MODEL RISK MANAGEMENT
FOR MARKET RISK

by

Marjorie Grootenboer (CID: 01442532)

Department of Mathematics
Imperial College London
London SW7 2AZ
United Kingdom

Thesis submitted as part of the requirements for the award of the MSc in Mathematics and Finance, Imperial College London, 2017-2018.

Declaration

The work contained in this thesis is my own work unless otherwise stated.

Signature and date:

Marjorie Grootenboer

September 11th, 2018

Acknowledgements

First of all, I would like to express my deepest gratitude to Dr. Harry Zheng, lecturer in Mathematical Finance at Imperial College London, who kindly supervised my Master thesis. I particularly appreciated his broad reachability. His guidance was a source of inspiration in my work. I was indeed very lucky to have him as my internal supervisor.

I would like to take this opportunity to thank every teacher I had this year at Imperial College London, each one of them being attentive, clear and keen. A special appreciation to Professor Mikko Pakkanen, lecturer in Mathematical Finance and Statistics at Imperial College London, who made me discover the risk management area. His *Quantitative Risk Management* course was one of the reasons I chose model risk as my subject.

A special gratitude goes to Vivian Vu, Quantitative Finance Analyst at Mazars UK, for her constant help and support. She shared with me insightful suggestions and she spent a great amount of time challenging my ideas. I am also thankful to Xavier Larrieu, Senior Manager and Head of Quantitative Services at Mazars UK, who trusted me and offered me an internship at Mazars UK this summer. His remarks and his expertise helped me tackle my subject as a whole.

Finally, I would like to thank my family for their unconditional support, and more specifically my parents. They taught me to aim high for every project I undertake, and to try hard to succeed. I wouldn't have accomplished my academic and professional achievements without them.

Contents

Introduction

Context

In the aftermath of the 2008 crisis and with the growing complexity of financial markets, institutions have been focusing more and more on risk forecasting. Often associated with a negative fallout, risk represents a major stake for financial bodies, as it describes the possibility of financial loss. Its measure is often used in decision making - to determine risk capitals or to manage the riskiness of a position for example [26, Slide 35].

In order to assess financial risks, banks have been developing models of all types. A model can be defined as follows [33]: a « model refers to a quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates ». Their accuracy is particularly crucial, as models can have knock-on effects on the bank's soundness and even on the economy. Let us remember that the subprime mortgage crisis of 2007 was partly due to inappropriate models [15]. Credit rating agencies provided irregular high scores to debt securities whose quality was undeniably lower. Indeed, a significant proportion of the loans belonged to homebuyers who had a poor credit situation or undeclared incomes. Hence, the model soundness is essential.

Since the global crisis, financial institutions have been relying even more on models, notably with the desire to automate processes and thus to gain time and precision. For instance, banks have been developing algorithmic trading: large trades are pre-programmed and are then executed thanks to automated platforms. Concerning credit risk, banks now question their model of probability of default (PD) in order to extend credit or not. For pricing and valuation, many tools are widespread and traders use them on a day-to-day basis.

Model Risk Definition

In response to this model-dependent environment, regulations have emerged to raise awareness about the so-called *model risk*. In April 2011, the Federal Reserve System delivered a Supervisory Guidance (SR 11-7) for the purpose of providing guidelines on how to manage efficiently this risk. In 2013, the Capital Requirements Directives (CRD) have been introduced to present the European Union directives on capital requirements. In the third article of CRD, model risk is defined as:

> « the potential loss an institution may incur, as a consequence of decisions that could be principally based on the output of internal models, due to errors in the development, implementation or use of such models ».

This definition implies two main sources of model risk [33]:

- The model itself is imperfect and generates incorrect outputs. The errors can be located anywhere in the development. It can be caused by a theoretical error, a misspecification, an operational mistake or even an approximation that was chosen to simplify the implementation.
- The model is used inappropriately or wrongly. The soundness and accuracy of the outputs do not guarantee a risk-free model. A model is often designed for a specific environment and using it outside of its limits may produce erroneous results. Moreover, users must be trained properly and regularly.

Like any other risk, model risk must be analysed and more specifically its sources and consequences should be clearly identified and communicated. In addition to working on *individual* model risk, one should also consider *aggregate* model risk [33]. As financial institutions work in multi-model environments, relationships and dependencies among models must be considered to truly encompass model risk. This additional dimension constitutes a further challenge.

Model Risk Quantification

Model risk can be seen as the risk of occurrence of a significant difference between a model value and an observed value [32]. Hence, it should be possible to identify, mitigate and, above all, quantify this gap. Currently, the emphasis is on the general management - how models are developed and used for instance - rather than on the quantification. In the Supervisory Guidance SR 11-7, we find three steps in the model risk framework that can impact model risk:

1. The model development, implementation and use,
2. The model validation,
3. The model governance, policies and controls.

We want to quantify model risk in order to forestall its potential consequences. Let us think about how to calculate model risk. What will be the unit of the output? Is it possible to have a single method which would be applicable to different models? Regarding the first question, one should consider model risk as any other risk: it should be simply added to other risk measures in order to evaluate a total risk. Risk forecasts being often computed as ratios with returns, one could suggest currency or log currency as the reference unit [16]. With regards to the second question, we can put forward two approaches to monitor model risk: computing as many models as possible and then evaluating the gap of the outputs, or assessing model risk against a benchmark model [3].

However, as models tend to be more and more sophisticated, it becomes particularly difficult to measure the inherent model risk. This complexity can especially be found in credit risk and operational risk [32]. Market risk measurement is nevertheless more developed, and evaluating discrepancies between models or specifying a benchmark model is reasonable. In this way, we shall concentrate on the measure of model risk within market risk models.

Model risk among market risk models implies that model outcomes are inconsistent [14]. For example, if we consider several models of market risk measure, it means that, for the same dataset and time period, we would not have the same results for the different models. Before trying to evaluate model risk within market risk methods, let us focus briefly on market risk and more specifically on the measures we will study. In 1996, the Basel Committee (BCBS) put forward the Value-at-risk (VaR) and asked banks to compute it using 99th percentile. VaR is thus a reference measure in market risk. In 2013, the BCBS published the Fundamental Review of the Trading Book (FRTB). One of its key reforms was to replace VaR with Expected Shortfall (ES) at 97.5%. We shall work with both measures and try to differentiate their results.

Research Structure

This thesis aims at illustrating model risk for market risk measures. More precisely, we quantify model risk for eight models of 99% VaR and 97.5% ES. The two risk measures are also compared.

In Section 1, we choose and compute eight models of markets risk measures. We start with two historical methods (filtered and unfiltered), then five analytical models (specific loss distributions, GARCH volatility estimation and Extreme Value Theory), and we finish with a Monte Carlo approach.

Section 2 is dedicated to the quantification of model risk. First, we backtest each model of the previous section in order to assess the quality of the methods via statistical tests. Secondly, we develop three approaches to quantify individual model risks: the risk ratio, the worst-case measure and the benchmark model comparison. For the benchmark, we compare each method to a model that is chosen according to the backtest results at first, and we then introduce a measure based on the Bayesian perspective. Those three approaches enable to assess the level of disagreement among the models [14] and give trends on the model outputs. In order to observe if those methods give similar information, we then focus on the correlation between model risk measures. In the next step, we analyse the models' sensitivity by changing parameter values. Finally, we focus on the model risk management, linking the quantification to the qualitative regulatory framework that financial institutions have to apply.

1 Market Risk Measure: VaR and ES Models

Market risk is associated with unexpected moves in market prices of financial assets such as stock prices and interest rates [11, Slide 1067]. Those fluctuations can lead to financial losses in a position. The evaluation of market risk consists in measuring the riskiness of this loss. Most risk measures rely upon statistical quantities which connect to the loss distribution [26, Slide 39]. It is worth noticing that this distribution is often determined in a backward-looking approach, which implies that the events that did not appear in the history might not be detected.

VaR is an industry-wise standard for market risk. It is a statistical measure of potential portfolio losses, based on the loss distribution. Already used by JP Morgan in the 1990s through its RiskMetrics system, VaR was introduced in the regulatory framework in 1996, in Basel I [6]. The idea behind VaR is to set a large probability α and to find a level l such that the probability that the loss surpasses l is equal to $1-\alpha$, a small probability [26, Notes on risk measures]. In this way, 99% VaR - the risk measure mentioned in regulations - is the loss that should only be exceeded 1% of the time. VaR can also be considered as a measure of capital adequacy [18, Chap 9], which is used to control the level of capital that banks have to hold to sustain losses.

Nevertheless, VaR presents disadvantages that were outlined during the 2008 crisis - its inability to detect the severity of losses above VaR notably. ES was presented in 1993 [30] as a 'coherent' [4] alternative to VaR and 97.5% ES was in this way put forward in FRTB [8] (2013). To estimate ES, we need to compute VaR first and then take the expectation on the values of the loss distribution that exceed VaR.

As mentioned in the introduction, we focus on eight risk forecast models. The first obvious choice is to implement historical simulation (HS) as it is one of the simplest and one of the most commonly used in the industry [23]. JPMorgan and Bank of America indeed estimate their trading risk through HS. Secondly, we introduce a filtered historical simulation combined with an exponentially weighted moving average (EWMA) volatility model. Next, we take two basic analytical methods based on a specific distribution: Normal and Student-t. We then consider GARCH family models. We include a Normal GARCH method (GARCH N), as it is a popular volatility estimator, and a Student-t GARCH (GARCH St) which is known for accurately capturing extreme events [14]. Thereafter, we examine extreme value theory (EVT) and we compute a model based on the generalized Pareto distribution. It enables us to focus on the tail of the loss distribution. Last but not least, we implement a Monte Carlo approach (MC).

1.0.1 Definitions

We consider the loss distribution at time $t+1$: $L_{t+1} := -(V_{t+1} - V_t)$, with V_t the portfolio value at time t. As V_{t+1} is not known at time t, L_{t+1} is a random variable. In this project, α denotes a real number in $(0,1)$.

Definition 1.1. The Value at Risk (VaR) of loss L at confidence level α is defined as:

$$VaR_\alpha(L) := \inf\{l \in \mathbb{R} : \mathbb{P}(L > l) \leq 1 - \alpha\}.$$

We can rewrite the definition: $VaR_\alpha(L) = \inf\{l \in \mathbb{R} : \mathbb{P}(L \leq l) \geq \alpha\} = q_\alpha(L) = F_L^{-1}(\alpha)$, with F_L the cumulative distribution function. In this way, $VaR_\alpha(L)$ is the α-quantile of the loss L.

Definition 1.2. The Expected Shortfall (ES) of loss L at confidence level α is defined as:

$$ES_\alpha(L) := \frac{1}{1-\alpha}\int_\alpha^1 VaR_u(L)du.$$

Remark 1.3. ES is well defined when: $\mathbb{E}(|L|) < \infty$.

Lemma 1.4. *Assume that the loss L has continuous F_L and that we have: $\mathbb{E}(max\{L,0\}) < \infty$. The Expected Shortfall (ES) of loss L at confidence level α can be expressed as follows [26, Notes on risk measures, Lemma 3.4]:*

$$ES_\alpha(L) = \mathbb{E}(L|L \geq VaR_\alpha(L)).$$

Remark 1.5. This lemma illustrates the name "expected shortfall": ES is the expected loss given that VaR is violated.

As the definition of VaR and ES directly involves the loss distribution, we need to estimate it in order to have access to both risk measures. Three groups of methods can be considered to forecast the loss distribution [26, Slide 30] [21]:

- Historical simulation,
- Analytical method,
- Monte Carlo method.

The eight models aforementioned are presented according to their method type. In order to compute the models, we use **R** and we start from a code implemented during the *Quantitative Risk Management* course at Imperial College London, as part of courseworks [20] [24].

1.0.2 Choice of Portfolio

We work with a portfolio invested in ten US stocks of large institutions:

Symbol	Company	Sector
AAPL	Apple Inc.	Information Technology
BAC	Bank of America Corp.	Financial Services
BOE	Boeing Co.	Aerospace Industry
BP	BP PLC	Energy
DIS	Walt Disney Co.	Entertainment
GE	General Electric Co.	Diversified Industrials
JNJ	Johnson & Johnson	Pharmaceuticals
MCD	McDonald's Corp.	Restaurants
NESN	Nestle SA	Consumer Goods
WMT	Wal-Mart Stores Inc	Retail

We set the investment period from 10/12/2005 to 08/06/2018 and we harmonize the portfolio by removing the dates that do not contain information on the ten stocks.

Two criteria were used to select the stocks. The first requirement was that the stocks must have been traded on the US markets during our investment period. The second one was that the stocks must come from different industries in order to have a diversified portfolio. We decide to rebalance the portfolio on a daily basis to have an equal weights portfolio.

As logarithmic prices are often used as risk factors [26, Slide 15], we consider logreturns to be the risk factor changes:

$$r_{t+1} := logS_{t+1} - logS_t.$$

We can then define the linearized loss corresponding to the stock portfolio with equal weights ($w_{t,i} = \frac{1}{10}$) [26, Slide 16]:

$$\bar{L}_{t+1} = -V_t \sum_{i=1}^{10} w_{t,i} r_{t+1,i}.$$

In this project, we work with the daily linearized loss relative to the portfolio value which is handy to manipulate [27, Slide 104]:

$$L_{t+1} = \frac{\bar{L}_{t+1}}{V_t}.$$

In order to forecast the loss distribution, we use a rolling window of N simulations. It means that for each estimation of a loss distribution, we base our calculations on the N previous losses. We set $N = 500$ which is approximatively equal to two years of data.

1.1 Historical Methods

Historical methods rely on the hypothesis that history will be repeated and thus that the future will be a combination of past events. We use the empirical distribution to estimate the loss distribution.

1.1.1 Historical Simulation (HS)

This approach relies on the assumption that returns are independent and identically distributed (iid), which constitutes its main limitation. Indeed, returns are known to be dependent as they demonstrate stylized facts such as volatility clustering [25]. Moreover, HS gives the same weight to all returns [14] and might not capture extreme events if they did not happen in the historical dataset [26, Slide 32]. Yet, it is worth noticing that this method has a very simple implementation and is widespread in the industry. Contrary to parametric models, HS does not specify a distribution for the returns. Let us note that the method implicitly assumes that the empirical distribution is accurate and comprehensive: it should be a sound picture of the future returns [23].

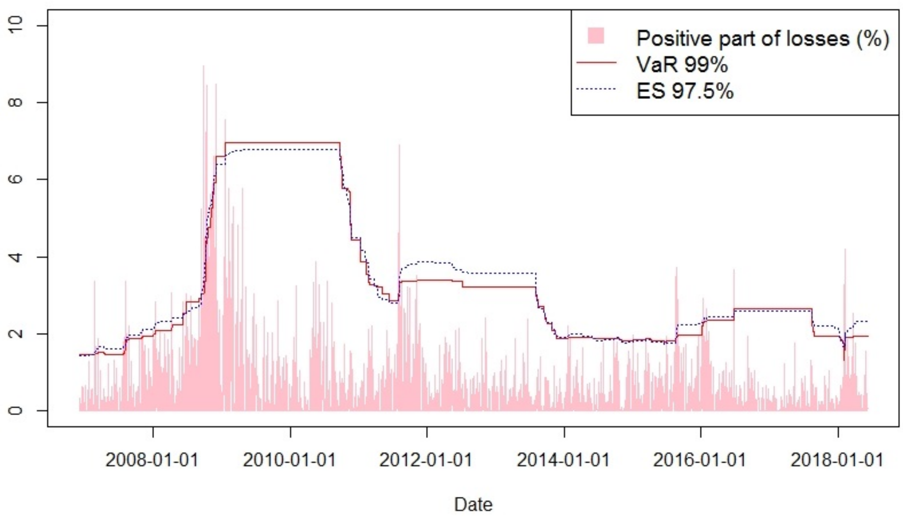
Consider the empirical α-quantile $\hat{q}_\alpha$ of a sample of N observed losses [26, Slide 45]. The number of VaR violations during the previous N trading days is given by: $M_{t,\alpha} := \#\{i \in \{0, 1, ..., N-1\} : L_{t-i} \geq VaR_\alpha(L_{t+1})\}$. VaR and ES are estimated with the following formulae:

$$VaR_\alpha(L_{t+1}) = \hat{q}_\alpha$$

$$ES_\alpha = \frac{1}{M_{t,\alpha}} \sum_{i=0}^{N-1} L_{t-i} \mathbb{1}_{\{L_{t-i} \geq VaR_\alpha(L_{t+1})\}}.$$

We obtain the following results for VaR and ES:

Figure 1: VaR 99% and ES 97.5% with Historical Simulation



We notice that the tracking of the positive part of the portfolio value is not accurate. Both VaR and ES are slow to react to the profit and loss (P&L) changes. This can be explained by the fact that HS gives equal weights to the whole forecasting window. As a consequence, when the losses decrease, the estimates for VaR and ES remain too conservative, as situations of large losses are still present in the window [20]. Furthermore, the method has difficulties to predict the first high losses of a peak such as during the 2008 crisis. Hence, this method does not seem to be optimal.

1.1.2 Filtered Historical Simulation (EWMA)

Definition 1.6. A white noise is defined as a covariance-stationary process $(\epsilon_t)_{t\in\mathbb{Z}}$ with the following autocorrelation function [27, Slide 10]:

$$\rho(h) = \begin{cases} 1, & h = 0 \\ 0, & h > 0. \end{cases}$$

Innovations refer to random variables with a white noise process.

