BANKING REGULATION:
A BAYESIAN NETWORK APPROACH TO RISK MANAGEMENT

Research submitted by

EDEN GROSS

in fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Department of Finance and Tax

University of Cape Town

September 2024

Supervisors:

Associate Professor Ryan Kruger

Associate Professor Francois Toerien

Abstract

The ever-evolving regulation surrounding banks and market risk, coupled with increased computing power, make for favourable conditions in employing machine learning techniques to estimate and forecast market risk metrics such as value at risk (VaR) and expected shortfall (ES). This study consists of three sections. First, this study comprehensively examines the performance of various market risk models when producing VaR and ES, and their stressed counterparts, using Standard and Poor's (S&P) 500 index returns from 1991 to 2020. The initial results show that autoregressive models are the most accurate of the traditional market risk models. Second, the first section's results are then used as the basis against which a novel and comprehensive Bayesian network (BN) methodology for producing VaR and ES forecasts, and those of their stressed counterparts, is assessed in the context of banking regulations, using four learning algorithms. The forecasts generated by the BNs are not found to offer any improved accuracy when incorporated into the market risk metric calculations, primarily due to the limited weight of the forecast in the return distribution relative to the historical returns in the return probability density function. Finally, a novel integrated forecast dynamic Bayesian network (IFDBN) methodology is developed, whereby, for each metric, the best-in-class autoregressive model and the best-in-class BN learning algorithm are coupled to produce market risk forecasts. The results of the IFDBNs are mixed, with the stressed ES metric IFDBN being the only IFDBN to produce more accurate forecasts relative to its traditional autoregressive counterpart. While certain market risk metrics may benefit from using IFDBNs in the forecasting process, this result is not universal, and the risk practitioner must evaluate the usefulness of IFDBNs on a case-by-case basis.

UNIVERSITY OF CAPE TOWN
FACULTY OF COMMERCE
Igniting Knowledge and Opportunity

Declaration

COMPULSORY DECLARATION:

1. This dissertation has been submitted to Turnitin (or equivalent similarity and originality checking software) and I confirm that my supervisor has seen my report and any concerns revealed by such have been resolved with my supervisor.
2. I certify that I have received Ethics approval (if applicable) from the Commerce Ethics Committee.
3. This work has not been previously submitted in whole, or in part, for the award of any degree in this or any other university. It is my own work. Each significant contribution to, and quotation in, this dissertation from the work, or works of other people has been attributed, and has been cited and referenced.

Student number	GRSEDE001
Student name	Eden Gross
Signature of Student	Signed by candidate
Date:	11 September 2024

Acknowledgements

התזה הזו מוקדשת לסבא וסבתא שלי, שתמיד האמינו ותמכו בי.

This thesis is the product of the support of so many people. First and foremost, my loving partner and family. I could not have done this without your continued, unconditional, and unwavering love and support.

Second, I'd like to thank my supervisors, Associate Professor Ryan Kruger and Associate Professor Francois Toerien. This was a long journey, with many surprises along the way. This line does not do justice to your contributions, academically, professionally, and, most importantly, personally.

Third, I'd like to thank Ryan, again, for making sure that I have the resources to see this thesis to its completion. I am eternally grateful.

Last, to the Department of Finance and Tax, for allowing me to undertake this research.

Table of Contents

List of Figures

List of Tables

1. Introduction

Financial risk management is neither a new topic nor a fast-evolving one. Many approaches to financial risk management, especially within the regulated banking sector, have been around for decades. However, these well-established approaches have a poor record when it comes to avoiding the collapse of banks, especially during financial crises, as evidenced by the 2008 global financial crisis, the collapse of Lehmann Brothers, and the more recent series of bank failures that plagued the United States (US) and elsewhere.

Despite the increasing complexity of models offered by the literature, most banks still use simplistic models, such as the historical simulation model (Pérignon & Smith, 2010) to measure market risk[1]. Drawbacks of these models include the slow updating to increased market volatility (Daníelsson, 2002) and the underestimation of risk[2]. However, even other, more sophisticated and complex models may still suffer from a fundamental weakness: They fail to incorporate any forward-looking information.

In recent years, various machine learning models have become a major feature in both the academic and the practitioner worlds. This recent increase in the uses and applications of machine learning models highlights the level of computing power now available and the ease of use of these models, even when the user is not necessarily familiar with the underlying mathematics and statistics that underpin their workings. Thus, there is a clear opportunity to build on the existing (and, likely, outdated) practices of financial risk management and banking regulation by introducing machine learning models and techniques to quantify banks' market risk and, most importantly, to incorporate forward-looking predictions.

A Bayesian network (BN) is a machine learning application of Bayesian inference in the form of a graphical representation of probabilistic computations and processes. BNs, and Bayesian inference, make use of statistics and relationships between variables to model conditional dependencies and, hence, causation graphically. More technically, a BN is a directed acyclic graph, i.e., a graph whose edges are all pointing in specific directions, which

[1] Market risk is the risk that losses are incurred in a portfolio due to adverse movements in the market prices of investments held in the portfolio (McNeil & Frey, 2000).

[2] This is true for the unscaled forecasts, i.e., one-day forecasts, as opposed to 10-day forecasts. See, for example, Berkowitz and O'Brien (2002), Pérignon, Deng, and Wang (2008), Berkowitz, Christoffersen, and Pelletier (2009), Pérignon and Smith (2010), and O'Brien and Szerszeń (2017), among others.

does not have any cycles in it, i.e., it is not possible to return to any node when one travels along the graph's edges (Stephenson, 2000).

Common applications of BNs are found in the fields of medicine, weather prediction, and speech recognition. In medicine, BNs are often used in the process of diagnostic reasoning. Once a patient is diagnosed with a disease, BNs are applied to the various possible causes of the disease to test the constructed hypotheses, and are further employed to avoid misdiagnoses, due to the obvious underlying uncertainty involved (Lucas, van der Gaag, & Ameen, 2004). In weather prediction, prediction models often need to be updated continually to accurately reflect the data obtained from weather balloons. BNs are then employed to update the statistical view of the weather, given the new information that is fed into the system from said balloons (Cano, Sordo, & Gutiérrez, 2004).

Asset pricing relies on the efficiency of financial markets, which is defined as the extent to which the relevant information available in the market is reflected in the price of the asset in question. The process of asset pricing requires an asset pricing model (the arbitrage pricing theory model, or APT, for example) to be applied so that a relative value of an asset can be determined. The continual adjustment of asset prices due to the availability of new information (or increased efficiency of the market) results in profits or losses in asset values which, in turn, are translated to returns.

As mentioned, banks operate within a regulatory framework, primarily due to their perceived systemic importance in national and global financial systems. While this framework is applied on a national level, it is often in line with the supranational regulatory framework of the Basel Committee on Banking Supervision (BCBS), known as the Basel Accords. These, collectively, make a set of guidelines for banks to follow when quantifying and managing their financial risks, namely market, credit, and operational risks.

