 2001.

Longin, F. & Solnik, B. (2001), *Extreme Correlation of International Equity Markets*, The Journal of Finance, No. 2, April 2001.

Lopez, J. (1998), *Methods for Evaluating Value-at-Risk Estimates*, Economic Policy Review, October 1998, 119-64.

Lopez, J. (1999), *Regulatory Evaluation of Value-at-Risk Models*, Journal of Risk 1, 37-64.

Tsai, K.-T. (2004) *Risk Management Via Value at Risk,* ICSA Bulletin, January 2004.

Wiener, Z. (1999), *Introduction to VaR (Value-at-Risk)*, Risk Management and Regulation in Banking, Kluwer Academic Publishers, Boston.

www.riskmetrics.com

Appendices

Appendix 1: Error Probabilities under Alternative Coverage Levels

Model is accurate		
Exceptions	Coverage = 99%	
(out of 250)	exact	type 1
0	8.1 %	100.0 %
1	20.5 %	91.9 %
2	25.7 %	71.4 %
3	21.5 %	45.7 %
4	13.4 %	24.2 %
5	6.7 %	10.8 %
6	2.7 %	4.1 %
7	1.0 %	1.4 %
8	0.3 %	0.4 %
9	0.1 %	0.1 %
10	0.0 %	0.0 %
11	0.0 %	0.0 %
12	0.0 %	0.0 %
13	0.0 %	0.0 %
14	0.0 %	0.0 %
15	0.0 %	0.0 %

Model is inaccurate: Possible alternative levels of coverage								
Exceptions	Coverage = 98%		Coverage = 97%		Coverage = 96%		Coverage = 95%	
(out of 250)	exact	type 2	exact	type 2	exact	type 2	exact	type 2
0	0.6 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
1	3.3 %	0.6 %	0.4 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
2	8.3 %	3.9 %	1.5 %	0.4 %	0.2 %	0.0 %	0.0 %	0.0 %
3	14.0 %	12.2 %	3.8 %	1.9 %	0.7 %	0.2 %	0.1 %	0.0 %
4	17.7 %	26.2 %	7.2 %	5.7 %	1.8 %	0.9 %	0.3 %	0.1 %
5	17.7 %	43.9 %	10.9 %	12.8 %	3.6 %	2.7 %	0.9 %	0.5 %
6	14.8 %	61.6 %	13.8 %	23.7 %	6.2 %	6.3 %	1.8 %	1.3 %
7	10.5 %	76.4 %	14.9 %	37.5 %	9.0 %	12.5 %	3.4 %	3.1 %
8	6.5 %	86.9 %	14.0 %	52.4 %	11.3 %	21.5 %	5.4 %	6.5 %
9	3.6 %	93.4 %	11.6 %	66.3 %	12.7 %	32.8 %	7.6 %	11.9 %
10	1.8 %	97.0 %	8.6 %	77.9 %	12.8 %	45.5 %	9.6 %	19.5 %
11	0.8 %	98.7 %	5.8 %	86.6 %	11.6 %	58.3 %	11.1 %	29.1 %
12	0.3 %	99.5 %	3.6 %	92.4 %	9.6 %	69.9 %	11.6 %	40.2 %
13	0.1 %	99.8 %	2.0 %	96.0 %	7.3 %	79.5 %	11.2 %	51.8 %
14	0.0 %	99.9 %	1.1 %	98.0 %	5.2 %	86.9 %	10.0 %	62.9 %
15	0.0 %	100.0 %	0.5 %	99.1 %	3.4 %	92.1 %	8.2 %	72.9 %

Source: Basel Committee (1996)