The filtered historical simulation combines a non-parametric modelling of innovations ϵ_t with a parametric modelling of the volatility σ_t.

We choose an exponentially weighted moving average (EWMA) volatility filter for the parametric part of this market risk model. The estimate for the volatility, $\hat{\sigma}_t$, is computed with the EWMA scheme (Formula 1.1). We take: $\hat{\epsilon}_t = r_t - \hat{\mu}_t$, with $\hat{\mu}_t$ an estimate for the conditional mean of the risk factor changes.

Definition 1.7. The EWMA iterative scheme for the volatility is defined as [27, Slide 99]:

$$\hat{\sigma}^2_{t+1} = \alpha_1 \hat{\epsilon}_t + \beta_1 \hat{\sigma}^2_t. \tag{1.1}$$

As we use daily data in this project, we choose $\alpha_1 = 0.06$ and $\beta_1 = 0.94$ [27, Slide 99]. Furthermore, to simplify the calculation, we take $\hat{\mu}_t = 0$.

Remark 1.8. The term $\alpha_1 \hat{\epsilon}_t$ can be interpreted as the "intensity of reaction" to market evolution. The term $\beta_1 \hat{\sigma}^2_t$ represents the "persistence of volatility" [2]. Hence, by taking $\alpha_1 = 0.06$, the volatility forecasted the previous day has a 94% weight and the residual for that day has a 6% weight. It makes the result react quickly to the change in volatility.

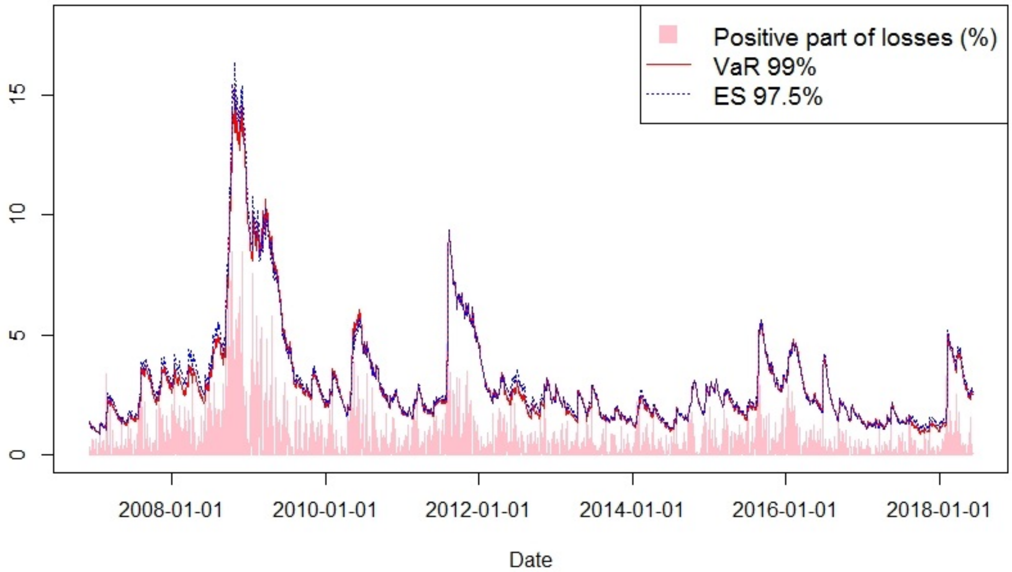
VaR forecasts are calculated using the following formula [27, Slide 102]:

$$VaR_\alpha(L_{t+1}) = \hat{\sigma}_{t+1} q_\alpha(\hat{F}_Z)$$
$$ES_\alpha(L_{t+1}) = \hat{\sigma}_{t+1} ES_\alpha(\hat{F}_Z).$$

As we are implementing a filtered historical simulation, $\hat{F}_Z$ is the empirical distribution of the standardized residuals.

We obtain the following results:

Figure 2: VaR 99% and ES 97.5% with Filtered Historical Simulation and EWMA



The VaR and ES forecasts give a precise tracking of the positive losses: both measures are reacting to changes in the position value in short periods of time. We also remark that the risk measures is sensitive to market conditions and thus is quite volatile. The EWMA recursion scheme only considers the previous days of volatility forecasts and thus has less "memory" than other models such as HS [20]. Hence, one should consider EWMA when the volatility is not constant [29]. Yet, from a computational point of view, filtered historical simulation can manage large portfolios and it remains a precise predictive method as it is supported by empirical information [23].

1.2 Analytical Methods

Analytical methods consist in selecting a model for the risk factor changes distribution so that the loss distribution can be analytically determined. The historical data is then fitted in order to evaluate the different parameters.

1.2.1 Normal and Student-t Distributions

Normal Distribution

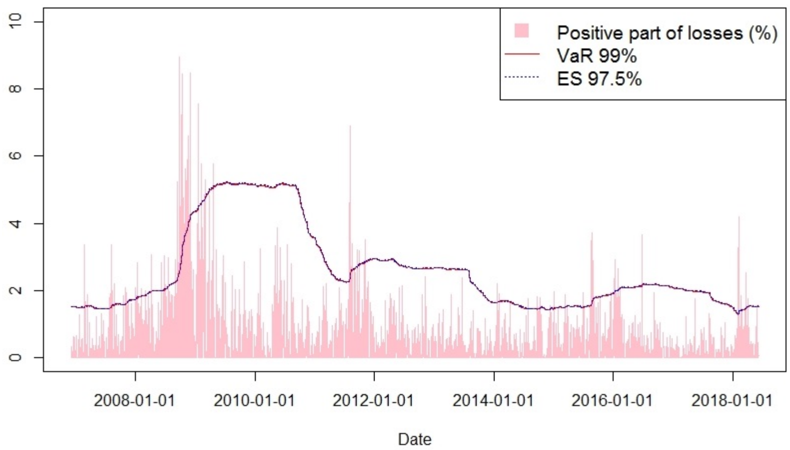
We assume that the risk factor distribution is Normal $N(\mu, \sigma^2)$. For each window of N observations, we estimate μ and σ by taking respectively the sample mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$. VaR is forecasted by [26, Slide 47]:

$$VaR_\alpha(L_{t+1}) = \hat{\mu} + \hat{\sigma}\Phi^{-1}(\alpha).$$

In the same way, we can estimate ES via its integral definition 1.2 which gives the following expression:

$$ES_\alpha(L_{t+1}) = \hat{\mu} + \hat{\sigma}\frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}.$$

Figure 3: VaR 99% and ES 97.5% with Normal Distribution



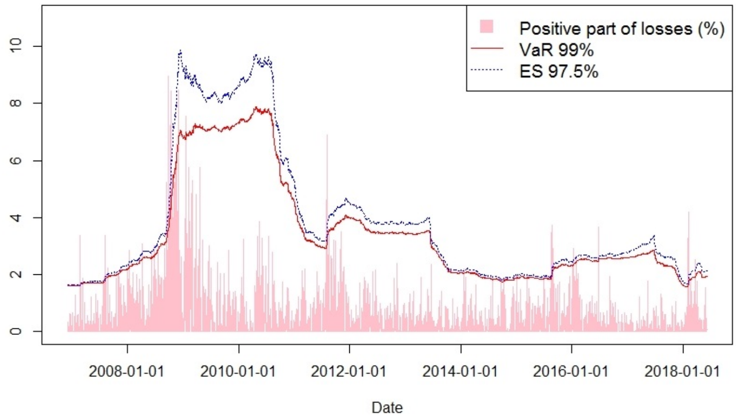
We observe that both measures are very similar yet not equal. One can easily notice that the Normally distributed method is not tracking in an accurate way the positive part of the daily losses. The highest observed loss is about 9% yet the highest estimate is below 6% in the entire forecasting period. The most important losses are thus not forecasted: all the peaks are underestimated, in particular during the 2008 crisis and in the end of 2011. The only reaction to those significant losses is an almost constant medium high risk forecast after the stress period.

Student-t Distribution

We fit the Student-t distribution $t(\nu, \mu, \sigma^2)$ to the N observations of the window. It gives us estimates for the three parameters. VaR and ES are given by [26, Slide 49]:

$$VaR_{\alpha}(L_{t+1}) = \hat{\mu} + \hat{\sigma} t_{\hat{\nu}}^{-1}(\alpha)$$
$$ES_{\alpha}(L_{t+1}) = \hat{\mu} + \hat{\sigma} \frac{g_{\hat{\nu}}(t_{\hat{\nu}}^{-1}(\alpha))}{1-\alpha} \left(\frac{\hat{\nu} + t_{\hat{\nu}}^{-1}(\alpha)^2}{\hat{\nu} - 1} \right).$$

Figure 4: VaR 99% and ES 97.5% with Student-t Distribution



The Student-t law has fatter tails than the Normal law. Hence, peaks in positive part of losses induce larger values in Student-t fit of VaR and ES. Let us also remark that this method is more sensitive and more conservative than the previous one. However, the highest losses are not well predicted: the portfolio may suffer important losses without the model preventing it. As a consequence, both Normal and Student-t models do not seem to be reliable risk measures.

1.2.2 GARCH Models

Generalized autoregressive conditional heteroskedastic (GARCH) processes were introduced to capture non-linear dependence of asset returns. Hence, this method models volatility clustering and persistence [27, Slide 44].

Definition 1.9. We define a $GARCH(p,q)$ process $(r_t)_{t\in\mathbb{Z}}$ by [27, Slide 45]:

- $(r_t)_{t\in\mathbb{Z}}$ is strictly stationary.
- $r_t = \sigma_t Z_t, \qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$
  with $(Z_t)_{t\in\mathbb{Z}}$ a strict white noise (iid variables) with mean 0 and standard deviation 1.

  We have $\alpha_0 > 0$, $\alpha_1, ..., \alpha_p \geq 0$ and $\beta_1, ..., \beta_q \geq 0$.
- $(\sigma_t)_{t\in\mathbb{Z}}$ has positive values and is strictly stationary.

Remark 1.10. Let us notice that this writing emphasizes the similarity between the EWMA forecast recursion (Formula 1.1) and the integrated $GARCH(1,1)$ expression, where $\alpha_1 + \beta_1 = 1$ and $\alpha_0 > 0$:

$$\hat{\sigma}_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 \hat{\sigma}_{t-1}^2.$$

We take $\alpha_0 = 0$ in the EWMA scheme.

In this project, we choose to work with $GARCH(1,1)$ as it gives a reliable fit of logreturn data [27, Slide 47]. We consider the following algorithm to estimate VaR and ES:

1. We choose a distribution for the innovations to fit the GARCH model.

2. We fit a static $GARCH(1,1)$ model on the first 500 trading days.

3. We refit the $GARCH(1,1)$ model every 50 days after the start of the forecasting period to get a volatility estimate for the whole forecast window. We use a 500 days rolling window for this step [20].

4. We estimate the α-quantile by taking the prior 500 days to calculate the empirical quantile.

5. We implement VaR and ES via the following formulae, with $\hat{\mu}_t$ an estimator of the conditional mean of the risk factor change r_t:

$$VaR_\alpha(L_{t+1}) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1} q_\alpha(\hat{F}_Z)$$

$$ES_\alpha(L_{t+1}) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1} ES_\alpha(\hat{F}_Z).$$

We display the results for VaR at 99% and ES at 97.5% with GARCH, using Normal (GARCH N), then Student-t innovations (GARCH St):

Figure 5: GARCH with Normal Innovations

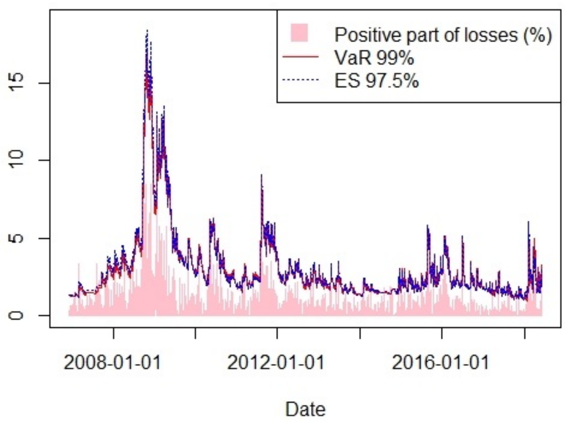
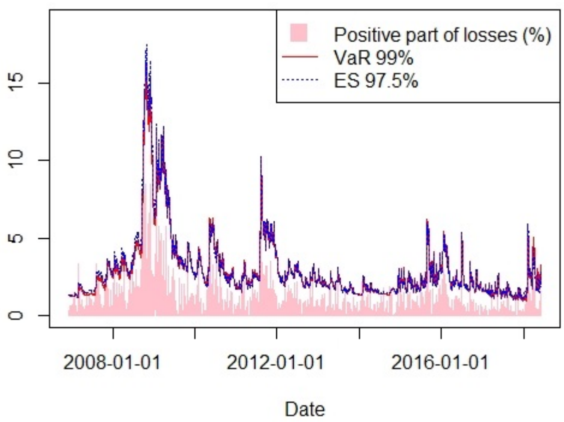


Figure 6: GARCH with Student-t Innovations



We observe that both GARCH models are responding quickly to the loss changes: there is an efficient tracking of the volatility dynamics. Regarding the output values, the results are quite similar for Normal and Student innovations. Let us remark that overall the Student-t model produces higher values which is logically explained by the fact that the distribution has fatter tails than the Normal distribution. However, between 2009 and 2010, we tend to observe higher forecasts for the Normal innovations model. Finally, the higher forecast is about 18%, which is twice the maximum value of the observed P&L: this method is conservative and thus introduce few violations.

1.2.3 Extreme Value Theory (EVT)

The objective of extreme value theory is to model the tails of loss distribution. In this way, this method is particularly relevant in risk management as tails are linked to the probability of extreme outcomes.

EVT relies on the assumption that the data is independent and identically distributed (iid). This hypothesis cannot be effective with most financial time series such as returns as we observe stylized facts. In order to overcome this problem, the dataset can be modified such that the iid assumption is almost met and then the EVT method is applied: this approach is called conditional EVT [28, Slide 49].

We consider the generalized Pareto distribution (GPD) method to model the behaviour of the right tail of the loss distribution (extreme losses) which relies on threshold exceedances [28, Slide 32].

Definition 1.11. The generalized Pareto distribution with shape $\xi \in \mathbb{R}$ and scale $\beta > 0$ has the following distribution function:

$$G_{\xi,\beta}(x) := \begin{cases} 1 - (1 + \frac{\xi x}{\beta})^{-1/\xi}, & \xi \neq 0 \\ 1 - exp(-\frac{x}{\beta}), & \xi = 0. \end{cases}$$

If $\xi \geq 0$, we must have $x \geq 0$ and if $\xi \leq 0$, we have $0 \leq x \leq -\frac{\beta}{\xi}$.

Let $X_1, ..., X_n$ be iid observations with distribution function F. We would like to have: $F_u(x) = G_{\xi,\beta}(x)$ with $F_u(x) = \frac{F(x+u)-F(u)}{1-F(u)}$ the excess distribution and $x \in [0, sup\{x \in \mathbb{R} : F(x) < 1\}]$.

We consider the losses relative to our portfolio: $L_{t+1} = \mu_{t+1} + \sigma_{t+1} Z_{t+1}$. We apply the following method to estimate VaR and ES [28, Slide 64]:

1. We model our losses with a $GARCH(1,1)$ model, with constant mean.

2. We use a rolling window to refit $GARCH(1,1)$ to the 500 previous losses via quasi-maximum likelihood estimation, which means that we assume standard normal errors and as a consequence we can apply EVT.

3. We select the threshold u: we arbitrary take the 93% quantile of the residuals [28, Slide 64] in the forecasting window.

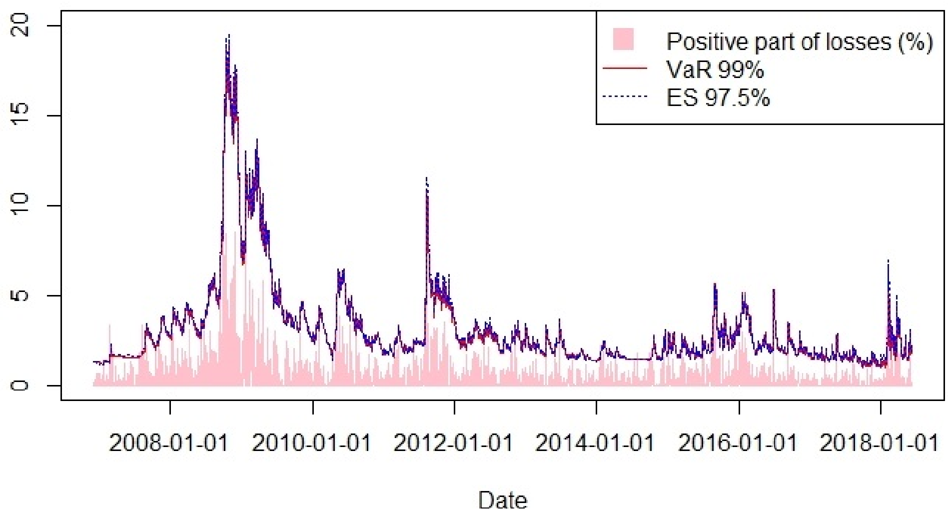
4. We fit a GPD to the standardized residuals and then we forecast VaR and ES with the following formulae:

$$VaR_\alpha(L_{t+1}) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}[u + \frac{\beta}{\xi}((\frac{1-\alpha}{1-F(u)})^{-\xi} - 1)] = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\hat{q}_\alpha(Z_{t+1}),$$
$$ES_\alpha(L_{t+1}) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\left(\frac{\hat{q}_\alpha(Z_{t+1})}{1-\xi} + \frac{\beta-\xi u}{1-\xi}\right),$$

with ξ and β estimated parameters values of the fitted GPD.