The BCBS has introduced several sets of regulations designed to regulate banks. With its first publication in 1988, the BCBS has introduced regulations that govern credit risk with its Basel Accord (Basel Committee on Banking Supervision, 1988). In 1996, the quantification and monitoring of market risk was introduced with the introduction of an amendment to the Basel Accord (Basel Committee on Banking Supervision, 1996). The Basel II Accord, introduced in 2006, incorporated operational risk into the risk framework applicable to banks (Basel Committee on Banking Supervision, 2006). The introduction of the 1996 amendment to the Basel Accord also moved the banks away from using a prescribed standardised formula to

the possible use of internal models (Basel Committee on Banking Supervision, 1996), which facilitated banks' use of models such as the historical simulation model.

The daily and intraday profit and loss figures captured for each of a bank's trading desks (i.e., the different asset classes in which a bank invests and trades) contribute to the bank's overall profit and loss account. This account is subject to the supervision of national regulators, as often governed by the guidelines set out by the BCBS, ensuring that excess losses can be covered by the bank through the means of capital reserves. The continual adjustment of asset prices and, hence, the profit and loss account, takes place via the incorporation of new information relating to the assets (and, by extension, the asset classes) as soon as this new information is disseminated to the market. BN framework models are well suited for modelling this exact process.

A common risk management metric employed by banks to measure market risk (and enforced by regulatory guidelines such as those proposed by the BCBS) is that of value at risk (VaR). VaR is a portfolio risk measure that stipulates the maximum expected loss over some target horizon within a pre-determined confidence interval (Shenoy & Shenoy, 2000). While a more mathematical definition is introduced in Chapter 2, VaR is a tail risk metric, calculated using the profit and loss probability density function (PDF). Applying a BN application to market risk produces this PDF, enabling the calculation of tail risk metrics such as VaR.

For banks within the US, the US Securities and Exchange Commission (SEC) expanded existing disclosure requirements of US banks in 1997. The SEC requires banks to publish material information on their holdings in market risk-sensitive instruments, using both quantitative and qualitative supplements. The SEC requires US banks to publish VaR forecasts and specifically sets out the disclosure of assumptions and accompanying qualitative information to accompany such VaR forecasts (United States' Securities and Exchange Commission, 1997).

As a result, US banks report their VaR forecasts, and this disclosure process accomplishes several goals. From a public information standpoint, VaR is a simple forecast to quote, capturing a probability, a time horizon, and a loss figure. This makes it an easy forecast for the public to both understand and use for comparative purposes (to, for example, compare banks' risk levels). From a regulatory standpoint, VaR forecasts are the key input in determining how much capital a bank must hold in reserves, while also serving as input for evaluating the

performance of the internal model used by banks via a process known as backtesting. These concepts are explored further in Chapter 3.

In recent publications, the BCBS decided to move from the use of the VaR as a risk management metric to the use of the expected shortfall[3] (ES) as a risk management metric (Basel Committee on Banking Supervision, 2013). Taking risk management a step further, the ES of a portfolio, put simply, is a conditional measure of loss outputting a loss figure which exceeds the VaR loss forecast (Yamai & Yoshiba, 2005), i.e., should a portfolio exhibit losses that exceed the VaR loss predicted (with a certain degree of certainty), its losses will equal the ES forecast. Hence, ES, too, is a tail risk metric, and the PDF output of a BN allows for this metric to be calculated. Although ES is not easily backtested[4], approximate backtesting techniques are discussed further in Chapter 3.

In the context of the regulatory requirement for US banks to report their VaR forecasts, and considering the change introduced by the BCBS in its Basel III framework to now require banks to report ES forecasts as well as the stressed versions of VaR and ES, this study examines four market risk metrics. Each market risk metric is backtested using its applicable techniques, as identified in the literature. These backtesting techniques are discussed further in Chapter 3.

The use of BNs to calculate market risk was proposed by early studies, such as those of Shenoy and Shenoy (2000) and Demirer, Mau, and Shenoy, (2006). More recently, Apps (2020) developed a simplified BN methodology to predict the directional move of a portfolio's returns and its impacts on VaR. BNs can be used to learn the causal relationships between various variables that influence a trading desk's returns. These relationships can then be used to produce forward-looking return predictions and, in turn, use these forward-looking predictions to produce more accurate tail risk metric forecasts, specifically in the context of market risk.

Hence, the use and incorporation of BNs in producing market risk forecasts may overcome some of the challenges experienced in quantifying and managing market risk. First, due to the causal relationships learned within the network and their frequent updating, the BN may result

[3] Expected shortfall is also commonly known as expected tail loss (ETL), conditional value at risk (CVaR), and average value at risk (AVaR).

[4] ES is not elicitable, meaning that a strictly consistent scoring function does not exist for it (Gneiting, 2011). This means that ES does not lend itself to backtesting directly.

in market risk metric forecasts that are more sensitive to changes in market volatility, thereby overcoming the slow adaptation displayed by currently used models such as the historical simulation model (Pérignon & Smith, 2010). Second, this frequent updating using modelled causal relationships may lead to market risk metric forecasts that are more in line with market movements, thereby reducing the underestimation of risk when using one-day forecasts. Third, the use of BNs will bring financial risk management techniques to the modern age, by initiating an exploration of forward-looking forecasting techniques beyond the tools available in the toolbox of the traditional risk manager at a bank.

1.1. Study Objectives and Contribution

This study's objectives are split according to its three empirical chapters.

- In the first empirical chapter, this study aims to establish how accurate and appropriate the traditional market risk models are within the existing regulatory framework put forth by the BCBS. Moreover, the first empirical chapter also aims to establish whether the use of ES over VaR has a significant impact on model appropriateness and accuracy.
- In this study's second empirical chapter, the aim is to introduce BNs to market risk modelling, and to determine whether these machine learning models can improve on the existing implementations, as proposed by early studies theorising on the matter (see, for example, Shenoy and Shenoy, 2000).
- Finally, the objective of this study's third empirical chapter is to introduce a novel hybrid BN methodology to market risk management and to assess whether it is able to improve on the existing implementations of the traditional market risk models.

Using the Standard and Poor's (S&P) 500 index as a test case, this study provides a comprehensive comparison of the accuracy of traditional market risk models in producing forecasts for VaR, ES, and their stressed counterparts (see Chapter 2 for more detail). Market risk forecasts are produced using two statistical distributions to enhance the comparison. Using several backtesting techniques, forecasting error measures, and statistical tests, this study determines the traditional model that produces the most accurate forecasts for each of the four market risk metrics considered. These best-in-class models are then used as the basis against which the corresponding BN's performance is evaluated.

This study then builds on the theoretical basis presented by Shenoy and Shenoy (2000) and Demirer, et al., (2006), and the recent simplified methodological BN attempt by Apps (2020) and develops a comprehensive BN methodology to model the causal relationships between

financial and economic variables (e.g., the consumer price index, commodity prices, and foreign exchange rates) to produce one-day-ahead forecasts of a market index using a rolling period methodology and several BN learning algorithms (for more detail, see Chapter 4). This implementation is performed using R. This mimics the performance of a US bank's equities trading desk, and enables the calculation of the four market risk metrics for each day of the out-of-sample period. These market risk forecasts are evaluated using the same backtesting techniques, forecasting error measures, and statistical tests mentioned above. The best-in-class models are compared to their traditional counterparts.