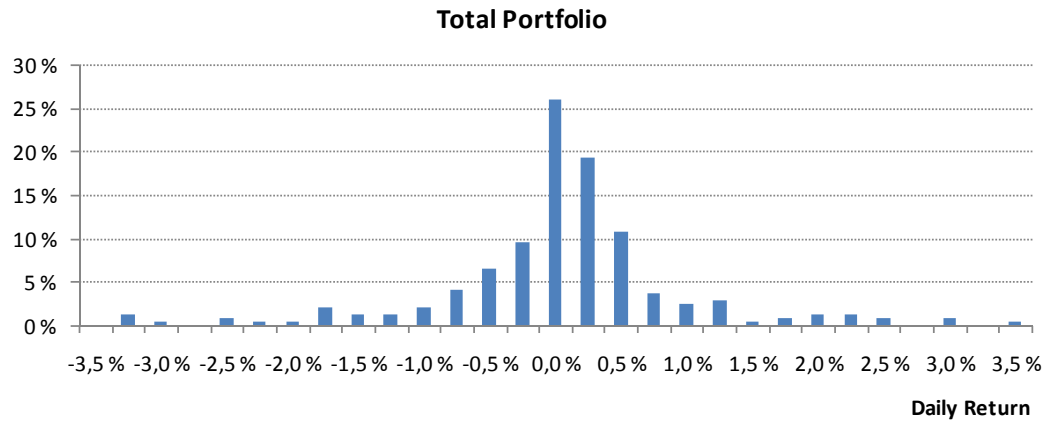
The table presents error probabilities for an accurate model (99% coverage) and for several inaccurate models (98%, 97%, 96% and 95% coverages). The column 'exact' reports the probability of obtaining exactly the stated number of observations in a sample of 250 observations and the columns 'type 1' and 'type 2' report the possibility of committing a type 1 error (rejecting an accurate model) or committing a type 2 error (accepting an incorrect model)

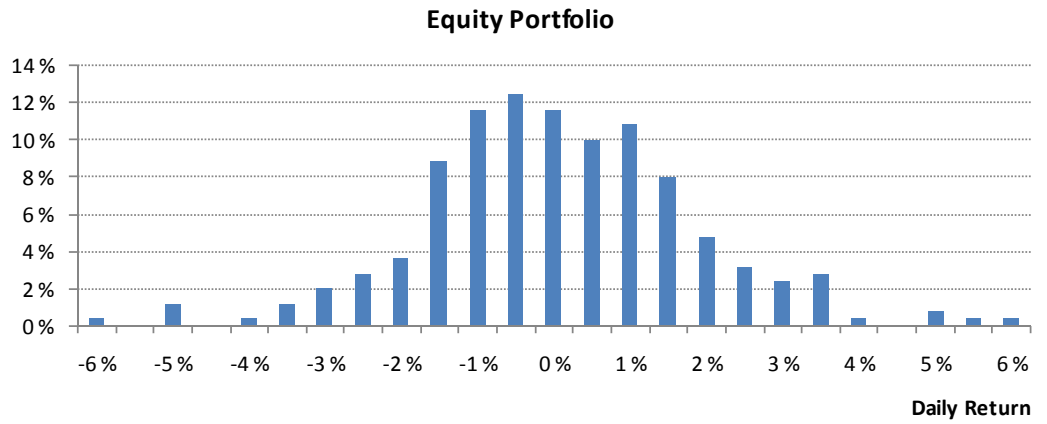
Assuming 250 observations and a reported confidence level of 99%, one would expect an accurate model to yield 2.5 exceptions on average. If the cutoff-point of rejecting a model is set to 5 or more exceptions, there is a 10.8% probability of rejecting an accurate model. On the other hand, suppose that the true coverage of the model is 97%, there is a 12.8% probability of accepting an incorrect model. If the model's coverage is 98%, there is a high probability of 43.9% of accepting a false model. This clearly shows that the Basel framework has relatively low power in distinguishing good models from bad ones.

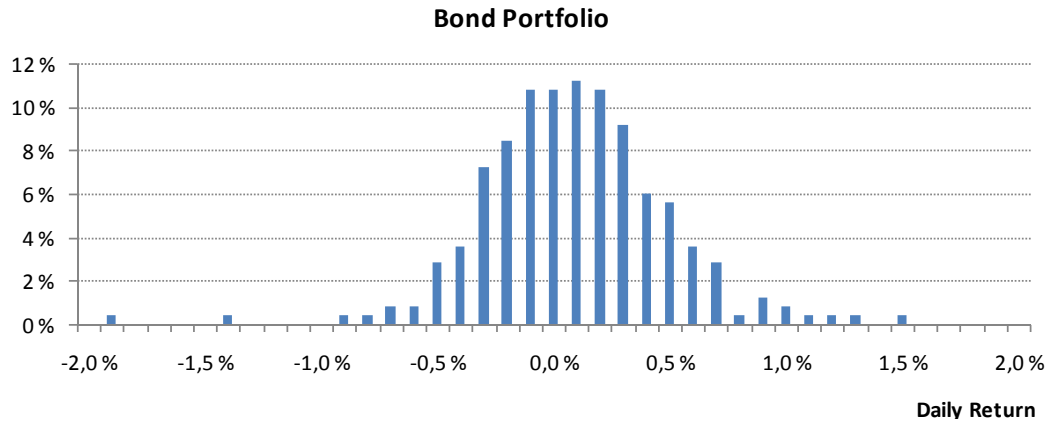
Appendix 2: Critical Values for the Chi-Squared Distribution