We obtain the following results for the EVT model:

Figure 7: VaR 99% and ES 97.5% with Extreme Value Theory



One can notice that the VaR and ES forecasts reflect reality. Indeed, the EVT outputs are volatile enough to track well the observed losses. Even the highest loss peaks are well predicted. This can be explained by the fact that EVT especially emphasizes on the tail and thus captures well extreme events. We observe that the risk measures are quite conservative. The maximum value is just below 20%, which is the highest estimate from the different methods so far. Hence, this method seems to be sound and reliable.

1.3 Monte Carlo Method

In the same way as for the analytical methods, the Monte Carlo method (MC) relies on an explicit distribution for the risk factor changes [26, Slide 33]. The general premise of Monte Carlo is to generate random samples that are used to compute an empirical mean. In the case of VaR, risk factor changes are estimated with a random element and the loss function is defined as the opposite of the sample mean of the logreturns.

The Monte Carlo approach has the property to handle different return distributions such as Normal or Student-t distributions. However, to capture rare events, this method requires a large number of replications and can thus be very time-consuming. For now, we only work with 10 000 simulations as the times of computation are important. We increase this number in the subsection 2.6 dedicated to the sensibility of the models.

We focus on a "basic Monte Carlo method" [19]. We make the assumption that the change in risk factor ΔS over Δt, conditional on historical data, has a multivariate Normal distribution $N(0, \Sigma_S)$.

We describe the method used to compute the loss distribution and the daily VaR. It then gives access to the daily ES:

1. In a first step, we need to build the portfolio covariance matrix Σ_S, which is a symmetric matrix of dimension m x m, with m the number of stocks in the portfolio.

   We choose the exponentially weighted moving average method to evaluate the matrix. It puts more weight on recent data, with the "smoothing constant" λ, and as a consequence it removes the "ghost feature" that can be observed in a classic historic covariance matrix method [1].

   Remark 1.12. This ghost feature can be observed when changing the length of the forecasting window of a historical method. As mentioned in Section 1.1.1, equal weights are given to each past data. It leads to a high range of values which significantly fluctuate in a short horizon, and average results which are less volatile. We observe this feature in Section 2.6, when analysing the sensitivity of HS to the number of days in the rolling window.

   Definition 1.13. Let us assume we have $t-1$ observations of a time series $(x_t)_{t\in\mathbb{Z}}$. The exponential weighted moving average of $(x_t)_{t\in\mathbb{Z}}$ is defined as [1]:

$$EWMA(x_{t-1}, ..., x_1) = \frac{x_{t-1} + \lambda x_{t-2} + \lambda^2 x_{t-3} + ... + \lambda^{t-2} x_1}{1 + \lambda + \lambda^2 + ... + \lambda^{t-2}}. \tag{1.2}$$

To estimate the volatility thanks to the EWMA method, we replace the time series $(x_t)_{t\in\mathbb{Z}}$ by the squared returns $(r_t^2)_{t\in\mathbb{Z}}$ in the equation 1.2 [1]. Furthermore, as $1+\lambda+\lambda^2+... = (1-\lambda)^{-1}$, the expression for the volatility forecast is: $\hat{\sigma}_t^2 = (1-\lambda)\sum_{i=1}^{\infty}\lambda^{i-1}r_{t-i}^2$.

In this way, the EWMA scheme can be changed into the recursive expression:

$$\hat{\sigma}_t^2 = (1-\lambda)r_{t-1}^2 + \lambda\hat{\sigma}_{t-1}^2.$$

With $\lambda = 1-\alpha_1$, we find the same volatility scheme as in the filtered historical simulation part (1.1.2). We forecast the covariance by taking the cross product of two returns $r_A r_B$.

Let us consider the vector R_t of asset returns on day t. To evaluate the covariance matrix, we compute the following matricial expression:

$$\Sigma_{S,t+\Delta t} = (1-\lambda)R_t R_t^T + \lambda\Sigma_{S,t}.$$

In the literature, several values for the smoothing constant can be found. A method to estimate λ could be to minimize the mean squared error between the observed squared return and the variance forecast [2]. However, most of the time, λ is chosen subjectively according to the length of the period of forecast. In this project, we want to compute daily VaR and thus we need daily covariance matrices. Hence, according to C. Alexander [2], we choose $\lambda = 0,94$ and we set the historical data length to 500 days.

2. In this second step, we perform the Cholesky decomposition of Σ_S. This method is quite efficient for modelling dependence structures between several assets.

   Definition 1.14. Let Σ_S be a positive definite matrix. The Cholesky decomposition of Σ_S is defined as:

   $$\Sigma_S = CC^T,$$

   where C is a lower triangular matrix with real positive diagonal elements.

   Remark 1.15. Let us notice that if Σ_S is positive definite then this factorization is unique.

3. In the third step, we generate a vector of independent standard normal variates Z. Its size is equal to our portfolio's. It allows us to compute $\Delta S = CZ$ [19] which is a vector of correlated variates. The change in risk factor over a short period Δt has a multivariate Normal distribution $N(0,\Sigma_S)$. Hence, in this step, we transform a vector of standard Normal variables into a vector of returns that are correlated to the 500 market observations used to compute the covariance matrix.

Remark 1.16. At this stage, we introduce the randomness that distinguishes Monte Carlo from the other methods.

4. We need to re-evaluate our position at $t + \Delta t$ for each stock. We introduce the Black-Scholes model to calculate the terminal stock price: we assume it follows a Geometric Brownian Motion. ΔS^i has a Normal distribution with mean 0 and standard deviation σ_i, with σ_i being the EWMA volatility that we computed during the step 1, for the covariance matrix. μ_i is the expected logreturn. We have the following formula:

$$S^i_{t+\Delta t} = S^i_t exp\left((\mu_i - \frac{\sigma_i^2}{2})\Delta t + \sqrt{\Delta t}\Delta S^i\right).$$

Remark 1.17. In this thesis, we consider a stock portfolio and we assume that the stock prices have a lognormal distribution. Hence, we have a closed form solution and the portfolio evaluation is simple.

To simulate the value of a more complex portfolio, another Monte Carlo procedure should be used within the VaR Monte Carlo approach. In our example, we could have used the following discretization scheme for our stock prices:

$$S^{i,k} = S^{i,k-1}(1 + \mu_i\delta_t + \sqrt{\delta_t}\Delta S^{i,k-1}), \qquad k \in [\![1, N]\!],$$

with $\delta_t = \frac{\Delta t}{N}$ with N being the number of simulations to estimate the terminal stock price.

In this project, we only need to generate one Brownian Motion to value the portfolio in one loss function simulation. With a more complex portfolio, we would have to generate a new standard Normal variate N times for one loss function simulation. This explains why the valuation step is considered to be the most time-consuming [19][18].

5. After computing the individual logreturns of the terminal stock prices, we evaluate the loss function. As we want a portfolio with equal weights, the loss function is defined as:

$$L_{t+\Delta t} = -\frac{1}{m}\sum_{i=1}^{m} r^i_{t+\Delta t}.$$

6. We repeat the steps 3, 4 and 5 n times - with $n = 10\ 000$. It allows us to estimate the loss distribution:

$$\mathbb{P}(L > x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{L_i > x\}}.$$

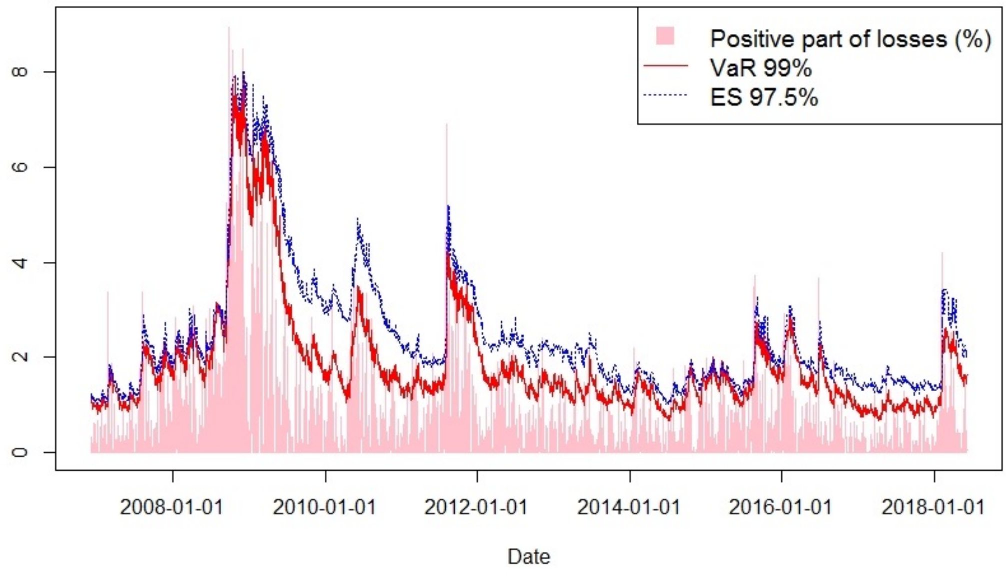
7. The final step consists in calculating the risk measures. To obtain the daily VaR, we sort the n values of the loss function by decreasing order. The VaR is therefore defined as: $\mathbb{P}(L > VaR_\alpha(L)) = 1 - \alpha$. We simply take the $n(1-\alpha)^{th}$ loss function value from the beginning.

   Lastly, to estimate the ES, we refer to the Lemma 1.4 and we calculate the mean of the losses such that the losses are greater than the VaR.

Remark 1.18. With this method, let us observe that the VaR is less stable than the ES as we take the $n(1-\alpha)^{th}$ loss function value with n being arbitrary, contrary to the ES that is an average.

As we want to compare this Monte Carlo model to the other market risk methods, the complete algorithm needs to be repeated for each day of our dataset. We display the results below:

Figure 8: VaR 99% and ES 97.5% with Monte Carlo



The risk measures are reacting quickly to the changes in the portfolio position. We observe that the VaR tracking of the positive part of the portfolio losses is quite precise. Nevertheless, VaR measures tend to underestimate regularly the losses. The ES has more conservative values yet, when the position observes large losses, the forecasts downplay the situation. As the largest losses are not well predicted by this MC method, we can suppose that the backtest results will not be good: in a stress period, the model does not give high enough forecasts and thus the portfolio suffers important losses.

We come back on Remark 1.17 about the time of computation. We focus on the first 6 steps, before computing the risk measures. We evaluate the average times and the total times for three steps of the algorithm: the calculation of the covariance matrix, the revaluation of the portfolio and the determination of the loss function. The times are displayed below:

Algorithm Step	Covariance Matrix	Portfolio Valuation	Loss Function
Average Time (seconds)	7.56e-02	8.20e-04	2.17e-04
Total Time (seconds)	219.0	2 374.5	629.2

Table 1: Times of Computation

The valuation of the portfolio is repeated n times and is therefore the most-time consuming step, with 57.6% of the total time of the first 6 steps. Yet, one can notice that the covariance matrix is the longest step of the algorithm in average, but it is only computed once.

We managed to compute a Monte Carlo approach for both VaR and ES. The results are mixed as it seems that a lot of violations occur. To strengthen this approach, an other method could be applied: the delta-gamma approximation [19]. This model relies on the assumption that the risk factors and the portfolio value have a quadratic relation.

Let us now compare qualitatively the eight market risk methods that we implemented in the beginning of this Section 1.

1.4 Qualitative Analysis

The purpose of Section 1 was to produce a one-day-ahead forecast for the linearized loss. We chose two risks measures to do so: VaR at 99% confidence and ES at 97.5%. We implemented eight different methods and we illustrated the three types of loss distribution modelling: analytical, historical and Monte Carlo methods.

By construction, the models do not rely on the same assumptions. Analytical and Monte Carlo approaches need to specify a distribution for the risk factors changes. It is an asset as those models can work for different distributions. However, it may present a risk as financial returns do not follow a specific distribution. HS is also based on a strong assumption for the distribution - returns are iid - which is not verified in real-life. Monte Carlo presents three characteristics it is worth emphasizing: this method introduces random scenarios - which differs, for instance, from historical methods relying only on past data, - the portfolio is revaluated using a specified model - it is an additional hypothesis, thus an additional source of risk, - and the time of computation that is required is more important than the other methods. Finally, the eight models share a common feature: we need a significant amount of data to implement them.

The graphs allow a qualitative analysis of the methods. First, we can notice that HS, Normal and Student-t do not track very quickly the changes in the portfolio value. It takes time before those models take into account what is happening in the position and thus change their outputs. It leads to overestimation of the observed losses most of the time and violation of important peaks. An other trend concerns Normal and MC: the two methods tend to have lower results and are therefore too optimistic. On the contrary, EVT is the most conservative with, in general, higher values of risk measures.

The former regulations put to the fore VaR at 99%. FRTB [8] now requires banks to move on to ES at 97.5% confidence. This project enables us to compare those two metrics. VaR and ES are quite similar for the different models. Although it has a lower confidence level, ES gives slightly higher estimates than VaR.

In order to have more precise and quantitative comparisons, we move on to Section 2 which is dedicated to the quantification and management of model risk.

2 Model Risk Quantification and Management

As mentioned in the Supervisory Guidance SR 11-7, there are several sources of model risk. In this section, we first focus on the quality of the models' outputs via backtests. We then compute model risk measures to assess quantitatively the coherence of the eight models. At this point, the correlation between the model risk results is observed in order to evaluate whether these methods give similar information. Next, we perform a sensitivity analysis on specific parameters. At last, we describe a management framework that should be applied by financial institutions to reduce their model risk.

We introduce model risk with different points of view: by examining individually each model, by analysing the outputs' divergence of the models, by assessing model risk thanks to a benchmark model, and by looking at the "parameter uncertainty" [3]. These analysis also enable us to compare VaR and ES as regulatory market risk measures.

2.1 Backtesting Market Risk Models

Backtesting consists in comparing the observed losses with the risk estimates of a model in order to assess its quality [8]. It is a common way to evaluate the accuracy of a risk model. It enables us to address individual model risk.

Backtesting is usually conducted through the analysis of violations [14]. We define the violation indicator as [26, Slide 58] :

$$I_t := \begin{cases} 1, & L_t > VaR_\alpha(L_t) \\ 0, & L_t \leq VaR_\alpha(L_t). \end{cases}$$

In 1995, the Basel Committee [5] introduced a factor $(3+k)$, which would be multiplied to the risk forecast, in order to include model risk in capital requirements. This multiplication factor is directly based on the backtest results of the risk measure RM - **in this project RM refers to VaR or ES**. The purpose of it is to impose a minimum of 3 when models are accurate and to add a multiplier k - a violation penalty - when too many exceptions are observed.

Basel II presents three zones of confidence. The value of the multiplication factor and the penalty zone are linked to the number of violations [12].

We display in the following table the penalty zones and the values of the multiplication factors, according to the number of violations in 250 business days:

Zone	**Violations**	**Multiplication Factor**
Green Zone	0 to 4	3.00
Yellow Zone	5	3.40
	6	3.50
	7	3.65
	8	3.75
	9	3.85
Red Zone	10+	4.00

Table 2: Basel Accord Penalty Zones

In this way, the daily capital requirements CR are defined as [5]:

$$CR_t := max\{RM_{t-1}, (3+k)\overline{RM}_{60}\},$$

with RM_{t-1} the previous day's risk measure and $\overline{RM}_{60}$ an average of the risk measures on each of the last sixty trading days.

Hence, if a model does not satisfy the minimum backtest results, the multiplication factor is increased. We understand the importance of having models that fall into the green zone: in the yellow or red zone, the penalty makes the capital requirements higher and the model is asked to be reviewed.

The risk measure that is mentioned in regulations is VaR [7]. While backtesting our results, we forecast the penalty zones for both VaR and ES to complete our analysis. We implement three backests: the unconditional coverage test (for VaR), the joint test (for VaR) and a Normal ES backtest.

We analyse the results of each backtest to forecast the penalty zones. We also focus on the p-values, which measure how close the result is to a defined null hypothesis - a high value corresponds to a high consistency [26, Slide 57]. If the backtest is accepted, the p-value is highlighted in green; alternatively, we present in red the backtests that are rejected.

For the VaR tests, we use a code provided by M. Pakkanen during the *Quantitative Risk Management* lecture in the context of a coursework [20].

2.1.1 Unconditional Coverage Test

Kupiec's introduced the unconditional coverage test. The idea is to check whether the number of violations is coherent with the confidence level [26, Slide 59].

We state the null hypothesis H_0 of the test: *The VaR forecasts are computed using a model that is correctly specified.* We assume that the indicators $I_1, ..., I_T$ are Bernoulli variables with probability π. Under H_0, we have: "$\pi = 1 - \alpha$". We define the likelihood function [26, Slide 59]:

$$L(\pi, I_1, ..., I_T) := \prod_{t=1}^{T} (1-\pi)^{1-I_t} \pi^{I_t}.$$

The maximum-likelihood estimator is given by: $\hat{\pi} = \frac{1}{T} \sum_{t=1}^{T} I_t$. We can then define the test statistic LR_{uc}, which follows asymptotically a $\chi^2(1)$ distribution [26, Slide 60]:

$$LR_{uc} := -2log \frac{L(1-\alpha, I_1, ..., I_T)}{L(\hat{\pi}, I_1, ..., I_T)}.$$

Hence, the p-value is given by the formula: $p_{UC} = 1 - F_{\chi^2(1)}(LR_{uc})$.