Finally, this study introduces a novel methodology combining both BNs and the autoregressive properties of some traditional market risk models. For this methodology, the term 'integrated forecast dynamic Bayesian networks' (IFDBNs) is coined. The results of the IFDBNs for the various market risk metrics are then compared to the results of both traditional models and BNs to assess the accuracy offered when producing market risk metric forecasts.

This study provides several contributions to the literature. First, it offers a comprehensive review of the performances of various models to produce market risk metrics in the context of a US bank's market risk, as related to its equities trading desk. This contribution is especially important when considering the stressed metrics, as they have very little coverage in the literature, especially with respect to differences relating to the use of different underlying statistical distributions. In addition, the manner in which stressed periods are calibrated in this study is considered to be the most conservative approach, and this, too, is a contribution to the literature for the stressed metrics.

Next, this study contributes to the literature by providing a comprehensive BN framework and methodology for asset pricing, with a specific focus on financial risk management. This study also expands on existing studies (Shenoy & Shenoy, 2000; Demirer, et al., 2006; Apps, 2020) by producing VaR forecasts using a variety of BN learning algorithms. Moreover, this study takes this a step further by producing forecasts for three additional market risk metrics.

Finally, and most importantly, this study develops a novel methodology for BNs, termed IFDBNs, combining the weight-assigning properties of autoregressive models and the forward-looking prediction properties of BNs. This method can be used in other applications, combining time series data (as is the case for dynamic BNs) and predictions in a way that places

more emphasis on the prediction and more recent observations. This methodology is not limited to financial risk management.

1.2. Structure Outline

The remainder of this study is structured as follows: Chapter 2 introduces the Basel Accords and provides a brief overview of the regulatory framework applicable to US banks' market risk. Chapter 3, the first empirical chapter of this study, calculates VaR and ES forecasts for the study period using traditional techniques. The data, methodology, and results are presented, followed by the chapter's conclusions. Chapter 4, the second empirical chapter of this study, develops a BN methodology for market risk management of a US bank's equities trading desk and constructs such BNs. The discussions surrounding such construction, together with the data, methodology, and results, are presented, followed by the chapter's conclusions. Chapter 5 develops the novel methodology of IFDBNs, and analyses their performances when producing market risk metric forecasts relative to the results of the preceding two chapters. Chapter 6 concludes this study, while Chapter 7 discusses the limitations of the research discussed herein and concludes with suggestions for future research.

2. The Basel Accords and Banks' Market Risk Management Framework

This chapter introduces the Basel Committee on Banking Supervision (BCBS) and the Basel Accords. This risk management framework is discussed, with reference to key aspects such as the measurement and quantification of risk, the use of internal models, and the standardised formula. Moreover, the framework provided to assess the predictive accuracies and performances of internal models is discussed to present a basis for the evaluation of the various value at risk (VaR) models and expected shortfall (ES) models considered in this study.

In essence, VaR is a measure that captures the profit and loss account of an institution. That profit and loss figure, however, is a random variable. Hence, consider a $100\alpha\%$ h-day VaR forecast at time t, a number $x_{ht,\alpha}$ of the random variable X representing the α quantile of the profit and loss account must be found such that

$$Pr\left[X_{ht} < x_{ht,\alpha}\right] < \alpha \tag{1}$$

This, then, results in a VaR at time t being

$$VaR_{ht,\alpha} = -x_{ht,\alpha} \times P_t \tag{2}$$

where P_t represents the total portfolio value at time t (Alexander, 2008). VaR and its most commonly used models are further discussed in Chapter 3. ES, on the other hand, is essentially the conditional expected loss, conditional on the loss incurred exceeding the VaR forecast. Both a formal definition and a discussion surrounding ES are presented in Chapter 3.

The BCBS is a committee housed in the Bank of International Settlements, located in Basel, Switzerland. The committee oversees banking rules and regulations, which are, in turn, optionally enforced by national banking regulators who are members of the Bank of International Settlements. Amongst other aspects, the committee requires member countries' banks to maintain adequate capital (in the form of reserves) to ensure that relevant financial stakeholders are not placed at risk.

As mentioned in Chapter 1, the BCBS has introduced several sets of regulations (known as Accords) and amendments to such regulations. The Accords started by only covering credit risk (in the original Basel Accord, see Basel Committee on Banking Supervision, 1988), to now including credit, market, and operational risks (see the 1996 amendment to the Basel Accord, Basel Committee on Banking Supervision, 1996; Basel II, Basel Committee on Banking Supervision, 2006).

Under the Basel II Accord, the notion of 'adequate capital' is often calculated by making use of a bank's internal VaR model (Basel Committee on Banking Supervision, 2004), although banks have the choice between the standardised formula, as dictated by the BCBS, or an internal model, which has to be 'proved' to be adequate by means of backtesting[5], as stipulated in the Market Risk Amendment issued by the BCBS in 1996 (Basel Committee on Banking Supervision, 2004). Banks are required to calculate 99% 10-day VaR forecasts at both the trading desk level and firm-wide level and report these daily, at the beginning of each trading day. This is achieved through an approximation, whereby a bank calculates its 99% one-day VaR forecast and scales this forecast by the square-root-of-time as a scaling factor[6], i.e., $\sqrt{10}$. The BCBS further stipulates that the sample period used to calculate VaR must be of a minimum length of one year.

The BCBS leaves the use and calibration of internal models in the hands of the banks and avoids prescribing how any internal model is to be calibrated to produce market risk metrics. The BCBS leaves the issue of suitability of the internal model to the process of backtesting – a process which, itself, is not subject to much prescription.

When it comes to calibrating internal models, some of which are discussed in Chapter 3, the lack of prescription surrounding the methodology of calibration of such models leaves room for ambiguity and for a bank to adjust a model's outputs until the desired outcome is achieved. For example, when calibrating any model whose output is dependent on fitting a distribution to the historical returns achieved by a bank (or a bank's trading desk), the choice of distribution is left to be decided by the bank. Hence, should a distribution chosen at first yield 'unfavourable' outcomes (e.g., excessive breaches, as determined by the bank), an adjusted distribution may be used in the hopes of achieving a 'better' number of breaches (or whatever else the bank would like to target). In the same vein, for any model that relies on the calibration of model parameters (as is the case for the autoregressive models considered in Section 3.1.3),

[5] Backtesting is the process of comparing a sample of VaR forecasts as predicted by the model for a historical period to the actual profit and loss figures of that period (Pérignon, Deng, & Wang, 2008).