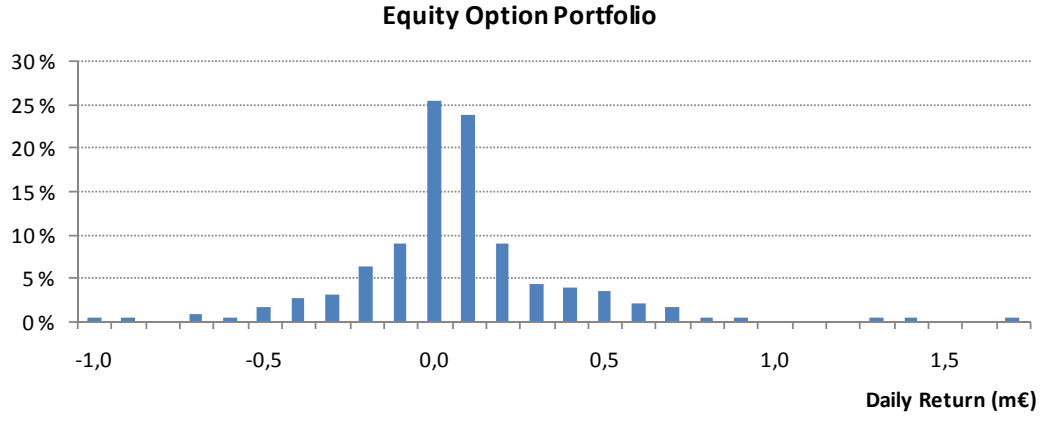
	p value												
f	**0.995**	**0.99**	**0.975**	**0.95**	**0.9**	**0.75**	**0.5**	**0.25**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
31	14.46	15.66	17.54	19.28	21.43	25.39	30.34	35.89	41.42	44.99	48.23	52.19	55.00
32	15.13	16.36	18.29	20.07	22.27	26.30	31.34	36.97	42.58	46.19	49.48	53.49	56.33
33	15.82	17.07	19.05	20.87	23.11	27.22	32.34	38.06	43.75	47.40	50.73	54.78	57.65
34	16.50	17.79	19.81	21.66	23.95	28.14	33.34	39.14	44.90	48.60	51.97	56.06	58.96
35	17.19	18.51	20.57	22.47	24.80	29.05	34.34	40.22	46.06	49.80	53.20	57.34	60.27
36	17.89	19.23	21.34	23.27	25.64	29.97	35.34	41.30	47.21	51.00	54.44	58.62	61.58
37	18.59	19.96	22.11	24.07	26.49	30.89	36.34	42.38	48.36	52.19	55.67	59.89	62.88
38	19.29	20.69	22.88	24.88	27.34	31.81	37.34	43.46	49.51	53.38	56.90	61.16	64.18
39	20.00	21.43	23.65	25.70	28.20	32.74	38.34	44.54	50.66	54.57	58.12	62.43	65.48
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77
41	21.42	22.91	25.21	27.33	29.91	34.58	40.34	46.69	52.95	56.94	60.56	64.95	68.05
42	22.14	23.65	26.00	28.14	30.77	35.51	41.34	47.77	54.09	58.12	61.78	66.21	69.34
43	22.86	24.40	26.79	28.96	31.63	36.44	42.34	48.84	55.23	59.30	62.99	67.46	70.62
44	23.58	25.15	27.57	29.79	32.49	37.36	43.34	49.91	56.37	60.48	64.20	68.71	71.89
45	24.31	25.90	28.37	30.61	33.35	38.29	44.34	50.98	57.51	61.66	65.41	69.96	73.17
46	25.04	26.66	29.16	31.44	34.22	39.22	45.34	52.06	58.64	62.83	66.62	71.20	74.44
47	25.77	27.42	29.96	32.27	35.08	40.15	46.34	53.13	59.77	64.00	67.82	72.44	75.70
48	26.51	28.18	30.75	33.10	35.95	41.08	47.34	54.20	60.91	65.17	69.02	73.68	76.97
49	27.25	28.94	31.55	33.93	36.82	42.01	48.33	55.27	62.04	66.34	70.22	74.92	78.23
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49