The results of this test are displayed in Table 3:

Model	Violations (Expected)	LR_{uc}	p_{UC}-*value*	Penalty Zone
Normal distribution (Normal)	86 (28)	74.232	0	Yellow
Student-t distribution (Student-t)	48 (28)	10.541	1.168e-03	Yellow
GARCH - Normal (GARCH N)	39 (28)	3.165	7.525e-02	Green
GARCH - Student-t (GARCH St)	42 (28)	5.198	2.262e-02	Green
Extreme Value Theory (EVT)	35 (28)	1.189	2.756e-01	Green
Historical simulation (HS)	53 (28)	16.169	5.793e-05	Yellow
Filtered historical simulation (EWMA)	39 (28)	3.165	7.525e-02	Green
Monte Carlo (MC)	128 (28)	185.756	0	Red

Table 3: Unconditional Coverage Test for VaR 99%

We first have to react on the MC results. As we saw in the part 1.3, MC has a good track of the losses, yet, being too optimistic, it introduces more exceptions than the other models. Hence, the backtest rejects the MC VaR and MC is associated with the red penalty zone. The capital requirements are in this way more important and the model should be strengthen. On the contrary, three models stand out from the other: EVT, GARCH N and EWMA. They have a low number of violations - about 1.28% for the whole forecasting window - and a p-value that is close to 1.

2.1.2 Joint Test

One drawback of the unconditional coverage test is that it does not detect if the violations are evenly distributed or if they happen in clusters which could imply dire consequences for a financial institution.

In this way, we introduce a second backtest, the independence test, which was first presented by Christoffersen [26, Slide 64]. This test relies on the hypothesis that VaR violation indicators form a Markov chain.

We define the corresponding parameters and the transition matrix:

$$\pi_{11} := \mathbb{P}(I_t = 1|I_{t-1} = 1), \qquad \pi_{01} := \mathbb{P}(I_t = 1|I_{t-1} = 0),$$
$$\Pi := \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix}.$$

The null hypothesis of this test is that the indicator functions are mutually independent which can be written as: H_0 : "$\pi_{01} = \pi_{11}$". The maximum likelihood estimator $\hat{\pi}$ used for the unconditional coverage test stays unchanged. The test statistic, which under H_0 is also following asymptotically a $\chi^2(1)$ distribution, is therefore given by [26, Slide 66]:

$$LR_{ind} := -2log\frac{L(\hat{\Pi}_0, I_2, ..., I_T)}{L(\hat{\Pi}, I_2, ..., I_T)}.$$

In order to have a comprehensive backtest, which covers both the independence and the correct coverage, we introduce the joint test [26, Slide 67]. The null hypothesis of the joint test is that the violation indicators are both independent and following a Bernoulli distribution.

$$H_0 : "\pi_{01} = \pi = \pi_{11}, \ \ \pi = 1-\alpha".$$

We define the joint test statistic as: $LR_{joint} := LR_{uc} + L_{ind}$.

This statistic follows asymptotically the $\chi^2(2)$ distribution and its p-value can be computed with the following formula:

$$p_J = 1 - F_{\chi^2(2)}(LR_{joint}).$$

The results of the joint test are displayed in Table 4:

Model	LR_{joint}	p_J-$value$
Normal	82.347	1.314e-18
Student-t	14.307	7.823e-04
GARCH N	3.515	1.725e-01
GARCH St	5.415	6.671e-02
EVT	1.775	4.117e-01
HS	25.155	3.448e-06
EWMA	5.679	5.846e-02
MC	185.779	4.557e-41

Table 4: Joint Test for VaR 99%

As predicted with the unconditional coverage test, MC VaR fails again the backtest. It is worth noticing that the Normal, Student-t and HS are also rejected. It was foreseeable as we observed in Section 1 that their tracking was poor. With the joint test, more comprehensible, we note that GARCH St does pass this time. EVT and GARCH N are also emphasized, as they have a higher p-value than with the unconditional coverage test. EWMA results remain robust.

2.1.3 Backtesting ES

In order to backtest ES, we introduce a Normal test. Let us first come back to the definition of the risk measure. As $\mathbb{E}\left[\mathbb{1}_{\{L_t > VaR_\alpha(L_t)\}}\right] = 1 - \alpha$, we have:

$$\mathbb{E}\left[(L_t - ES_\alpha(L_t))\mathbb{1}_{\{L_t > VaR_\alpha(L_t)\}}\right] = 0.$$

Backtesting ES is equivalent to observing when the loss is greater than the expected shortfall forecasts on days when there is a VaR violation [26, Slide 69].

We define a new variable: $k_t := (L_t - ES_\alpha(L_t))\mathbb{1}_{\{L_t > VaR_\alpha(L_t)\}}$. H_0 is accepted if both ES and VaR are equal to their forecasts.

We consider the following test statistic which follows asymptotically the standard Normal distribution under H_0 [26, Slide 70]:

$$TS := \frac{\sum_{t=1}^{T} k_t}{\sqrt{\sum_{t=1}^{T} k_t^2}}.$$

We can compute the p-value with the following formula: $p = 1 - \Phi(TS)$.

The results of the Normal test are displayed in Table 5:

Model	Violations	TS	*p-value*	Penalty Zone
Normal	85	5.536	1.544e-08	Yellow
Student-t	35	-2.215	9.866e-01	Green
GARCH N	39	0.084	4.663e-01	Green
GARCH St	34	0.166	4.341e-01	Green
EVT	35	0.479	3.160e-01	Green
HS	46	1.891	9.231e-02	Green
EWMA	33	0.219	4.133e-01	Green
MC	66	-2.016	9.781e-01	Yellow

Table 5: Backtest for ES 97.5%

It is worth noticing that the Monte Carlo ES gives better results than VaR. Next, as we noticed in Section 1, most 97.5% ES are more conservative than 99% VaR. As a consequence, we observe less violations and that most models are placed in the green zone, which is what we are looking for. The best p-value is awarded to Student-t, yet, one should not forget that most exceptions occurred for significant losses. Hence, even if the backtest results are quite good, it is important to conduct other tests and observations, and not to conclude only with those statistical tests.

One should keep a critical opinion on the backtest results. The tests rely on statistical distributions and we saw previously that financial data does not follow a particular distribution. Moreover, backtests assessed the number of violations rather than their location within the time period. It would be preferable to have uniform exceptions: if a model presents a small number of violations - and passes the backtest requirements - the violations could occur at the same moment. This implies that the financial institution suffers important losses [23]. Hence, backtesting may not be the most efficient way to compare our models.

To conclude, in the model risk framework, backtesting is an important step to assess both market risk and individual model risk. Indeed, the p-values can be understood as model risk measures. Unfortunately, it is not comprehensive and qualitative analysis should be conducted to complete the quantitative results.

2.2 Risk Ratio

The paper *Model Risk of Risk Models* (2016) [14] introduces a new method called risk ratio to illustrate the "level of disagreement amongst different chosen models". We rely on it in this section.

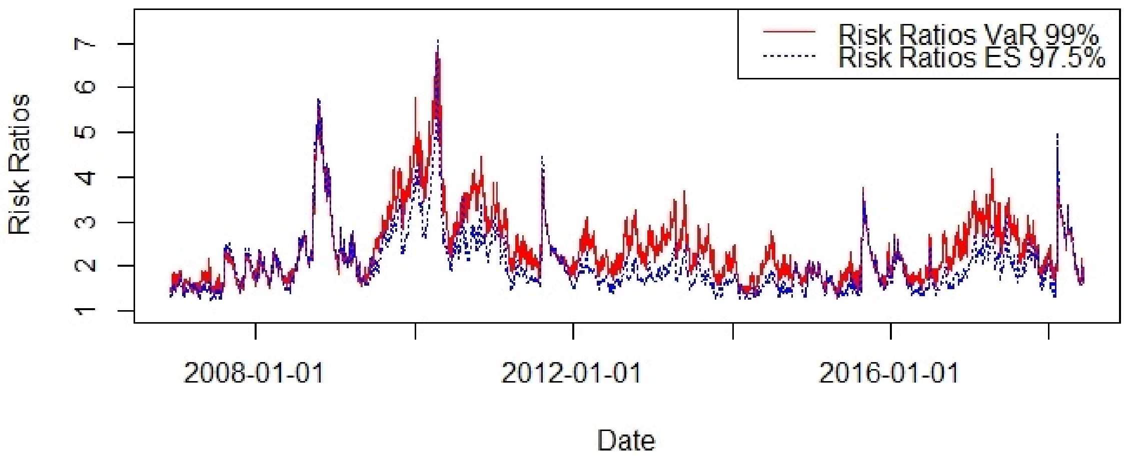
Definition 2.1. Let us consider K models of risk forecast. Each model gives a risk measure RM_{t+1}^k on day $t+1$ by using data available at time t. We define the risk ratio RR as the ratio between the highest and the lowest measure [14]:

$$RR_{t+1} := \frac{max\{RM_{t+1}^k\}_{k=1}^K}{min\{RM_{t+1}^k\}_{k=1}^K}.$$

Remark 2.2. The risk ratio provides an easy way to quantify model risk. If the chosen models give similar forecasts, then the risk ratio is close to 1. On the contrary, if the forecasts are significantly different, then the risk ratio can capture this divergence and quantify it.

We start by calculating the risk ratio for the eight chosen models on a daily basis. The results are displayed below:

Figure 9: Risk Ratios for VaR 99% and ES 97.5%



We observe two major peaks: 5.682 (for VaR) in the end of 2008 (2008-10-16) and 7.062 (for ES) in 2010 (2010-04-27). The first is undoubtedly due to the 2008 financial crisis. The second peak - the highest - may be linked to the models that do not track well the losses, such as the Normal and HS models. Indeed, we look at the portfolio stock prices that are displayed in the Appendix (Figure 24) to check if a market event would have explain this drop. If we do not observe a particular decline in prices at this period, the P&L displays a large peak in the end of 2011. Hence, the explanation must be linked to both the P&L peak in 2011, and the gap between the models that are slow to react after a large loss and the models that have a good tracking after the crisis.

Other peaks can be noticed at the end of the forecasting period. The largest one, spread over 2016 and 2017, could be linked to the Brexit and Donald Trump's presidential election. Indeed, these events were not foreseen by the main polls and led to instability on financial markets.

To complete the analysis of the risk ratios, we examine the maximum and minimum VaR as well as ES forecasts [14]. The purpose is to find trends in the models to answer the following question: are some models' outputs always superior to the other values? We take about six points per year to make the graph more readable:

Figure 10: Highest and lowest daily 99% VaR forecasts

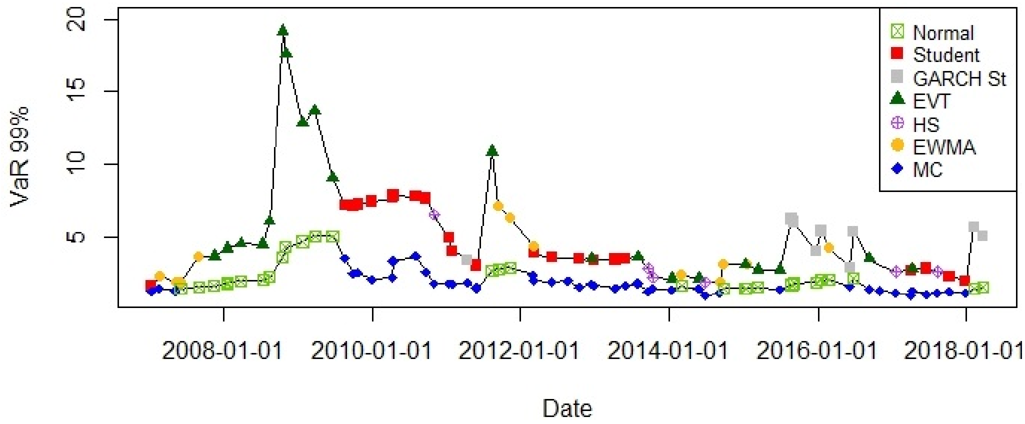
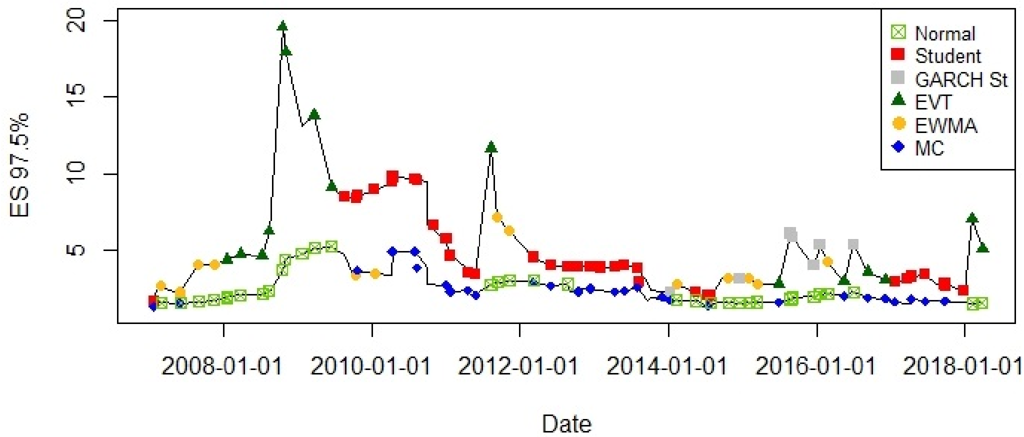


Figure 11: Highest and lowest daily 97.5% ES forecasts



We first observe that GARCH N does not appear on both figures, as its values are not the most extreme ones. In the same way, HS is not present for ES. Then, we observe for both VaR and ES that the Normal and MC models constantly generate the lowest estimates. We note that the highest observations tend to be produced by EVT and Student-t, fat-tailed models, which was foreseeable.

Next, we analyse the sensitivity of the risk ratios when we exclude one model [14]. In particular, we look at the maximal value, the mean of the ratios and the standard deviation (sd) on the whole period of analysis. The results are displayed below in Table 7. We underline the highest values.

Excluded Model	None	HS	EWMA	Normal	Student-t	GARCH N	GARCH St	EVT	MC
Max	6.969	6.969	6.969	6.969	6.211	6.969	6.969	6.969	5.805
Mean	2.047	2.026	1.989	1.972	1.987	2.047	2.045	2.027	1.631
Sd	1.131	1.120	1.137	1.090	1.086	1.131	1.130	1.119	0.902

Table 6: Risk Ratio Sensitivity to an Excluded Model - VaR 99%

Excluded Model	None	HS	EWMA	Normal	Student-t	GARCH N	GARCH St	EVT	MC
Max	7.062	7.062	7.062	7.062	5.787	7.062	7.062	6.157	7.062
Mean	1.756	1.750	1.669	1.636	1.633	1.754	1.750	1.746	1.716
Sd	0.976	0.977	0.930	0.880	0.881	0.975	0.973	0.958	0.982

Table 7: Risk Ratio Sensitivity to an Excluded Model - ES 97.5%

Overall, ES displays less variability and a lower mean than VaR, as ES is more stable. Yet, the maximum risk ratio is higher than VaR. Surprisingly, we notice that this value is produced by the ratio between Student-t and EVT, which are two fat-tailed models and should produce more conservative values. The maximum risk ratio for VaR is the ratio between Student-t and MC, which is what we could have expected, as we previously analysed the maximum/minimum.

One can also notice that, if EVT often produces the highest estimates, the mean of risk ratios is sparsely changed. The standard deviation is yet decreased. We observe that excluding MC leads to lower means - the lowest for VaR - which is logical as MC produced most of the minimal risk measures.

From the risk ratio implementation, we note that most of the fluctuations between the models are observed during the 2008 crisis. As a consequence, in the following analysis, we focus on this period.

2.3 Worst-case Approach

We implement a method presented in *Model Risk and Regulatory Capital* (2002) [22], called worst-case approach. As its name suggests it, this methods consists in taking the worst-case situation as a reference and then comparing it to the model of study.

Definition 2.3. Assume that we have a class $\mathcal{K}$ of different risk measures whose forecasts are given by RM^k. The worst-case market risk measure is computed via: $sup_{i\in\mathcal{K}}\{RM^i\}$. We define the worst-case model risk for the model k [22]:

$$WC_k := sup_{i\in K}\{RM^i\} - RM^k.$$

Remark 2.4. In *Model Risk and Regulatory Capital*, a subset of $\mathcal{K}$ is introduced in order to take into account the choice of the models. Hence, the model risk measure is defined according to a risk measure k and a subset of $\mathcal{K}$ that contains k. In order to analyse all the models we implemented, we assume that the subset of $\mathcal{K}$ is actually $\mathcal{K}$.

As seen in Section 2.2, we observed more fluctuations in the models during the 2008 crisis. This period of financial stress is appropriate to assess the strength of our models. Hence, we choose to focus on the period from 14-04-2008 to 01-07-2011 for this study.