[6] The scaling of VaR forecasts using the square-root-of-time rule produces VaR estimates which are misleading due to the incorrect (implicit) assumptions that the underlying returns are normally distributed, and that the volatility of such returns is homoscedastic in nature (Daníelsson, Embrechts, Goodhart, Keating, Muennich, Renault, & Shin, 2001). In fact, Daníelsson, Hartmann, and de Vries (1998) note that this scaling rule is only appropriate since any other scaling rule is just as arbitrary in nature.

the decision of how often to update these parameters, if at all, is also left to the bank's judgement. As a last step of abstraction when it comes to the application of models to calculate market risk metrics, the models themselves are, too, left to the bank's judgement. Finally, the backtesting techniques tasked with 'validating' a bank's internal model, the most prominent of which are discussed in Section 3.3.3, are, too, left to the bank's judgement, apart from the BCBS's traffic light test (see Section 3.3.3.1). Hence, a bank may simply use whichever combination of statistical distribution, parameter updating frequency, model, and backtest that produces the most favourable results. This study has made several assumptions, as explored in detail in Chapter 3, which are in line with these issues. These assumptions were made to provide the most accurate results, as opposed to those that could have been made to produce the most 'favourable' results.

The 99% confidence level has been criticised by Daníelsson, et al., (2001) in their response to a request for comments by the BCBS preceding the release of the Basel II Accord. The authors note that the regulatory framework introduced by the BCBS intends to protect banks (and the economy as a whole) against the risk of bank failure and the systemic risk which may be the result of such failure. At the 99% level, a bank would be expected to experience such failure at a frequency of 2.5 times a year – a frequency that is too frequent for the nature of events that the regulatory framework aims to prevent (Daníelsson, et al., 2001). Hence, the BCBS's framework may overlook regulatory risk, which is introduced due to the framework's use, or the consequences of such regulatory risk, while reducing systemic risk.

The standardised method requires the bank to hold capital equal to a product of risk weights and market values of notional portfolios, together with some consideration of the correlation between such notional portfolios (Basel Committee on Banking Supervision, 2013). Banks are required to calculate the standardised method's capital reserves in the case that the calculations obtained using the internal model are rejected (Basel Committee on Banking Supervision, 2013).

The standardised method has been developed as an easily implementable capital reserving method. It does, however, have its shortcomings in adequately capturing the intricacies of the various assets held by each specific bank and the diversification benefits enjoyed between trading desks (to their full extent). This may imply that a bank employing the standardised model may hold excess capital on hand, which results in high opportunity costs – costs that are high enough to result in banks opting to implement their own internal models.

Table 1: The Basel Committee on Banking Supervision's Traffic Light Test Breaches Penalty for the Basel II Accord

Zone	**Number of Breaches**	***k***
Green Zone	< 4	**0.00**
	5	**0.40**
	6	**0.50**
Yellow Zone	7	**0.65**
	8	**0.75**
	9	**0.85**
Red Zone	10 <	**1.00**

Note: This table depicts the classification of breaches of internal value at risk (VaR) models employed by banks according to the Basel II Accord. A breach is recorded when the actual loss figure for a specific day exceeds the VaR forecast produced by the bank's internal model. Depending on the number of breaches incurred by the model, a penalty variable k is determined. The number of breaches is tested over a period of at least 250 trading days.

If a bank chooses to implement an internal VaR model, that VaR model is tested for breaches (often referred to as violations, exceptions, or exceedances) over an out-of-sample period, often being a 250-trading-day period. A breach takes place when the loss incurred during any one trading day exceeds the VaR forecast for that day (Jiménez-Martín, McAleer, &. Pérez-Amaral, 2009). Basel II imposes a penalty on models that experience more breaches, measured by the penalty variable, k. A value between 0 and 1 (inclusive) is assigned to k based on the number of breaches experienced by the model. Breaches are further separated into 'zones' depending on the number of breaches experienced. The 'Green zone' captures any model that experiences between 0 and 4 breaches, the 'Yellow zone' captures any model that experiences between 5 and 9 breaches, while the 'Red zone' captures any model that experiences 10 or more breaches. The zone penalty system is depicted in Table 1, above.

The Basel II Accord then requires banks to calculate their minimum daily capital charges (DCC) as follows.

$$\min_{\{k,VaR\}} DCC_t = \max\{-(3+k)\,\overline{VaR}_{60}, -VaR_{t-1}\} \tag{3}$$

where VaR_t is the value at risk forecast at time t, calculated as per Equation (2); $\overline{VaR}_{60}$ is the average VaR for the past 60 trading days; and $k \in [0,1]$ is a penalty variable determined as detailed in Table 1. This formula states that the regulatory VaR on a given day is to be the higher of the previous day's VaR forecast and the product of the previous 60 days' VaR forecasts multiplied by a factor consisting of 3 and the penalty variable k.

If a bank's penalty variable is classified to be in the 'Red zone', this may lead a regulator to require the bank to adopt the standardised method to calculate VaR as stipulated in the Basel II

Accord (McAleer & da Veiga, 2008). This will, most likely, lead the bank to hold more capital in the form of reserves than needed, thereby leading to reduced profitability and a damaged reputation. The enforcement of the standardised method may be necessary as the large scaling, as captured by the penalty variable (acting as a scaling factor when added to the constant three[7]) in Equation (3), leads to the conclusion that a failure to cover a bank's VaR using its capital reserves should only occur for portfolios with extreme risks (Jackson, Maude, & Perraudin, 1997).

However, a criticism of the penalty variable k is that a bank may choose a value of k that works to its advantage. Begley, Purnanandam, and Zheng (2017) find that banks take advantage of the discontinuous nature of k (i.e., its progression in value based on the increasing number of breaches) to balance the added opportunity cost of holding additional capital versus the reputational risk associated with holding less capital. Hence, banks have the opportunity to use the regulatory framework to their advantage and to hold less capital than otherwise would be possible. Another criticism of the penalty variable is that it does not account for the magnitude of breaches, either individually or collectively (Jiménez-Martín, et al., 2009). In fact, Jiménez-Matín, et al., (2009) encourage banks to experience breaches, as long as these banks avoid the 'Red zone', to optimise their opportunity costs, and to have large breaches due to the lack of a limit on the size of each breach.

In response to the 2008 global financial crisis, the BCBS has complemented the VaR forecast with a stressed VaR (SVaR) forecast, being a one-year VaR calibrated under stressed conditions[8] (Basel Committee on Banking Supervision, 2009). Since SVaR is equal to or greater than VaR (Liu & Stentoft, 2021), the capital held by banks has at least doubled following the introduction of SVaR and, moreover, has stabilised given that the period used to calibrate SVaR changes less frequently than that used to calculate VaR.

The lack of prescription offered by the BCBS allows for flexibility when it comes to calculating the SVaR (or any other stressed metric). The BCBS states that the stress period must be at least one year in length, and the historical period used to calibrate the stress period must include 2007 (Basel Committee on Banking Supervision, 2023). However, the BCBS is

[7] The constant added, i.e., the number three, accounts for model error (Stahl, 1997).