Appendix 3: Daily Return Distributions









Appendix 4: Results of Kupiec's TUFF-Test

	Confidence Level	Number of Observations	Observed Number of Exceptions	Time Until First Exception	Kupiec's TUFF-Test: Test statistic LR_{TUFF}	Kupiec's TUFF-Test: Critical Value $\chi^2(1)$	Kupiec's TUFF-Test: Test Outcome
Top Portfolio							
	99 %	250	10	70	0.11	3.84	Accept
	95 %	250	25	23	0.02	3.84	Accept
	90 %	250	36	23	1.01	3.84	Accept
Equity Portfolio							
	99 %	250	10	9	3.09	3.84	Accept
	95 %	250	33	1	5.99	3.84	Reject
	90 %	250	50	1	4.61	3.84	Reject
Bond Portfolio							
	99 %	250	7	33	0.89	3.84	Accept
	95 %	250	18	3	2.38	3.84	Accept
	90 %	250	30	3	1.21	3.84	Accept
Equity Option Portfolio							
	99 %	236	12	33	0.89	3.84	Accept
	95 %	236	20	2	3.32	3.84	Accept
	90 %	236	29	2	2.04	3.84	Accept

Appendix 5: Summary of the Backtesting Results

TOP PORTFOLIO

Confidence Level	Exceptions / Observations	Frequency Tests			Independence Tests		Joint Tests	
		Traffic Light	TUFF-test	POF-test	Christof-fersen	Mixed Kupiec	Christof-fersen	Mixed Kupiec
99 %	10 / 250	Red Zone	Accept	Reject	Accept	Reject	Reject	Reject
95 %	25 / 250	Yellow Zone	Accept	Reject	Accept	Reject	Reject	Reject
90 %	36 / 250	Yellow Zone	Accept	Reject	Accept	Reject	Reject	Reject

EQUITY PORTFOLIO

Confidence Level	Exceptions / Observations	Frequency Tests			Independence Tests		Joint Tests	
		Traffic Light	TUFF-test	POF-test	Christof-fersen	Mixed Kupiec	Christof-fersen	Mixed Kupiec
99 %	10 / 250	Red Zone	Accept	Reject	Accept	Reject	Reject	Reject
95 %	33 / 250	Red Zone	Reject	Reject	Accept	Reject	Reject	Reject
90 %	50 / 250	Red Zone	Reject	Reject	Accept	Reject	Reject	Reject

FIXED INCOME PORTFOLIO

Confidence Level	Exceptions / Observations	Frequency Tests			Independence Tests		Joint Tests	
		Traffic Light	TUFF-test	POF-test	Christof-fersen	Mixed Kupiec	Christof-fersen	Mixed Kupiec
99 %	7 / 250	Yellow Zone	Accept	Reject	Accept	Accept	Accept	Reject
95 %	18 / 250	Yellow Zone	Accept	Accept	Accept	Accept	Accept	Accept
90 %	30 / 250	Green Zone	Accept	Accept	Accept	Accept	Accept	Accept

EQUITY OPTION PORTFOLIO

Confidence Level	Exceptions / Observations	Frequency Tests			Independence Tests		Joint Tests	
		Traffic Light	TUFF-test	POF-test	Christof-fersen	Mixed Kupiec	Christof-fersen	Mixed Kupiec
99 %	12 / 236	Red Zone	Accept	Reject	Accept	Reject	Reject	Reject
95 %	20 / 236	Yellow Zone	Accept	Reject	Accept	Reject	Reject	Reject
90 %	29 / 236	Green Zone	Accept	Accept	Reject	Reject	Accept	Reject