We obtain the following results for the worst-case model risk measure:

Figure 12: Worst-case approach for VaR 99%

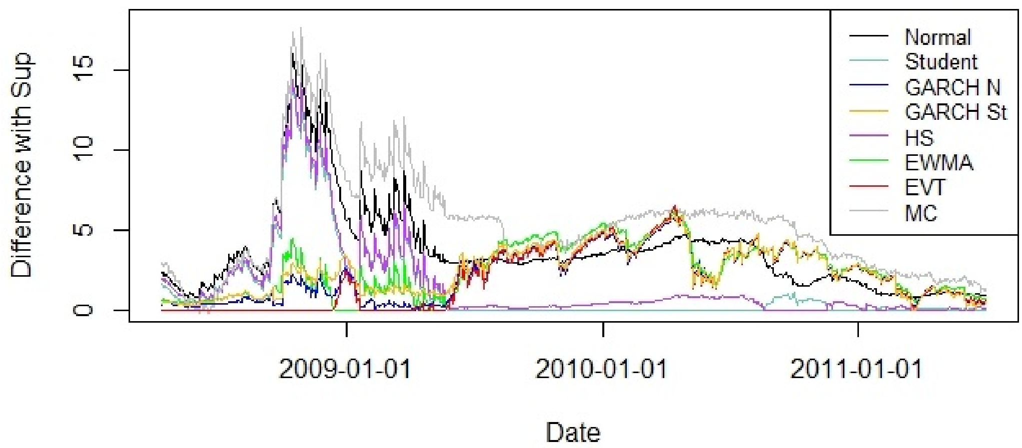
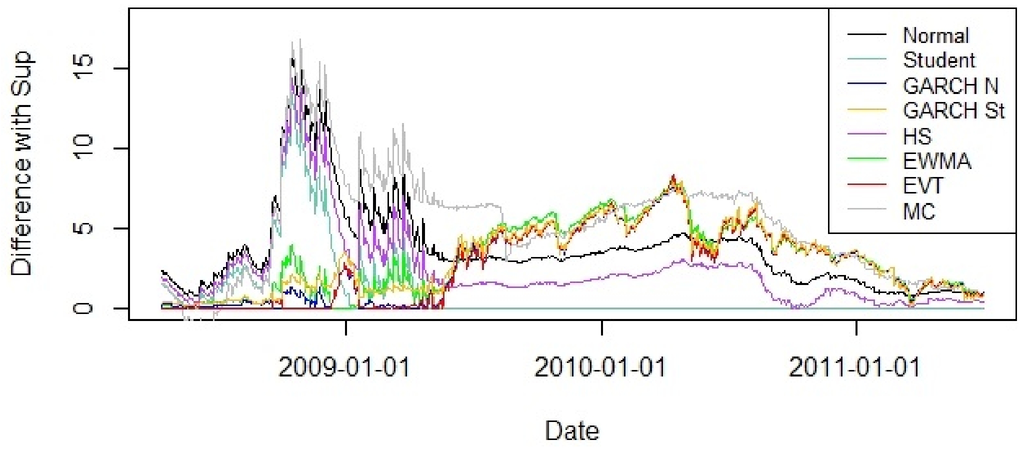


Figure 13: Worst-case approach for ES 97.5%



The maximum difference of market risk forecasts - which is equal to 17.537% for VaR and 16.889% for ES - is obtained by the MC model in the end of 2008. It is logical as MC produces most of the minimum values. We observe that four models have very similar results for both VaR and ES: EVT, GARCH N, GARCH St and EWMA. Student-t produces quasi-null model risk measures after mid 2009, as it is most of the time the maximum market risk estimate.

The worst-case model risk measure has more fickle results between the last third of 2008 and the first third of 2009. After this period, the outputs are more constant, being all included between 0 and 8%.

A drawback of this method is that it is not robust. Indeed, the maximum market risk estimate may be produced by a model that would have been eliminated during the backtest step. It is therefore dependent on the set of selected models. Moreover, if our set of models does not contain a conservative model, such as Student-t and EVT, then the values of this model risk measure are low and do not capture the risk of some models.

2.4 Benchmark Model

In this section, we want to quantify model risk thanks to a benchmark model [3]. In addition to quantifying the level of divergence between the models, we would like to assess trends in the models' outputs, and to answer the question: are some models systematically under or over estimating the observed losses? Ideally, a market risk forecast should indeed be preventing the observed loss. Hence, we would prefer to slightly overestimate the portfolio loss rather than the contrary.

Definition 2.5. Assume we have K models of risk measure and each model gives a forecast RM^k. We take the model m as the benchmark. We define a model risk β associated to the measure of study RM^k and the benchmark model RM^m:

$$\beta(m, k) := RM^k - RM^m.$$

In this section, we compute the measure β with two types of benchmark model:

- a benchmark model selected according to its backtest results,
- a benchmark model estimated from all models, based on a Bayesian approach.

As we did previously, we choose to focus on the 2008 crisis which contains more variations in the P&L than other times of the forecasting window.

2.4.1 Choice of Benchmark Model

In Section 2.1, we assessed the quality of the risk measures. We decide to take two models that presented good backtest results as benchmarks:

- Extreme Value Theory
- Filtered Historical Simulation with EWMA.

In this section, we analyse the model risk of our eight models by comparing their values to one of the two models mentioned above. It enables us to evaluate if some models are constantly underestimating the risk measures or on the contrary, if they tend to overestimate the values.

We implement the model risk β by looking at the difference between the model of study and the benchmark. We first analyse VaR models.

Figure 14: Comparison with EVT Benchmark for VaR 99%

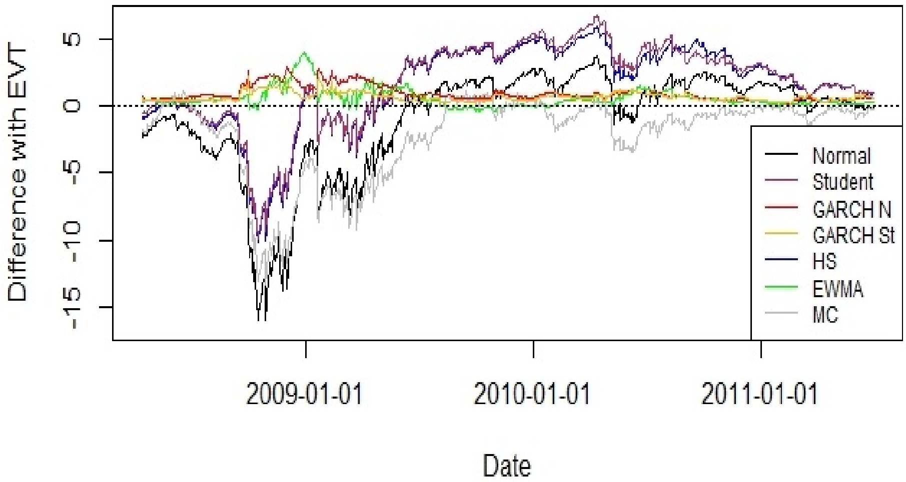
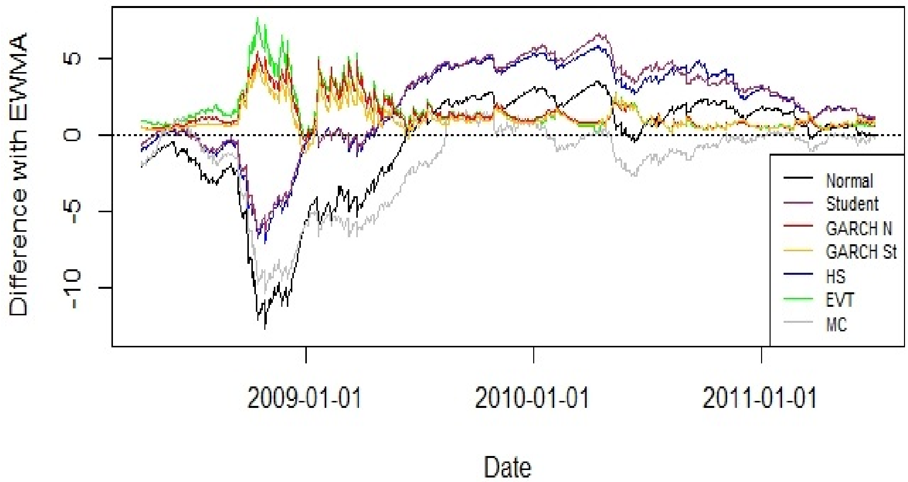


Figure 15: Comparison with EWMA Benchmark for VaR 99%



First, we observe that Student-t and HS models provide similar results for both EVT and EWMA benchmarks. It is interesting as we observed in Section 1 that the graphs for ES are quite different for the two methods. The two GARCH models have very similar model risk measures as well.

One can notice that the most important peaks are negative, which implies that the biggest gaps are observed when models are underestimating the benchmark. This can be dangerous for an unprepared financial institution, as we have already mentioned. HS, Student and Normal underestimate both the EVT and EWMA benchmarks until approximately mid 2009. Then, they overestimate the benchmarks. We notice that MC is almost always underestimating the benchmarks.

Next, we focus on ES models:

Figure 16: Comparison with EVT Benchmark for ES 97.5%

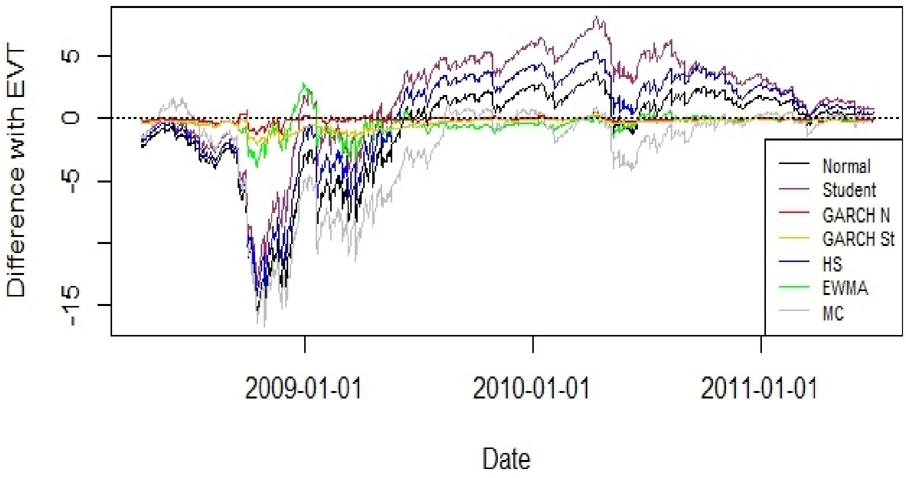
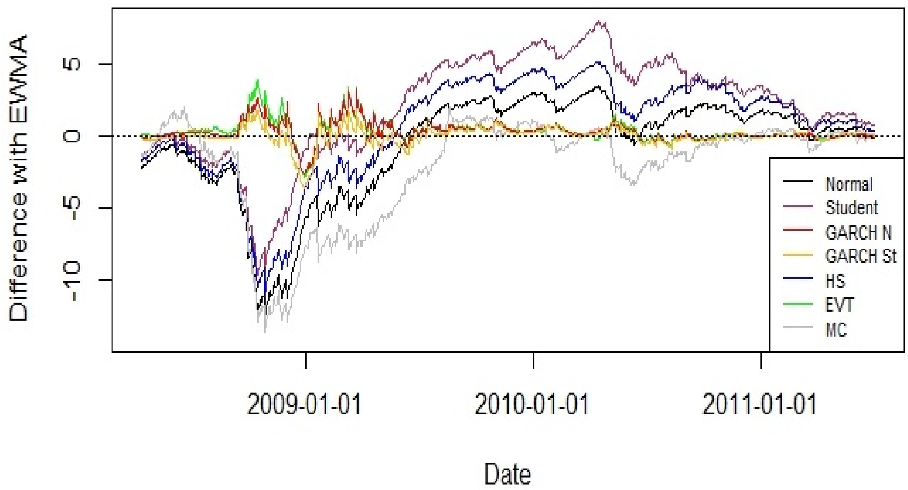


Figure 17: Comparison with EWMA Benchmark for ES 97.5%



We observe that with the EWMA benchmark, the two GARCH models and the EVT model give similar results. HS and Student-t are not comparable this time. This is explained by the fat tails of the Student-t distribution. Again, the higher peaks are observed for negative values, at the end of 2008. Finally, as for VaR models, we notice that MC is almost systematically underestimating the benchmarks.

Choosing EWMA and EVT was arbitrary, only based on the backtests and we said that these results should be carefully interpreted. We need another way to select the benchmark model.

2.4.2 Bayesian Model Averaging

Bayesian model averaging refers to a method that forecasts a predictive value which is estimated from a weighted average of several models outputs [34]. The weights are based on prior information. Hence, rather than choosing a reference model, which constitutes a source of risk, we construct a new one, by using this Bayesian approach.

A framework of this statistical point of view can be defined as follows [9]:

1. Before estimating a risk measure for the time $t+1$, we have some *prior knowledge*, based on the prior information, at time t.
2. We are able to analyse the data and thus extract relevant information.
3. We can update our prior knowledge and obtain a *posterior information.*

Assume that we have K models for calculating the risk measure. Each model has a set of parameters θ_k - those sets can be different according to the nature of the model. In order to assess a Bayesian model averaging, we need to have access to two pieces of information [13]:

- prior density $p(\theta_k|M_k)$ which contains our beliefs on the model parameters θ_k, given that the model M^k holds.
- prior probability $\mathbb{P}(M_k)$ that the model k is the "true model".

Definition 2.6. Let us assume we have a set of observations x. $p(x|M_k)$ is the integrated likelihood of the data for the model M_k. We define the posterior probability for the model M_k given x [13]:

$$\mathbb{P}(M_k|x) := \frac{p(x|M_k)\mathbb{P}(M_k)}{\sum_{j=1}^{K} p(x|M_j)\mathbb{P}(M_j)}.$$

Remark 2.7. $p(x) = \sum_{j=1}^{K} p(x|M_j)\mathbb{P}(M_j)$ corresponds to the uncertainty in the data x [9]. We can condense the previous definition as a proportionality expression: $\mathbb{P}(M_k|x) \propto p(x|M_k)\mathbb{P}(M_k)$.

In order to obtain the Bayesian averaging estimate, we calculate the risk measures for the different models, then we compute the posterior probabilities and finally we take an average of the resulting risk measures weighted by its posterior probability [32].

Definition 2.8. The Bayesian model average of the risk measures RM given the K models and the set of observations x is defined as:

$$\widehat{RM}(x) := \sum_{k=1}^{K} RM^k \mathbb{P}(M_k|x). \tag{2.1}$$

In order to apply this Bayesian methodology, we need to assume that both the model and the parameters are uncertain. Furthermore, a major difficulty consists in measuring the priors [34]. We develop a simplified method that is linked to this Bayesian approach, as we use *prior knowledge* to obtain a posterior probability that is used to weight our risk measure results.

Given a set of observations x, we want to assess the posterior probability $\mathbb{P}(M_k|x)$ that the model k is the most accurate.

Definition 2.9. Consider the portfolio's loss L_t at time t and K models of risk measures. RM_t^k is the risk forecast of the model k at time t. We define the weight associated to the risk measure k at time t:

$$w_t^k := \begin{cases} W_1, & L_t + 10\%L_t < RM_t^k \\ W_2, & L_t \leq RM_t^k \leq L_t + 10\%L_t \\ W_3, & L_t - 10\%L_t \leq RM_t^k < L_t \\ W_4, & RM_t^k < L_t - 10\%L_t. \end{cases}$$

This weights' breakdown allows us to rate our models on a scale of importance, according to their output values. For each output, we analyse if the value is higher than the observed loss and if it is overestimating or underestimating this loss by more than 10%. This is how we define the four cases defined in Definition 2.9.

Remark 2.10. It is worth noticing that the weight w_t^k is our *prior knowledge* of the model k. The weight values are based on our beliefs.

Definition 2.11. Given the weight w_t^k of the model k at time t, we define the posterior probability $\mathbb{P}(M_{t+1}^k|x)$ that the model k is the best model given the observations x:

$$\mathbb{P}(M_{t+1}^k|x) := \frac{p_t^k}{\sum_{i=1}^K p_t^i}$$

$$\text{with } p_t^i := \sum_{j=t-99}^{t} \frac{1}{i-j+1} w_j^i.$$

Remark 2.12. The choice of p_t^i was influenced by the EWMA premise. Indeed, we give more weight to the most recent observations and we decrease the influence exponentially when moving away from the time t.

In order to find an optimized set of weights, we need to identify criteria that will enable us to compare them. We decide to backtest the Bayesian averaging estimates using the formula 2.1. We focus on the number of violations and on the p-values associated with the backtests. We analyse two tests for VaR: unconditional coverage test (UC) and joint test (J).