[8] The stressed period is taken to be the bank's most severe one-year period of losses available (Liu & Stentoft, 2021). The SVaR forecast quoted is still a 10-day 99% figure.

not clear as to how this period is to be determined. This ambiguity offers an opportunity for banks and regulators to apply judgement – a judgement which may or may not be in line with the result intended by the BCBS. For example, a bank may choose to use a one-year period which includes the worst return observed in its historical period. This period may or may not be the worst stressed period in totality, as the other days included in the period may have been better in aggregate relative to another historical period which does not include that worst return observed, i.e., there may exist a period better suited as a stress period which does not include the particular day on which the worst return was observed. Moreover, it is possible for the stressed period to not consist of consecutive days at all. In the strictest and most conservative definition of a stressed period, a bank may amalgamate a collated set of days on which the worst returns were achieved in its historical period into a collated stressed period, thereby yielding the absolute worst calibration set. This, in fact, is the method adopted in this study to offer the most robust and conservative set of results.

The Basel III Accord was then introduced as an amendment to the Basel II Accord after the 2008 global financial crisis and stipulated a tighter definition of the term 'capital' and the requirements of such during stressed conditions in response to the failures of its predecessor (Basel Committee on Banking Supervision, 2017). This Accord moved banks from the use of VaR as the only market risk quantifying tool to both VaR and (liquidity-adjusted) ES following the BCBS's Fundamental Review of the Trading Book (Basel Committee on Banking Supervision, 2022). Moreover, the SVaR measure was to be complemented by the stressed ES (SES) measure[9], and SES can be calibrated over a reduced set of risk factors if a sufficiently long historical period is unavailable (Liu & Stentoft, 2021). The Basel III Accord still requires the calculation of a one-day 99% VaR forecast, but now also requires calculating a one-day 97.5% ES forecast to capture the bank's tail risk (Basel Committee on Banking Supervision, 2022), although this forecast cannot be scaled to reflect a 10-day rule[10] (Liu & Stentoft, 2021). Both forecasts are to be reported on a trading desk and firm-wide level.

Considering the liquidity issues experienced by banks during the 2008 global financial crisis, the move to use ES rather than VaR was accompanied by classifying instruments by

[9] The ES used in the Basel III Accord is the liquidity-adjusted SES, hereafter referred to simply as 'SES'.

[10] The BCBS facilitates the calculation of 10-day ES forecasts using overlapping forecasting periods (Liu & Stentoft, 2021).

their liquidity horizons. The instruments' horizons are grouped into buckets of length 10 (the base horizon), 20, 40, 60, and 120 days, whereby liquidity horizon LH_j, for $j = 1, 2, 3, 4,$ or 5, captures all instruments with a liquidity horizon of 10, 20, 40, 60, or 120 days or more, respectively. ES is then calculated as follows.

$$ES = \sqrt{[ES(1)]^2 + \sum_{j \geq 2} \left(ES(j) \times \sqrt{\frac{LH_j - LH_{j-1}}{10}} \right)^2} \tag{4}$$

where $ES(j)$ is the expected shortfall for a portfolio with instruments of the respective liquidity horizon and $\sqrt{(LH_j - LH_{j-1})/10}$ is a liquidity-based scaling factor (Basel Committee on Banking Supervision, 2019b).

The introduction of ES as a substitute and as the key measure of the Basel II Accord's framework has also led the BCBS to revise its penalty variable, k, presented in Table 1. The new penalty variable, which will be still labelled as k in this study, is now as presented in Table 2, below.

Table 2: The Basel Committee on Banking Supervision's Traffic Light Test Breaches Penalty for the Basel III Accord

Zone	**Number of Breaches**	***k***
Green Zone	< 4	**0.00**
	5	**0.20**
	6	**0.26**
Yellow Zone	7	**0.33**
	8	**0.38**
	9	**0.42**
Red Zone	10 <	**0.50**

Note: This table depicts the classification of breaches of internal stressed expected shortfall (SES) models as employed by banks according to the Basel III Accord. A breach is recorded when the actual loss figure for a specific day exceeds the SES forecast produced by the bank's internal model. Depending on the number of breaches incurred by the model, a penalty variable k is determined. The number of breaches is tested over a 250-trading-day period.

The Basel III Accord requires banks to calculate their internally modelled capital requirements with no risk constraints (IMCC) for the entire bank for calibrated period t is as follows.

$$IMCC(C_t) = ES_{R,C_\tau} \times \max\left\{ \frac{ES_{F,C_t}}{ES_{R,C_t}}, 1 \right\} \tag{5}$$

where ES_{R,C_τ} is the expected shortfall forecast for a reduced set of risk factors, calibrated over the stressed period τ; ES_{F,C_t} is the expected shortfall forecast for the full set of risk factors, calibrated over the current period t; and ES_{R,C_t} is the expected shortfall forecast for the reduced set of risk factors, calibrated over the current period t (Basel Committee on Banking Supervision, 2019b).

As mentioned earlier, ES is calculated at both the company level and at the trading desk level. The former is then calculated from the latter using the correlations coefficients of correlated trading desks. The Basel III Accord acknowledges that the correlations between trading desks tend to increase during times of crisis and these increased correlations, in turn, reduce the bank's IMCC in Equation (5). The individual trading desk's IMCC for desk i, denoted $IMCC(C_{i,t})$, is then calculated as follows.

$$IMCC(C_{i,t}) = ES_{R,C_{i,\tau}} \times \max\left\{\frac{ES_{F,C_{i,t}}}{ES_{R,C_{i,t}}}, 1\right\} \tag{6}$$

where the variables are as those defined for Equation (5), but with reference to the individual trading desk i.

The bank's total IMCC at time t, i.e., $IMCC_t$, is then calculated as follows.

$$IMCC_t = \rho \times IMCC(C_t) + (1-\rho) \times \sum IMCC(C_i, t) \tag{7}$$

where $\rho = 0.5$ (Basel Committee on Banking Supervision, 2019b).

Finally, the Basel III Accord stipulates that the bank's capital charge at time t is, therefore, as follows.

$$Capital\ Charge_t = \max\{IMCC_{t-1}, (1.5+k) \times \overline{IMCC}_{60}\} \tag{8}$$

where $IMCC_{t-1}$ is the previous day's IMCC value as calculated by Equation (7); k is the penalty factor, as detailed in Table 2; and $\overline{IMCC}_{60}$ is the average IMCC value for the past 60 trading days (Basel Committee on Banking Supervision, 2019b).

The revised penalty variable, as presented in Table 2, has not done away with the criticism and analyses presented by Begley, et al., (2017). Moreover, the penalty variable still does not take cognisance of the magnitude of the breach, as discussed by Jiménez-Matín, et al., (2009).

Daníelsson (2013) finds that 97.5% ES forecasts are more volatile than their corresponding 99% VaR forecasts. Chang, Jiménez-Martín, Maasoumi, McAleer, and Pérez-Amaral (2019)

theorise that necessarily due to ES's ability to account for tail events, its values are expected to be more volatile, at least depending on how extreme the losses captured by the tail are. They find this to not be the case, at least in the extreme events modelled in the authors' study. The adaptation of the Basel Accords, first from VaR to SVaR, and then again from SVaR to (S)ES, has stabilised the reserve levels held by banks (Liu & Stentoft, 2021). Some authors, such as He, Kou, and Peng (2022), suggest using other risk metrics that measure the tail, such as medial shortfall, although these measures are out of the scope of this study.