		99% VaR			97.5% ES	
Test	Weights	Violations	p_{UC}-*value*	p_J-*value*	Violations	*p-value*
1	$W_1 = W_2 = W_3 = W_4$	35	1.996e-01	1.452e-02	27	3.362e-04
2	$10W_1 = W_2 = W_3 = 10W_4$	37	1.020e-01	1.177e-02	30	9.355e-04
3	$W_1 = 10W_2 = 10W_3 = W_4$	35	1.986e-01	1.452e-02	**26**	3.234e-04
4	$W_1 = W_2 = 10W_3 = 10W_4$	33	3.526e-01	1.557e-02	**26**	5.078e-04
5	$10W_1 = 10W_2 = W_3 = W_4$	57	1.313e-06	2.841e-06	43	2.540e-05
6	$10W_1 = W_2 = 10W_3 = 50W_4$	32	4.541e-01	1.526e-02	27	5.990e-04
7	$10W_1 = W_2 = 20W_3 = 50W_4$	32	4.541e-01	1.524e-02	27	6.096e-04
8	$20W_1 = W_2 = 10W_3 = 50W_4$	**31**	**5.714e-01**	**1.203e-01**	29	8.403e-04
9	$25W_1 = W_2 = 5W_3 = 50W_4$	34	2.677e-01	1.046e-01	31	8.449e-04
10	$20W_1 = W_2 = 10W_3 = 100W_4$	**31**	**5.714e-01**	**1.203e-01**	28	**9.503e-04**

Table 8: Comparison of 10 sets of weights for 99% VaR and 97.5% ES

We observe that the last test gives the overall best results. It was what we were expecting as a good risk measure should be higher than the observed loss value but not too far from the real outcome. This is why a risk forecast between L_t and 10% of L_t has the largest weight. Then, the risk measure underestimating the loss value by less than 10% has the second most important weight. Next, a risk measure overestimating the loss by more than 10% has the third biggest weight. Finally, a forecast underestimating the P&L by more than 10% is not put aside but we assign it the smallest weight.

Remark 2.13. The backtest results cannot be directly compared to the results in Section 2.1 as we used the first 100 days to calculate the Bayesian estimator.

We assign the following weights, with respect to the 10th test of Table 8: $W_1 = 5, W_2 = 100, W_3 = 10, W_4 = 1$. We decide to analyse the means of the weights for our eight models. It reflects the influence of a given model in this Bayesian averaging estimate.

Model	Mean W_i - VaR 99%	Mean W_i - ES 97.5%
Normal	**5.748**	**5.782**
Student-t	5.318	5.347
GARCH N	5.580	5.330
GARCH St	5.421	5.518
EVT	5.425	5.467
HS	5.522	5.391
EWMA	5.400	5.274
MC	**5.716**	**5.594**

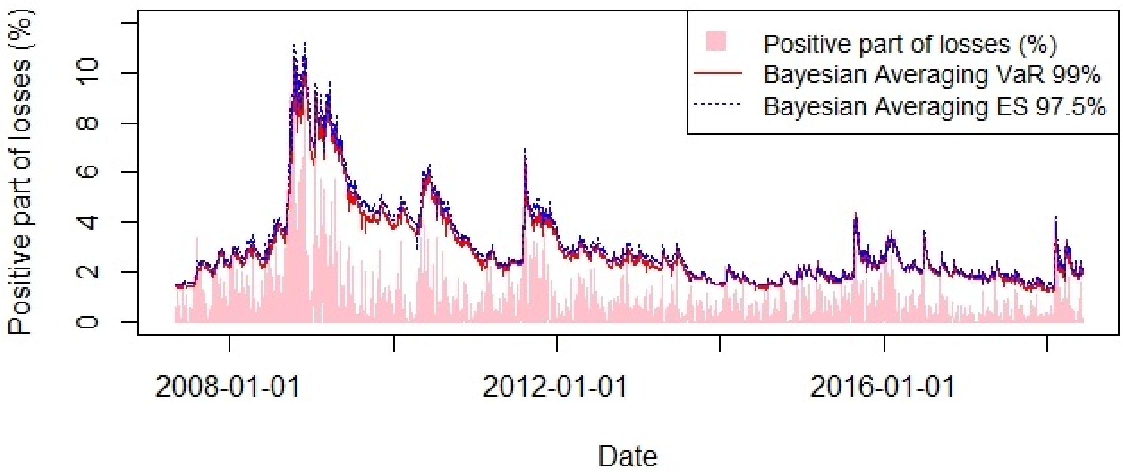
Table 9: Mean of weights used in the Bayesian Approach

For both VaR and ES, we observe that the Normal model and the MC model have the highest means in weights. Hence, in average, those two models have the greatest influence on the Bayesian estimator. On the contrary, the lowest influences are observed for the Student-t and EWMA models for the VaR, and GARCH N and EWMA models for the ES.

Remark 2.14. We observe that the means are higher but close to 5, which is the value of the weight W_1. It implies that most values are over 10% of the observed losses.

With the weights chosen above, we compute the Bayesian averaging estimate. We compare it to the observed losses below:

Figure 18: Bayesian Averaging estimate for VaR 99% and ES 97.5%



We observe a good tracking of the losses' volatility. Very few violations are observable. We can finally compute the model risk measure β with $\widehat{RM}$ taken as the benchmark:

Figure 19: Comparison with Bayesian Benchmark for VaR 99%

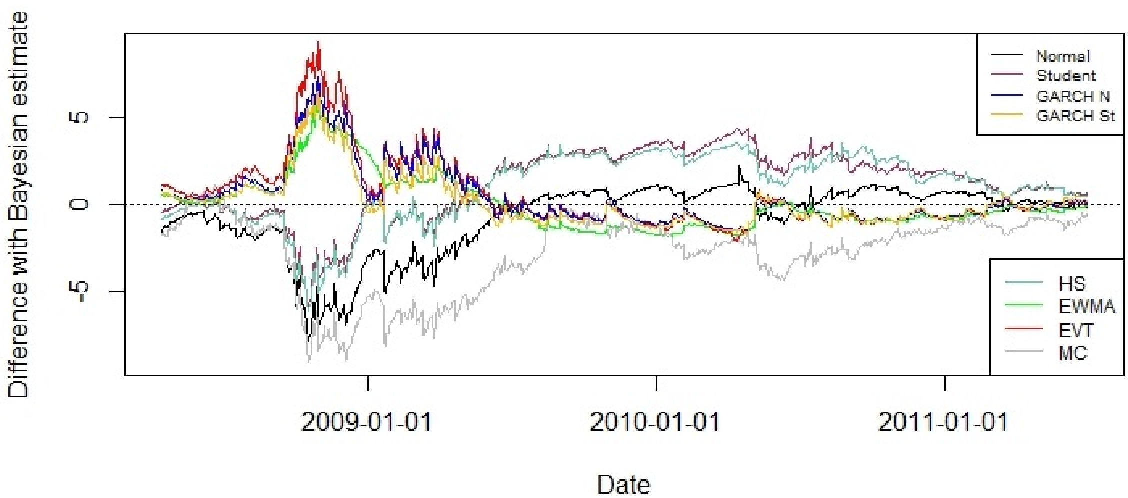
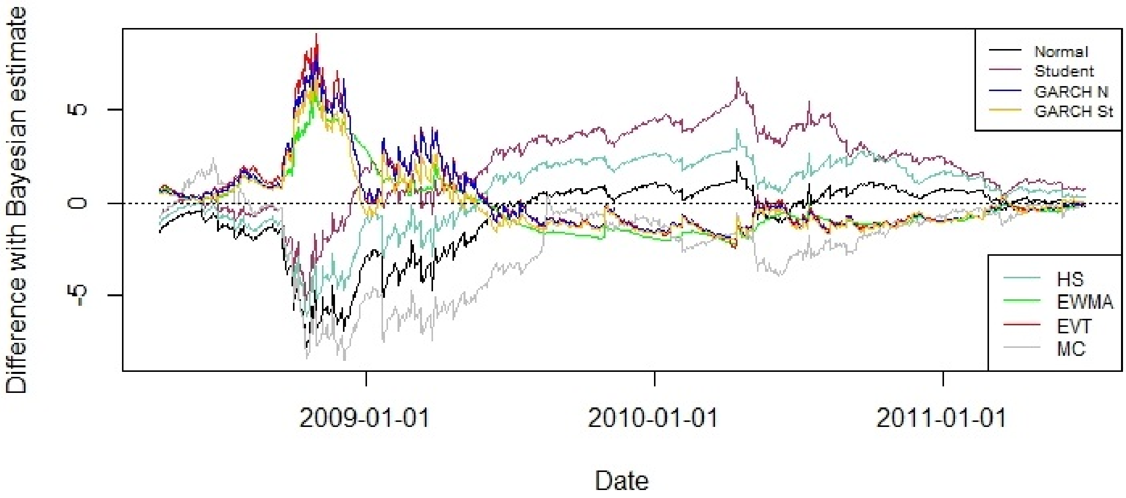


Figure 20: Comparison with Bayesian Benchmark for ES 97.5%



As we took a weighted average over all models, we observe that the differences are less important than for the EVT and EWMA benchmarks (between -9 to 9%). Moreover, the positive and negative values are similar, as the average smoothed the gaps. Most of the fluctuations are located at the end of 2008, like in the worst-case analysis and in the arbitrary-chosen benchmark comparison: half of the models is underestimating the Bayesian result and half is overestimating it.

Although the Bayesian averaging estimate is a fair risk measure and enables us to assess model risk, it requires to have a certain number of models at the outset, which is not systematically the case for financial institutions. Moreover, the weights were chosen from a set of ten tests: an optimization process should be implemented in order to minimize the number of violations. Finally, we note that this measure is not very robust as changing one model has repercussions on the Bayesian estimate and thus on the model risk measure β.

2.5 Correlation Analysis

After implementing different measures of model risk, for market risk models, we would like to analyse their correlation [14]. Indeed, we want to observe if the measures are going in the same direction and are, in a sense, giving the same warnings.

First, we need to compare the quantification methods of model risk that we implemented in Sections 2.2, 2.3 and 2.4. Risk ratio is providing a unique measure for the eight market risk models. The worst-case approach and the Bayesian benchmark are giving a measure for each market risk model. EWMA and EVT benchmarks produce seven measures as one of the model is used as the reference. The number of measures being different according to the model risk approach, we cannot analyse the correlation directly.

Before analysing the dependence among the model risk approaches, we focus on the correlation within one model at a time. We work on the whole forecasting period. We display the results for ES 97.5% in Appendix A. We define a scale of color in order to highlight the large correlations (positive or negative).

-1 to -0.75	-0.75 to -0.40	-0.40 to 0.40	0.40 to 0.75	0.75 to 1
(red)	(light red)	(white)	(light green)	(green)

Table 10: Color Scale for Correlation Matrices

We start with the worst-case approach. The correlation matrix is shown below:

	Normal	Student	Garch N	Garch St	HS	EWMA	EVT	MC
Normal	1	0.51	0.39	0.45	0.74	0.16	0.17	0.86
Student-t	0.51	1	-0.23	-0.19	0.55	-0.46	-0.41	0.22
GARCH N	0.39	-0.23	1	0.92	-0.03	0.58	0.88	0.55
GARCH St	0.45	-0.19	0.92	1	0.01	0.68	0.77	0.66
HS	0.74	0.55	-0.03	0.01	1	-0.29	-0.23	0.43
EWMA	0.16	-0.46	0.58	0.68	-0.29	1	0.50	0.49
EVT	0.17	-0.41	0.88	0.77	-0.23	0.50	1	0.36
MC	0.86	0.22	0.55	0.66	0.43	0.49	0.36	1

Table 11: Correlation Matrix for the Worst-Case Measures - VaR 99%

Overall, we observe that the correlations are positive and quite high. Especially, we notice a high dependence between the two GARCH models, due to their similar implementations. On the contrary, HS and GARCH St are not correlated at all.

We go on with the Bayesian estimate comparison:

	Normal	Student	Garch N	Garch St	HS	EWMA	EVT	MC
Normal	1	0.91	-0.70	-0.72	0.93	-0.76	-0.73	0.32
Student-t	0.91	1	-0.70	-0.75	0.97	-0.78	-0.73	0.05
GARCH N	-0.70	-0.70	1	0.94	-0.71	0.74	0.94	-0.08
GARCH St	-0.72	-0.75	0.94	1	-0.76	0.80	0.91	-0.05
HS	0.93	0.97	-0.71	-0.76	1	-0.80	-0.74	0.09
EWMA	-0.76	-0.78	0.74	0.80	-0.80	1	0.78	-0.04
EVT	-0.73	-0.73	0.94	0.91	-0.74	0.78	1	-0.05
MC	0.32	0.05	-0.08	-0.05	0.09	-0.04	-0.05	1

Table 12: Correlation Matrix for the Bayesian Benchmark Measures - VaR 99%

We easily notice that the correlations are very large - half of the models is correlated over 75% in absolute value. The negative correlations are due to the fact that half of the market risk models is underestimating the benchmark while the other half is overestimating it. We have similar results for the ES models.

Next, we analyse the two benchmark models - EVT and EWMA:

	Normal	Student	Garch N	Garch St	HS	EWMA	MC
Normal	1	0.95	-0.11	-0.18	0.96	-0.40	0.42
Student-t	0.95	1	0.04	-0.05	0.99	-0.32	0.27
GARCH N	-0.11	0.04	1	0.86	0.02	0.28	-0.67
GARCH St	-0.18	-0.05	0.86	1	-0.08	0.42	-0.69
HS	0.96	0.99	0.02	-0.08	1	-0.33	0.29
EWMA	-0.40	-0.32	0.28	0.42	-0.33	1	-0.44
MC	0.42	0.27	-0.67	-0.69	0.29	-0.44	1

Table 13: Correlation Matrix for the EVT Benchmark Measures - VaR 99%

	Normal	Student	Garch N	Garch St	HS	EVT	MC
Normal	1	0.97	0.26	0.21	0.97	0.12	0.49
Student-t	0.97	1	0.34	0.28	0.99	0.20	0.38
GARCH N	0.26	0.34	1	0.95	0.34	0.91	-0.27
GARCH St	0.21	0.28	0.95	1	0.27	0.85	-0.32
HS	0.97	0.99	0.34	0.27	1	0.19	0.39
EVT	0.12	0.20	0.91	0.85	0.19	1	-0.35
MC	0.49	0.38	-0.27	-0.32	0.39	-0.35	1

Table 14: Correlation Matrix for the EWMA Benchmark Measures - VaR 99%

We observe that the correlations are less important than for the Bayesian benchmark. It can be explained by the fact that the Bayesian estimate is taking, in average, every model and therefore each measure of β takes into account a certain proportion of each market risk models.

Now, we compare the correlation between the model risk measures we implemented. We cannot compare the worst-case for the MC model with the Bayesian approach measure for the EVT model as the market risk models have different trends. We have to focus on one market risk model at a time. Overall, the Normal model has the highest correlations for each model risk approach and is therefore a fair representative of the different model risk models. As a consequence, we display the correlation between the model risk measures, for the Normal market risk model:

	Risk Ratio	Worst-Case	Bayesian	EVT	EWMA
Risk Ratio	1.00	0.47	0.32	0.46	0.48
WorstCase	0.47	1.00	-0.43	-0.26	-0.27
Bayesian	0.32	-0.43	1.00	0.91	0.90
EVTBenchmark	0.46	-0.26	0.91	1.00	0.96
EWMABenchmark	0.48	-0.27	0.90	0.96	1.00

Table 15: Correlation Matrix for the Model Risk Measures of VaR 99% Models

	Risk Ratio	Worst-Case	Bayesian	EVT	EWMA
Risk Ratio	1.00	0.79	0.01	0.13	0.14
WorstCase	0.79	1.00	-0.38	-0.23	-0.22
Bayesian	0.01	-0.38	1.00	0.93	0.91
EVTBenchmark	0.13	-0.23	0.93	1.00	0.95
EWMABenchmark	0.14	-0.22	0.91	0.95	1.00

Table 16: Correlation Matrix for the Model Risk Measures of ES 97.5% Models

As one could have forecasted, the three benchmarks give high correlation values. This is linked to their similar structures. The worst-case approach gives negative correlations with every other model except the risk ratio, to which it is quite well correlated. We explain it by the fact that the benchmarks can produce negative measures contrary to the worst-case approach. Finally, we observe that the risk ratio method gives lower correlations with the benchmarks for ES than for VaR. To conclude, the model risk measures that we computed are going in similar directions.

2.6 Models' Sensitivity

A sensitivity analysis is usually performed by changing a particular feature of a model and then observing the differences in the model results. In particular, one can modify the parameters of the model or the methodology [22]. When changing a parameter, the sensitivity analysis gives a "measure of the impact" of an input on the model output [22]. We examine two measures:

- the percentage of violations
- the standard deviation of the outputs.

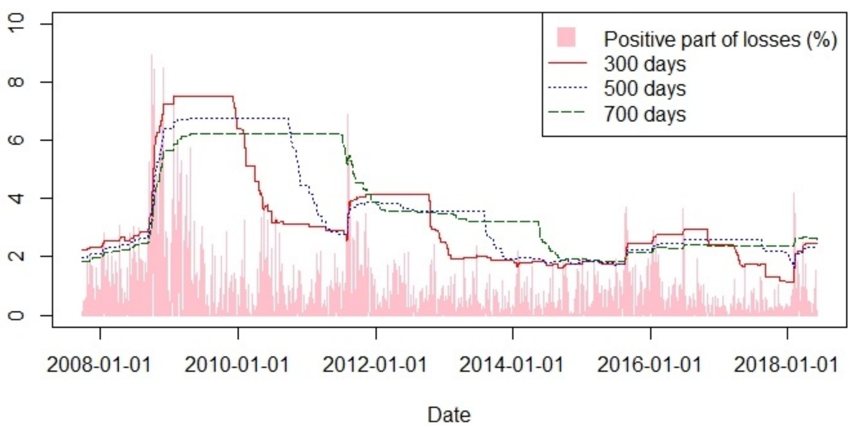
The violations are directly linked to the regulations and more precisely on the capital requirements. As a consequence, it is an important variable in the model risk analysis. The standard deviation enables to measure how a model diverges from the average risk forecast. Hence, it is a credible measure of model risk [14].

We have to keep in mind that the models we chose have not the same parameters. We focus on four models to assess their sensitivity to a particular parameter:

1. HS: sensitivity to the number of days in the rolling window
2. EWMA: sensitivity to α_1 in the EWMA volatility scheme 1.1
3. EVT: sensitivity to the threshold u which impacts the excess distribution
4. MC: sensitivity to the number of simulations.

We begin with the sensitivity analysis of HS. We select a lower and a higher number of days in the rolling window: 300 and 700 days. We observe similar results for VaR 99% and ES 97.5%. The results for VaR are shown in the Appendix A. We display below the ES analysis:

Figure 21: HS ES 97.5% Sensitivity Analysis to the Number of Days in the Rolling Window



We observe, as we were expecting, that the longer the rolling window, the slower the track of the observed losses. Indeed, HS gives the same weight to each day of the forecasting window, so with a larger window, it takes more time to 'forget' the high values. HS with 700 days gives forecasts that are far from the reality. Indeed, when the 2008 crisis appears, this model takes more time than HS with 300 or 500 days to increase its risk forecasts. Furthermore, its values are in average lower than the other models' results - which are below the observed peaks. The 300 days HS reacts quicker than the two other models but seems to produce more exceptions.

Remark 2.15. The Figure 21 is a good example of the "ghost feature" that we mentioned in Remark 1.12.

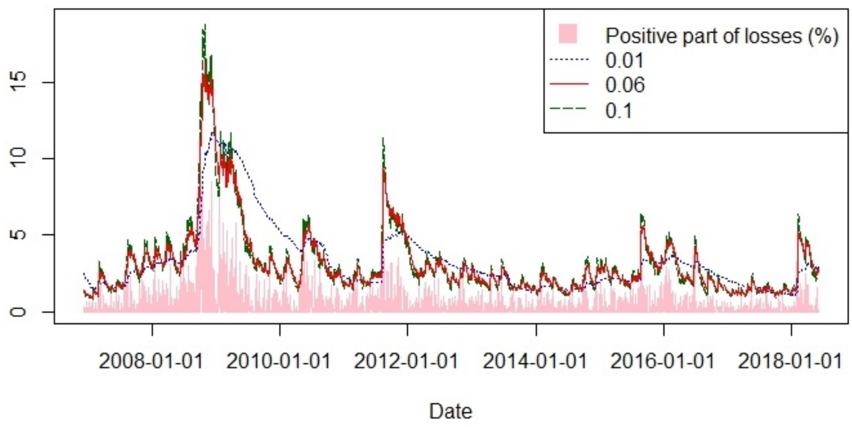
	% Violations		% Standard Deviation	
Number of days	VaR 99%	ES 97.5%	VaR 99%	ES 97.5%
300 days	1.899%	1.657%	1.921%	1.930 %
500 days	1.829%	1.588%	2.006%	1.966%
700 days	1.568%	1.520%	1.961%	2.060%

Table 17: HS Sensitivity to the Number of Days of in the Rolling Window

As predicted with Figure 21, HS with 300 days gives more violations than HS with 500 days and 700 days. However, it gives lower standard deviations. HS with 700 days has less violations than HS with 500 days while their standard deviations are similar. Yet, one should remember that the loss tracking is less precise: HS with 500 days - the model we chose - is the most robust.

In a second step, we come back on the filtered historical simulation model - EWMA - and we analyse its sensitivity with respect to the parameter α_1 in the volatility recursive formula 1.1.

Figure 22: EWMA ES 97.5% Sensitivity Analysis to α_1 in the Volatility Scheme



We observe that with the lowest value of α_1 the tracking is less precise and the values are more optimistic. On the contrary, EWMA with $\alpha_1 = 0.1$ gives more fickle outputs, with its highest value equal to approximately 18% during the 2008 crisis while $\alpha_1 = 0.01$ only produces a risk forecast near 12%. We now analyse the violation and standard deviation results:

	% Violations		% Standard Deviation	
Alpha	VaR 99%	ES 97.5%	VaR 99%	ES 97.5%
$\alpha_1 = 0.01$	1.381%	1.139%	2.311%	2.412%
$\alpha_1 = 0.06$	1.346%	1.139%	2.350%	2.426%
$\alpha_1 = 0.10$	1.312%	1.174%	2.386%	2.573%

Table 18: EWMA Sensitivity to the Value of Alpha

As one could have expected, a higher value of alpha leads to a higher standard deviation, as the "intensity of reaction" is increased (Remark 1.8). Nevertheless, it implies that it is less stable. We notice here that the number of violations do not fluctuate significantly when we modify α_1. EWMA with $\alpha_1 = 0.06$ is the right balance: it gathers a precise track of the losses' change with a reasonable standard deviation.

Next, we analyse the EVT model by modifying the threshold u. We display the graphs in Appendix A. Let us notice that globally, when $u = 0.99$, VaR is higher than ES while, for the other thresholds, the risk measures give similar forecasts.

	% Violations		% Standard Deviation	
Threshold	VaR 99%	ES 97.5%	VaR 99%	ES 97.5%
$u = 0.85$	1.346%	1.312%	2.562%	2.615%
$u = 0.93$	1.208%	1.208%	2.574%	2.627%
$u = 0.99$	1.105%	2.761%	2.673%	2.301%

Table 19: EVT Sensitivity to the Value of the Threshold u

We observe that EVT with $u = 0.99$ produces less violations for VaR than the two other thresholds. However, the number of exceptions is more than double for ES. The standard deviations of $u = 0.85$ and $u = 0.93$ are similar - slightly lower for 0.85 - but the violations are better for 0.93. Hence, $u = 0.93$ is a sound choice.

Finally, we move on to MC and its sensitivity to the number of simulations. The graphs are presented in Appendix A. We obtain the following results for the violations and the standard deviations:

	% Violations		% Standard Deviation	
Number of simulations	VaR 99%	ES 97.5%	VaR 99%	ES 97.5%
10 000 simulations	4.729%	2.106%	1.212%	1.367%
100 000 simulations	4.487%	2.175%	1.216%	1.372%
500 000 simulations	4.349%	2.209%	1.215%	1.372%

Table 20: MC Sensitivity to the Number of Simulations

We observe very similar results for the three models. As expected, with less simulations, we have less volatility for the outputs. For VaR especially, the number of violations does not seem to converge: the simulations should be even more increased. We only conclude that the ES is giving risk forecasts that present significantly less violations than the VaR measures.

The standard deviation is giving us information on how the outputs are dispatched, compared to the outputs' mean. However, it is not a precise model risk measure as the standard deviation of the market risk forecasts has to be compared to the standard deviation of the P&L to make sense.

With this sensitivity analysis, we understand that the parameter calibration is an essential part of model implementation. This is especially true as all parameters do not have the same sensitivity and therefore do not impact the outputs with the same intensity.

2.7 Model Risk Management Framework

The objective of this section is to build a general model risk framework. Regulations such as SS11-7 [33] and FRTB [8] are delivering guidance and advice on how to manage model risk. We try to gather processes and concepts to create a comprehensive framework.

As mentioned in the Introduction, the Supervisory Guidance SR 11-7 gives three main steps that can impact model risk. We keep the main expressions as in the text.

1. The model development, implementation and use,
2. The model validation,
3. The model governance.

Let us focus on each stage [33]. Model development contains notably the statement of purpose and the theory. The implementation consists on several elements such as the methodology, the numerical methods, the chosen approximations and the data assessment. It leads naturally to the model use. It enables to assess the performances of the model over time. Guaranteeing strong competences for the development and implementation team, as well as a complete and regular training of the users is essential in this first part of model management.

Model validation should be processed with an independent point of view: the validation team should not have interest in accepting a model. Three main goals are quoted in SR 11-7. The first one is to assess the theoretical concepts. The second objective is to conduct a continuous monitoring of the model. Finally, the validation step should evaluate the outputs' accuracy [33].

Last but not least, model governance must be conducted by the highest level of management to ensure a proper influence and diffusion of the decisions. The main roles of the governance are to define the risk management procedures, to allocate roles and responsibilities, and finally to make sure that the policies are implemented. The governance can also include the intervention of the "internal audit" and the use of "external resources" [33]. This management work has to be documented in order not to repeat the same mistakes.

We now introduce a widespread concept in risk governance: the three lines of defence [15] [17]. It describes the risk management structure that should be employed.

1. The first line of defence is represented by the *business units* which can be understood as an operational function. It consists on a daily control of model risk with the implementation of predefined processes.

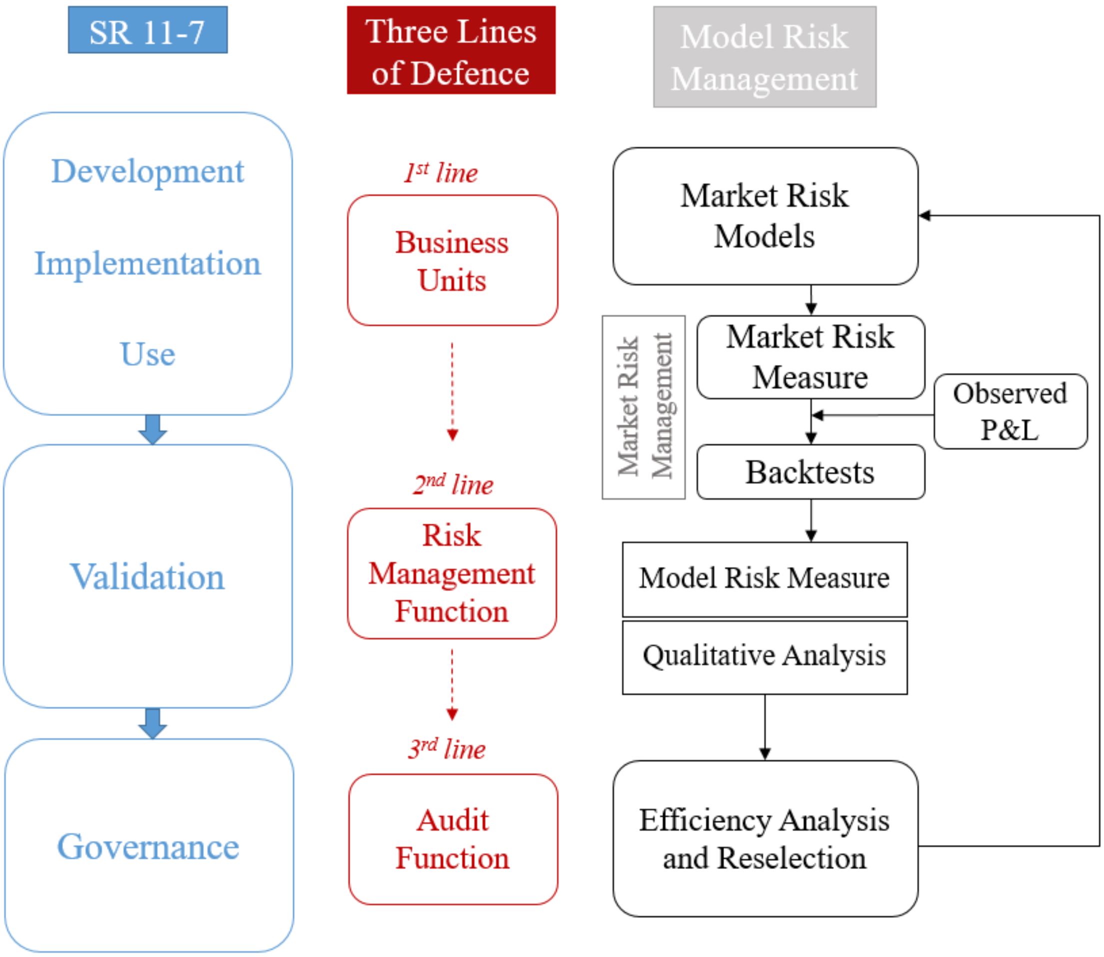
2. The second line is provided by the *risk management team.* The risk management is overseen and advice can be given to the first line.

3. Finally, the third line is composed by the *audit function.* It is oriented toward procedures and controls of the two first lines.

With this brief description of the two frameworks, we conclude that the three lines of defence are included in the SS 11-7 management framework. Indeed, the three levels of control are mainly focusing on responsibilities and not on the processes. Yet, it is worth noticing that the two frameworks can be put in parallel as the three elements are coherent one at the time.

In this way, we consider a model risk management framework applied for market risk. We take into account the environment presented in *Model Risk* (2010) in Chapter 15 [31] and we align it to the frameworks we saw previously.

Figure 23: Model Risk Management Framework



The first step of model risk management for market risk is to define the model, to state the hypothesis such as risk factor distribution [31] and then to implement it. The calibration of the parameters is an important stage of the implementation. Next, the market risk is estimated by the user and backtested after the observation of the portfolio position. This analysis gives a first hint on the model quality. The third step consists in conducting a model risk evaluation via the implementation of a quantitative measure - such as risk ratio, worst-case approach or benchmark comparison. It can be completed by a more qualitative analysis of the model risk outputs, in order to step back on the results and to compare it to the backtests. The conclusions of this step are then analysed, and potential remarks and model improvements can lead to an impact on capital requirements [31].

In practice, all the measures and tests should be reviewed regularly, and the models' assets and drawbacks should be conveyed to the whole hierarchy [31]. Finally, a key factor of a model risk management relies on the permanent challenge of the models. Each step should be conducted in order to challenge the models.

The Supervisory Guidance SR11-7 specifies that, in particular situations, a quantitative evaluation of model risk is not possible. A solution would be to modify the inputs and calculations so that the model results are more conservative. Another possibility would be to integrate a qualitative assessment of model uncertainty to the model risk analysis. It would be a complement to the numerical outputs, which would then be less emphasized [33].

Conclusion

In this project, we introduced and quantified model risk, applied to market risk models. We started with the implementation of eight VaR and ES methods, which are the most widespread market risk forecasts in the industry. Then, we backtested our results and we applied three different measures of model risk: risk ratio, worst-case approach and benchmark comparison.

A general remark regarding those measures is their instability: they all depend on the set of market risk models, chosen at the beginning of the project. Removing a model from this set can impact to a certain extent the final outputs and thus change the capital requirements of the institution in question. A great difference between the three model risk measures is that the risk ratio provides one single metrics for the whole set of models, whereas the other measures give one output per model. The Bayesian benchmark can also be differentiated as its implementation is less straightforward than the other risk measures.

The measures we developed in this thesis enabled us to assess the risk of suffering losses due to the use of models. In particular, we evaluated the discrepancies among models' outputs and we observed that those gaps are greater in times of financial stress. Model risk should also increase with the complexity and the prospective impact of models [33]. Moreover, as models have different parameters and implementations, model risk should be decomposed to capture those differences. For instance, the calibration, convergence and sensitivity influence on model risk could be subject to additional studies. Furthermore, model risk should be assessed in an aggregated manner, as most financial institutions are evolving in multi-model environments. Therefore, it should be calculated once rather than added independently from different units' results. As a consequence, the approaches we presented are not sophisticated enough to be a comprehensive measure of model risk.

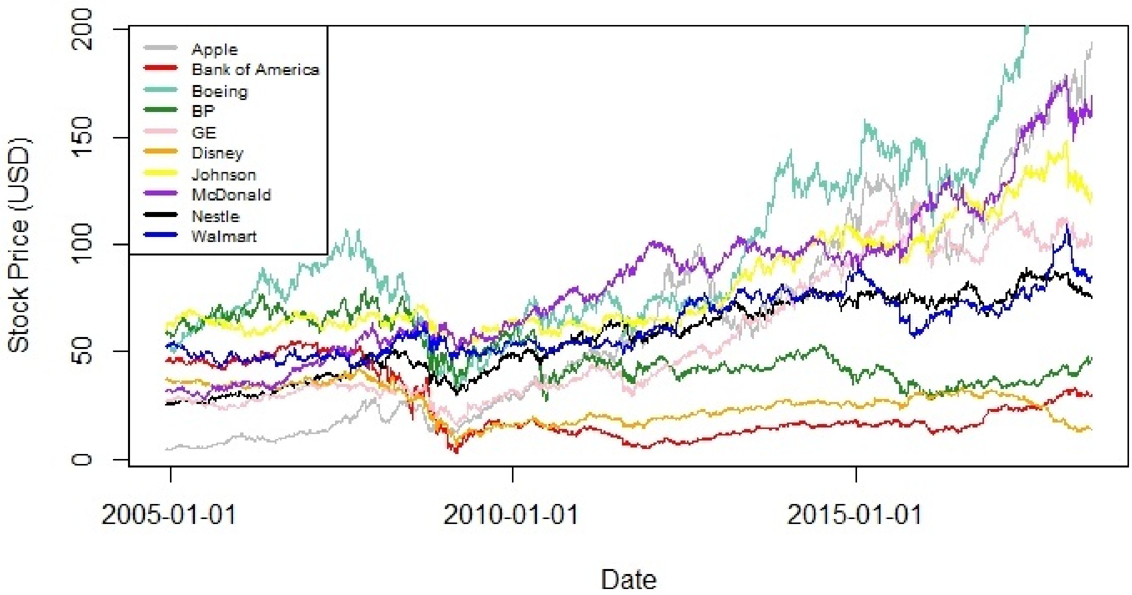
« All models are wrong » as George E. P. Box's aphorism states. A model approximates the reality, yet it is possible to optimize its risk thanks to an exhaustive framework. High model risks should be compensated by a strong management involvement [10]. More than that, it would be better to develop more efficient models than to try to remove negative effects a posteriori [31, Chapter 15].