The Basel Accords, and their continual updating to the changing needs and risks faced by the banking sector, have attempted to keep abreast with methodological shifts in market risk management. From expanding the number of risk classes with the introduction of the Basel II Accord to the use of stressed periods to stabilise banks' reserves, it is obvious that the structural integrity of the banking sector globally, in general, and in the United States, in particular, is crucial to banking regulators and supranational institutions such as the BCBS. However, the Basel Accords leave room for interpretation and a lack of prescription, both of which leave the banks with plenty of scope for subjective decisions that may aid in gaming the system in the hopes of achieving a balance between regulatory scrutiny and operational freedom. This study explores and discusses some of these subjective decisions and highlights these further in Chapter 3.

In Chapter 4, this study then introduces a general methodology to be implemented for the discovery of the probability density function (PDF) of a bank's equities trading desk's returns[11]. This PDF, in turn, can be used to calculate various tail risk metrics, including VaR and ES. The Bayesian network (BN) model may be superior in its updating abilities to changing market values and, therefore, be able to provide more robust VaR and ES forecasts, leading to increased stability in the bank's reserves and, hence, in the banking sector. Moreover, the BN model may remove some of the subjectivity in producing market risk metrics, although not all. In fact, it may introduce subjectivity in different areas of the market risk calculation process. The BN model may, nonetheless, offer preferable processes and more accurate results relative to the existing traditional models.

[11] While the methodology presented in this study is specifically applied to a bank's equities desk, it can similarly be applied to other desks, including fixed income, interest rates, foreign exchange, and others.

3. Market Risk Management using Traditional Models

This chapter considers the implementation of models to forecast value at risk (VaR) and expected shortfall (ES) in the process of risk management for banks in the United States (US), especially due to their prominence in the Basel Committee on Banking Supervision's (BCBS's) market risk framework. It also establishes the basis for this study by examining the existing literature surrounding market risk metrics such as VaR and ES, and to assess the performances of traditional models in producing market risk forecasts. This basis will be used in subsequent chapters of this study and built upon by introducing Bayesian networks (BNs) to calculate said market risk metrics.

From a novelty perspective, this chapter offers a comprehensive assessment of traditional models when producing VaR, ES, and their stressed counterparts. The normal distribution and the skewed Student's t distribution are used to produce all four market risk metrics using various traditional models. The former's use is prevalent throughout the literature, while the use of the latter has been suggested by some studies to provide a better fit to the data (see Sections 3.2 and 3.3 for more detail). This comprehensive assessment of the various models' performances using both statistical distributions is novel, as no other studies provide such a comprehensive analysis.

This chapter's second novel contribution is the variety of backtesting methodologies used to assess the stressed market risk metrics considered in this study. These analyses are also a novel contribution, given that few studies consider the performances of the various traditional models in producing the stressed metrics, especially when using the ES backtests considered in this study.

The remainder of this chapter is structured as follows. First, this chapter discusses VaR and ES as risk metrics, before introducing some of the commonly applied models which are often employed in forecasting the two risk measures. Each model is examined, and its pitfalls are outlined to highlight the potential advantages of a BN model, detailed in the next chapter, over each common model. A literature review of the various models is then provided with reference to their performances in forecasting 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts within the BCBS's risk management framework. The data and methodology employed, together with the required backtesting techniques for each of the risk metrics, forecasting error measures, and relevant statistical tests, are also provided, followed by the results of calculating

the 99% 10-day VaR forecasts and 10-day 97.5% ES forecasts of a US bank's equities trading desk. Finally, a conclusion surrounding the performances of the various models is provided.

3.1. Traditional Models

VaR is a risk metric often employed to capture market risk. In this study, the definition of market risk as used by the BCBS is adopted, as was presented in Equation (2). The BCBS defines market risk to be the risk of losses in portfolio value due to market price movements for both on- and off-balance sheet accounts (Basel Committee on Banking Supervision, 2019a).

As mentioned in Chapter 1, VaR is a loss amount quoted as an absolute value with a pre-specified degree of certainty for a specific period. For example, a company may quote a one-day VaR forecast of $100 million at a 95% confidence level. This means that it is expected that losses made by the company will not exceed $100 million on any single trading day over a one-year horizon 95% of the time or, alternatively, it is expected that the company's losses will exceed $100 million on any single trading day on only 5% of days, over a one-year horizon.

A VaR forecast is often calculated using a set of simplifying assumptions: First, the problem is simplified by assuming that the VaR forecast is only sensitive to a limited set of sensitivity factors, f. Second, a dependence relationship is approximated between the value of the instruments examined and the set f. Third, a distribution for the set f is assumed (Pritsker, 2006). The forecasts calculated are only as reliable as the assumptions made. For example, the worse the assumed statistical distribution fits the set of sensitivity factors f, the less accurate the VaR forecast will be.

The VaR metric is calculated by applying Equation (2), regardless of which model is used to evaluate VaR. The different models chosen by the bank calculating VaR simply change how the different components of Equation (2) are calculated, but not the equation itself.

The VaR metric is a very attractive risk-reporting figure for several reasons. First, it can be relatively easy to implement, given the chosen calculation model (Daníelsson, 2002). Second, it is a summary statistic which captures an absolute loss forecast, a time horizon, and a confidence level. This allows for the measure to be easily explained to the general population, regardless of its technical skills (Daníelsson, 2002). Last, the figure is also comparable across institutions, allowing for the standardisation of risk measurement, if employed.

On the other hand, the VaR metric has its shortcomings. First, it is a single-point estimation of a profit and loss account, which cannot describe what loss can be expected should the loss

amount exceed the VaR forecast. Recall the example outlined earlier in this section. While there exists a 5% probability of exceeding a loss of $100 million on any given trading day over a one-year horizon, should this loss amount be exceeded, the VaR forecast does not provide any detail as to what the level of loss incurred will be. Second, the VaR forecast is also easily manipulated, as it only relies on the profit and loss figures provided by the reporting bank (Daníelsson, 2002).

In addition to the points mentioned above, VaR suffers from its lack of sub-additivity, meaning that it cannot simply be calculated for two separate portfolios by adding their respective VaR forecasts together to determine the VaR of the combined portfolio. Hence, unless VaR is calculated for the combined portfolio from first principles (or by applying some marginal contribution analysis), the sum of the VaR forecasts of the two portfolios may be incorrectly reported, as it may differ from the VaR forecast of the combined portfolio (Daníelsson, 2002). The issue of sub-additivity and the related implications on the coherency of a risk measure are discussed below.