Finally, this project was the opportunity to compare VaR at 99% and ES at 97.5%, as FRTB requires banks to move from the first one to the other. Overall, ES produces more conservative risk forecasts than VaR and the backtest results are stronger - less violations are observed. We also note that model risk measures are similar for both VaR and ES models.

A Appendix

Risk Ratio

Figure 24: Stock Prices of the Portfolio



Correlation Analysis

	Normal	Student	Garch N	Garch St	HS	EWMA	EVT	MC
Normal	1.00	0.42	0.36	0.50	0.89	0.26	0.26	0.83
Student-t	0.42	1.00	-0.41	-0.35	0.54	-0.59	-0.50	0.05
GARCH N	0.36	-0.41	1.00	0.92	0.15	0.63	0.95	0.53
GARCH St	0.50	-0.35	0.92	1.00	0.26	0.74	0.88	0.71
HS	0.89	0.54	0.15	0.26	1.00	0.02	0.06	0.61
EWMA	0.26	-0.59	0.63	0.74	0.02	1.00	0.61	0.57
EVT	0.26	-0.50	0.95	0.88	0.06	0.61	1.00	0.46
MC	0.83	0.05	0.53	0.71	0.61	0.57	0.46	1.00

Table 21: Correlation Matrix for the Worst-Case Measures - ES 97.5%

	Normal	Student	Garch N	Garch St	HS	EWMA	EVT	MC
Normal	1.00	0.77	-0.69	-0.68	0.90	-0.72	-0.72	0.26
Student-t	0.77	1.00	-0.69	-0.77	0.90	-0.76	-0.70	-0.23
GARCH N	-0.69	-0.69	1.00	0.95	-0.77	0.80	0.97	0.11
GARCH St	-0.68	-0.77	0.95	1.00	-0.81	0.86	0.95	0.19
HS	0.90	0.90	-0.77	-0.81	1.00	-0.81	-0.79	-0.02
EWMA	-0.72	-0.76	0.80	0.86	-0.81	1.00	0.83	0.14
EVT	-0.72	-0.70	0.97	0.95	-0.79	0.83	1.00	0.11
MC	0.26	-0.23	0.11	0.19	-0.02	0.14	0.11	1.00

Table 22: Correlation Matrix for the Bayesian Benchmark Measures - ES 97.5%

	Normal	Student	Garch N	Garch St	HS	EWMA	MC
Normal	1.00	0.90	0.30	0.15	0.95	-0.11	0.43
Student-t	0.90	1.00	0.21	0.00	0.97	-0.17	0.18
GARCH N	0.30	0.21	1.00	0.49	0.22	0.04	0.36
GARCH St	0.15	0.00	0.49	1.00	0.06	0.43	0.35
HS	0.95	0.97	0.22	0.06	1.00	-0.15	0.26
EWMA	-0.11	-0.17	0.04	0.43	-0.15	1.00	0.18
MC	0.43	0.18	0.36	0.35	0.26	0.18	1.00

Table 23: Correlation Matrix for the EVT Benchmark Measures - ES 97.5%

	Normal	Student	Garch N	Garch St	HS	EVT	MC
Normal	1.00	0.94	0.40	0.49	0.97	0.33	0.50
Student-t	0.94	1.00	0.45	0.48	0.98	0.39	0.30
GARCH N	0.40	0.45	1.00	0.91	0.40	0.96	0.09
GARCH St	0.49	0.48	0.91	1.00	0.46	0.87	0.23
HS	0.97	0.98	0.40	0.46	1.00	0.34	0.37
EVT	0.33	0.39	0.96	0.87	0.34	1.00	0.03
MC	0.50	0.30	0.09	0.23	0.37	0.03	1.00

Table 24: Correlation Matrix for the EWMA Benchmark Measures - ES 97.5%

Models' Sensitivity

Figure 25: HS VaR 99% Sensitivity Analysis to the Number of Days in the Rolling Window

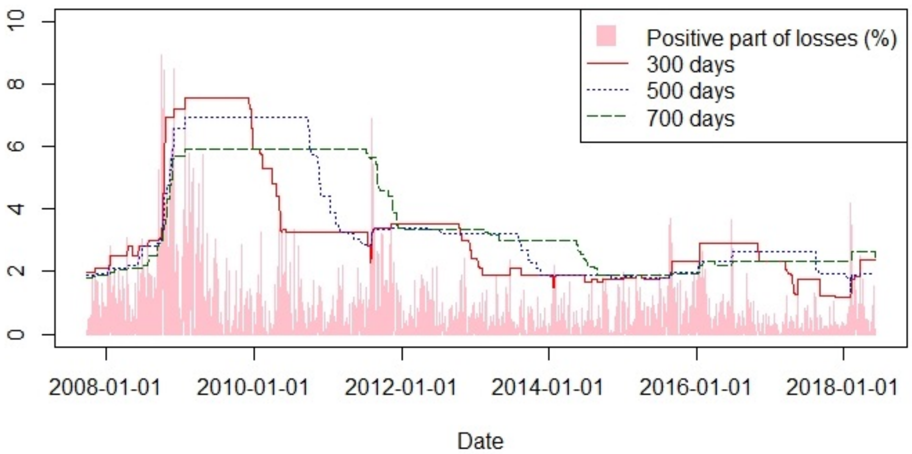


Figure 26: EWMA VaR 99% Sensitivity Analysis to α in the volatility scheme

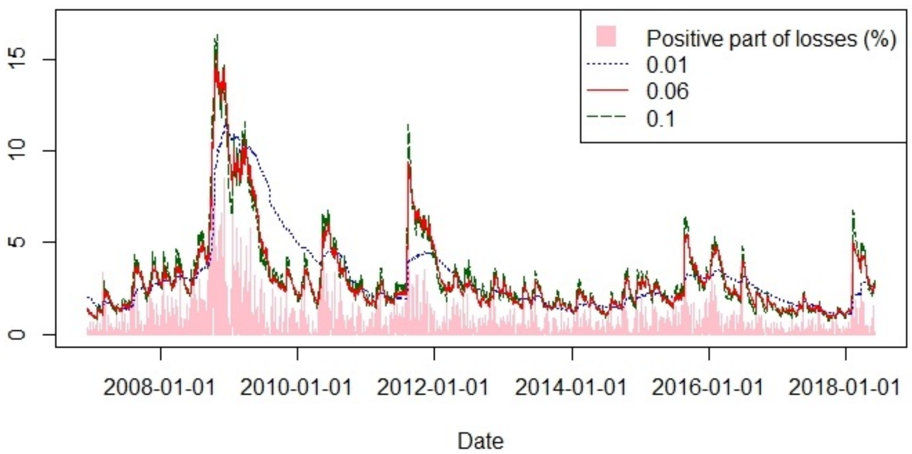


Figure 27: EVT VaR 99% Sensitivity Analysis to the Value of the Threshold

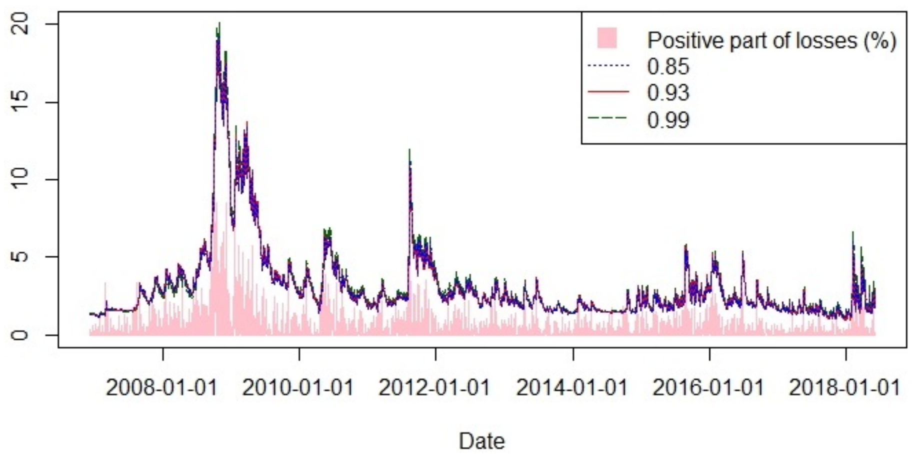


Figure 28: EVT ES 97.5% Sensitivity Analysis to the Value of the Threshold

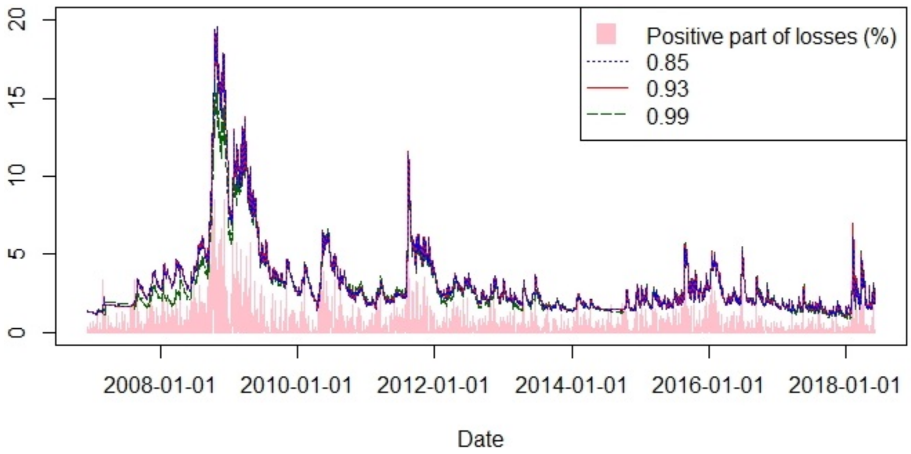


Figure 29: MC VaR 99% Sensitivity Analysis to the Number of Simulations

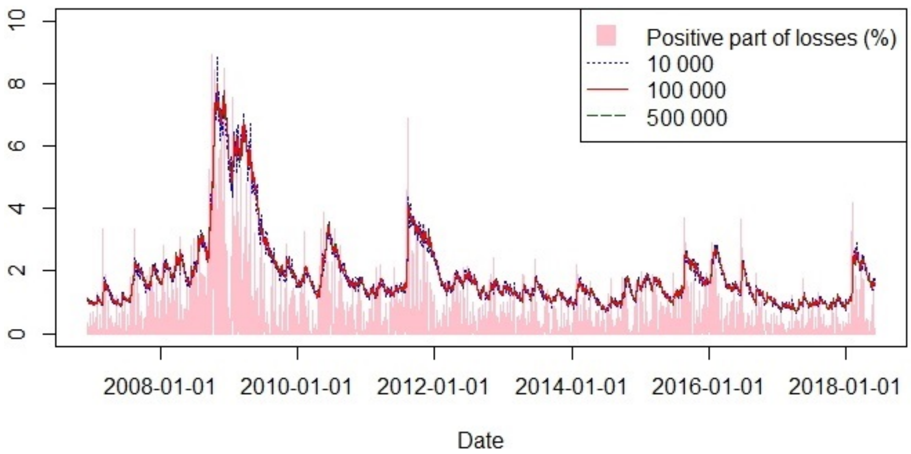
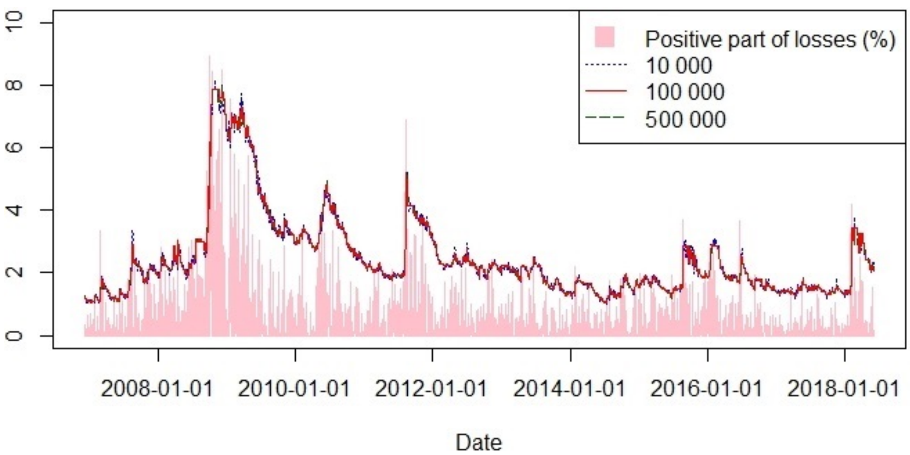


Figure 30: MC ES 97.5% Sensitivity Analysis to the Number of Simulations



References

[1] C. Alexander. *Volatility and Correlation: Measurement, Models and Applications.* University of Sussex UK, Wileys, 1998.

[2] C. Alexander. *Moving Average Models for Volatility and Correlation, and Covariance Matrices.* Wiley Press, 2008.

[3] C. Alexander and J.M. Sarabia. *Quantile Uncertainty and Value-at-Risk Model Risk.* Risk Analysis, Vol. 32 No. 8, 2012.

[4] P. Artzner, F. Delbaen, J.-M. Eber and D. Heath. *Coherent measures of risk.* Mathematical Finance, Vol. 9 No. 3, 1999.

[5] Basel Committee on Banking Supervision. *An Internal Model-Based Approach to Market Risk Capital Requirements,* 1995.

[6] Basel Committee on Banking Supervision. *Amendment to the 1988 Basel Capital Accord,* 1996.

[7] Basel Committee on Banking Supervision. *International Convergence of Capital Measurement and Capital Standards,* 2004.

[8] Basel Committee on Banking Supervision. *Fundamental review of the trading book: A revised market risk framework,* 2013.

[9] N. Bingham. *Bayesian Methods,* lecture notes, Statistical Methods for Finance M5MF38 Imperial College London, December 2017.

[10] R. Black (Chair), M. Beck, R. Green, J. Grewal, Z. Lim, I. Maclugash, N. Morjaria, A. Smith, A. Tsanakas and L. Witts. *Model risk - illuminating the black box.* Institute and Faculty of Actuaries, 2017.

[11] D. Brigo. *Interest Rate Models with Credit Risk, Collateral, Funding Liquidity Risk and Multiple Curves*, lecture notes, Interest Rate Models M5MF30 Imperial College London, January 2018.

[12] C. Chang, J.-A. Jimenez-Martin, M. McAleer and T. Perez-Amaral. *Risk Management of Risk under the Basel Accord: Forecasting Value-at-Risk of VIX Futures,* 2011.

[13] R. Count. *Model Uncertainty and its impact on the pricing of derivative instruments.* Mathematical Finance, Vol. 16, No. 3, 2006.

[14] J. Danielsson, K. James, M. Valenzuela and I. Zer. *Model Risk of Risk Models*, 2016.

[15] Deloitte US. *Model Risk Management Plaquette.* www2.deloitte.com. Deloitte Risk Advisory, 2017.

[16] D. Eliezer, S. Dash. *How-to Guide to Quantifying Model Risk.* www.numerix.com. Numerix, 2014.

[17] Financial Services Authority. *Enhancing frameworks in the standardised approach to operational risk,* Guidance Consultation, 2010.

[18] P. Glasserman. *Monte Carlo Methods in Financial Engineering.* Springer, 2003.

[19] P. Glasserman, P. Heidelberger and P. Shahabuddun. *Efficient Monte Carlo methods for value-at-risk.* Columbia University, 2000.

[20] M. Grootenboer, C. Maclver, J. McIndoe and C. Patel. *Stylised facts, GARCH Modelling, and Risk Measures.* Quantitative Risk Management - Assignment 1, 2017.

[21] O. Gueant. *Computing the Value at Risk of a Portfolio: Academic literature and Practionners' response.*

[22] J. Kerkhof, B. Melenberg and H. Schumacher. *Model Risk and Regulatory Capital.* Journal of Financial Stability, Forthcoming, 2006.

[23] H. Li, X. Fan, Y. Li, Y. Zhou, Z. Jin and Z. Liu. *Approaches to VaR.* Standford University.

[24] C. MacIver, C. Patel, H. Sorsky and M. Grootenboer. *Extreme value theory and Copula-based risk measures using Monte Carlo.* Quantitative Risk manegement - Assigment 2, 2017.

[25] M. Pakkanen. *Part 1: Risk management and stylised facts*, lecture notes, Quantitative Risk Management M5MF10 Imperial College London, October 2017.

[26] M. Pakkanen. *Part 2: Basic concepts of risk management*, lecture notes, Quantitative Risk Management M5MF10 Imperial College London, October 2017.

[27] M. Pakkanen. *Part 3: Univariate time series and volatility modelling*, lecture notes, Quantitative Risk Management M5MF10 Imperial College London, November 2017.

[28] M. Pakkanen. *Part 4: Heavy-tailed distributions and extreme value theory*, lecture notes, Quantitative Risk Management M5MF10 Imperial College London, November 2017.

[29] G. Piroozfar. *Forecasting Value at Risk with Historical and Filtered Historical Simulation Methods.* Uppsala University, 2009.

[30] P. Rappoport. *A new Approach: Average Shortfall.* J.P. Morgan Fixed Income Research Technical Document, 1993.

[31] H. Scheule and D. Rosch. *Model Risk: Identification, Measurement and Management.* Risk Books, 2010.

[32] P. Sibbertsen, G Stahl and C. Luedtke. *Measuring Model Risk.* The Journal of Risk Model Validation, Vol. 2, No. 4, 2008/09.

[33] SR 11-7 - Supervisory Guidance on Model Risk Management, Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, 2011.

[34] M. Wladkowski. *An Assessment of Value at Risk Measurements Using Bayesian Model Averaging.* Northcentral University, 2012.