Daníelsson (2002) points out while criticising (then and still) current VaR calculation models that a basic underlying assumption surrounding them is that of the stability of the statistical properties of financial data throughout time, regardless of the existence of a crisis. Daníelsson almost regards this to be an axiom of current modelling techniques, which proves a fundamental flaw in their constructions. It is, therefore, logical to assume that a model that does not assume the stationarity of the underlying process, i.e., one that can update its statistical relationships as time passes, would be superior to a model that cannot do so.

Regardless of the drawbacks outlined above, regulatory bodies such as the BCBS require banks to implement and calculate VaR as a risk management metric. While there are various models available for banks to estimate their VaR forecasts, this section focuses on several of the most commonly used models due to their relative simplicity of implementation by banks and low computing costs, namely the historical simulation model, the delta-normal model, three autoregressive models, and the RiskMetrics model. Each model is described below, its flaws highlighted, and possible reasons why estimating market risk metrics using a BN approach (as detailed in the next chapter) might be a superior alternative to each of the models are discussed.

ES, as a measure of tail loss, is often regarded as a substitute or a complement risk management measure to the VaR measure, since VaR does not quantify the risk in the tail of

the profit and loss distribution. Indeed, this is one of the primary reasons for the BCBS's shift from the use of the VaR measure to the use of the ES measure (Basel Committee on Banking Supervision, 2019a). Moreover, it is a (mathematically) superior risk metric when compared to VaR due to its coherency, a property of coherent risk measures which VaR lacks, although it was thought to pose challenges when it comes to backtesting.

While there are several criticisms of the VaR metric, it is still commonly used today by many banks and regulatory authorities, including the BCBS (Liu & Stentoft, 2021), even though some authors have criticised the measures chosen by regulators such as the BCBS (see, for example, Daníelsson, et al., 2001). One of the most prominent criticisms of the VaR measure is its failure to meet the criteria of a coherent risk measure. These criteria are made formal using four axioms, which crystallise the classification criteria for a measure to be considered a coherent risk measure. Any measure that is treated as a coherent risk measure, but fails at least one of the axioms, may lead to incalculable flaws in the risk management structure in which it exists (Acerbi & Tasche, 2002).

The definition of a coherent risk measure, adapted from Acerbi and Tasche (2002), is defined as follows: Consider a set of random variables R, whereby $\forall\, r \in R, r \in \mathbb{R}$. A coherent risk measure is a function $\varrho: \mathrm{R} \rightarrow \mathbb{R}$ which satisfies the following four axioms.

i. Monotonous: $X \in R,\ X \geq 0 \Rightarrow \varrho(X) \leq 0$

ii. Sub-additive: $X, Y, X + Y \in R \Rightarrow \varrho(X + Y) \leq \varrho(X) + \varrho(Y)$

iii. Positively homogeneous: $X \in R, r > 0, rX \in R \Rightarrow \varrho(rX) = r\varrho(X)$

iv. Translation invariant: $X \in R, a \in \mathbb{R} \Rightarrow \varrho(X + a) = \varrho(X) - a$

VaR fails to qualify as a coherent measure of risk since it fails to display the sub-additivity axiom. The measure fails to be sub-additive as it is defined to be a minimum loss incurred in the quantile examined (Acerbi & Tasche, 2002). The failure to be sub-additive implies that the VaR measure fails to capture the effects of diversification among the different portfolios held by a single entity.

The notion of sub-additivity is crucial for banks. Banks hold multiple portfolios through their various trading desks (equities, fixed income, derivatives, mortgages and other credit products, to name a few). Hence, the ability of a bank to capture diversification is paramount.

In addition to the coherency issue of VaR, by definition, it fails to recognise any loss incurred beyond its confidence level. This may be especially problematic as VaR has been

found to underestimate risk and be an unreliable risk measure under stressed market conditions, i.e., during a crisis (Yamai & Yoshiba, 2005). On the other hand, VaR is an elicitable risk measure (Acerbi & Székely, 2014), meaning that it lends itself naturally to being ranked in different applications and, consequently, allows for its backtesting to be applied directly.

ES was introduced by Artzner, Delbaen, Eber, and Heath (1997), and can be explained as the expected loss of a portfolio given the exceedance of the VaR forecast (Yamai & Yoshiba, 2005). Hence, by definition, ES is a measure which explains that risk beyond the VaR forecast. This explanation also makes it clear why ES is often referred to as conditional VaR (as the forecast is conditional on the exceedance of VaR) and expected tail loss (as it is the expected loss exhibited in the tail of the profit and loss distribution).

The definition of ES as a coherent risk measure, again adapted from Acerbi and Tasche (2002), is as follows: Consider the same profit and loss random variable X, as defined for VaR in the explanation preceding Equation (1), again with a probability level of $100\alpha\%$, specified over the time horizon t. The expected shortfall corresponding to $100\alpha\%$ is then equal to

$$ES^{(\alpha)}(X) = -\frac{1}{\alpha}\left[E\left[X\mathbf{1}_{\{X\leq x^{(\alpha)}\}}\right] - x^{(\alpha)}\left(F\left(x^{(\alpha)}\right) - \alpha\right)\right] \tag{9}$$

where $\mathbf{1}_{\{X\leq x^{(\alpha)}\}}$ is an indicator variable equal to 1 if $X \leq x^{(\alpha)}$ or 0 if $X > x^{(\alpha)}$, where $x^{(\alpha)} = \sup\{x|\ \Pr[X \leq x\} \leq \alpha\}$.

ES can also be defined explicitly in terms of VaR, as it is the expected loss given the exceedance of the VaR forecast. Hence, given a profit and loss distribution X and a probability level $100\alpha\%$, ES can be defined as the average loss exceeding the $100\alpha\%$ h-day VaR forecast, mathematically defined as follows.

$$ES_{ht,\alpha} = E\left[X\middle|X \geq VaR_{ht,\alpha}\right] = ES^{(\alpha)} \tag{10}$$

ES is, indeed, sub-additive, as has been shown in the literature (for details, see Acerbi and Tasche, 2002). In addition, ES has several important features such as completeness, universality, and simplicity (Acerbi & Tasche, 2002). Moreover, Acerbi and Tasche (2002) conclude their study with an important observation: The ES measure can be easily implemented by any bank which already calculates its VaR, with minimal additional computational costs (if any, at all).

ES is calculated as an average loss beyond the VaR forecast. Hence, it necessarily considers tail losses in all quantiles exceeding the VaR quantile. Therefore, ES is a risk measure that is

far more comprehensive than the VaR measure and, also, is necessarily computed after VaR is computed. The only drawback of ES as a risk measure is that it is not elicitable (Gneiting, 2011) as VaR is, meaning that it is not easily ranked in performance across different models. In fact, Diebold, Gunther, and Tay (1998) conclude that ES cannot be backtested at all. However, several studies have developed backtesting techniques for ES, both conditional and unconditional, which are explored in Section 3.3.3.

While the Basel III Accord requires ES to be calculated at various liquidity horizons (see Chapter 2), it is noted that the different liquidity horizons are of use when either calculating ES for the bank as a whole, and different trading desks' instruments would fall into different liquidity horizons, or when calculating ES for a desk whose instruments would fall into different liquidity horizons. As this study focuses on the equities trading desk of a US bank and, moreover, focuses on a market proxy (as discussed in Section 3.3.1), the only instruments considered in this study are what the Basel III Accord classifies as equity large capitalisation risks. The liquidity horizon attached to equity large capitalisation risks is ten days (Basel Committee on Banking Supervision, 2019b), meaning that this study does away with calculating ES for different liquidity horizons.

3.1.1. The Historical simulation Model

The historical simulation model is the simplest and least computationally intensive model. This makes it the most used model to calculate VaR and, by extension, ES, especially by banks (Pritsker, 2006).

The model makes use of the historical distribution f of the set of sensitivity factors to calculate profit and loss amounts for a portfolio of instruments as if this portfolio was held for a historical period of length N days (Pritsker, 2006). It is common to choose N to be 252 days, as there are approximately 252 trading days in any given year. The key assumption of this model is that past returns will reoccur in the future. The historical period allows for the calculation of the mean and standard deviation of the historical distribution, which can then be used in Equation (2).

While this model does not assume the distribution of the returns, it has several flaws. First, the model assumes an equal probability weight contribution for each trading day of N^{-1}. This explicitly assumes that the returns are independently and identically distributed throughout the period (Pritsker, 2006). This is often acknowledged by banks which implement this model. Second, the method requires a very large dataset to compute the distribution of past returns.

Third, the model makes the explicit assumption that past performance is indicative of future performance, thereby nullifying the efficient market hypothesis and weak form market efficiency. Fourth, the model disregards all notions of return volatility varying with time due to its first assumption. This means that, if 250 days are used, the VaR contribution from 250 days ago is as important as the VaR contribution from yesterday, which does not take account of market changes over time. It is expected that more emphasis should be placed on more recent contributions. Last, since the historical simulation model reflects an average VaR forecast for the historical period, a lag in the incorporation of new information as it becomes available is expected, thereby either costing the bank money (as excess capital in the form of reserves is held unnecessarily) or leading to further default days (breaches), where reserves calculated using VaR do not cover the loss experienced (as insufficient capital is held).

Displaying the above flaws of the historical simulation model, for example, JP Morgan Chase & Co., a US bank, notes that its historically simulated VaR forecasts rely on past data and are, therefore, inaccurate predictors of future losses (JP Morgan Chase & Co., 2022). The bank implicitly acknowledges and provides the necessary legal disclaimers relating to the calculation of VaR using the historical simulation model in its Form 10-K filed with the US Securities and Exchange Commission (SEC).

An evaluation of VaR utilising a BN, as proposed in this study and explored in the next chapter, would eliminate some of the flaws discussed above, while maintaining the benefit of the historical simulation model. Firstly, since a BN's output is a probability density function (PDF), the issue of an assumed distribution for the set f, as discussed earlier, is avoided. This overcomes one of the biggest hurdles when it comes to applying accurate calculations using the VaR forecasts, regardless of which model is used. Moreover, the BN will allow a forecasted return to be used in the return PDF achieved by the portfolio of exposures, incorporating forward-looking predictions. The BN model will, nevertheless, still require a large dataset to compute the return PDF. However, should the dataset be already available for the application of the historical simulation model, the BN application is believed to offer a more robust estimate of VaR that can be updated as soon as new information is available.

3.1.2. The Delta-Normal Model

The delta-normal model (also known as the variance-covariance model) assumes that each of a company's risk factors and, hence, returns, is governed by a normal distribution. This is

then extended to imply that the joint effects of these underlying factors are governed by a multivariate normal distribution (Linsmeier & Pearson, 2000).

To apply the model, the underlying factors which affect the returns of the portfolio under examination must be determined (Linsmeier & Pearson, 2000). It is then assumed that the contribution of risk of each factor is normally distributed, with a combined effect that is governed by a multivariate normal distribution.

While the former activity of identifying the risk factors may be a difficult task, it is not impossible. The latter, the normality assumption, is the most fundamental flaw of this model.

There exists ample evidence that the returns of financial instruments are not normally distributed but are, in fact, leptokurtic, meaning that their distributions' peaks are higher than those of normally distributed variables and, therefore, exhibit fatter tails. This, in turn, suggests the existence of more extreme outliers than predicted under the normal distribution assumption (Peiró, 1999). The assumption is further contradicted when it comes to smaller datasets which cover a shorter period, as it cannot be assumed that the central limit theorem applies then. While this model may not assume that the distribution exhibited by past returns will be repeated, it explicitly assumes a single standard distribution for all returns and is more computationally intensive than the historical simulation model (mainly due to the analysis performed on multivariate normally distributed variables).

In addition, banks may neglect to update and maintain the matrix of dependencies used in this model frequently enough to capture changing market conditions timely (Daníelsson, 2002). Other issues relating to the feasibility of this model include the increasingly large size of (and, consequently, the computational power required to calculate) the matrix. In its revised 2021 Form 10-K filing with the US SEC, Citigroup, for example, notes that it uses a Monte Carlo simulation with 'approximately 450,000 market factors' to forecast VaR (Citigroup Inc., 2022).

If a BN was applied to forecast VaR and ES instead, the normality assumption would not be required, as the belief network will output a (return from a) PDF from which moments and quantiles can be calculated. This may lead to more robust VaR forecasts.

3.1.3. Autoregressive Models

Since the introduction of the 1996 Market Risk Amendment[12], banks have been granted permission to make use of more sophisticated VaR models as their internal models, subject to the backtesting methodology stipulated by the BCBS (see Table 1). This, coupled with the increased computational capacity of the average computer, has made the use of autoregressive models more common in financial applications.

Some of the autoregressive models often found to be employed to forecast VaR include the autoregressive conditional heteroscedasticity (ARCH) model, the generalised autoregressive conditional heteroscedasticity (GARCH) model, and the exponential generalised autoregressive conditional heteroscedasticity (EGARCH) model, among others[13]. Each of these three autoregressive models is discussed below. However, before any mathematical definitions are provided, it is important to understand why these models may be useful in forecasting VaR. Note that these autoregressive models all deal with heteroscedasticity. Heteroscedasticity is the assumption that the square of the expected sum of error terms in an ordinary least squares model is not constant throughout the data, i.e., heteroscedasticity is the opposite of homoscedasticity (Engle, 2001), and assumes that the variance of a measure varies with time. ARCH models are then used to estimate this non-constant square of the expected sum of error terms, i.e., they are tasked with estimating the variance and, thereby, heteroscedasticity.

The heteroscedastic attribute of the autoregressive models discussed below allows them to incorporate non-constant volatility into their models and, therefore, any volatility changes experienced in the underlying time series (Giot & Laurent, 2004). This is a clear advantage of the autoregressive models when compared to the less sophisticated models outlined above (the historical simulation model and the delta-normal model), as those incorporate changes in volatility very inefficiently in comparison.

[12] See Basel Committee on Banking Supervision (1996).

[13] There are extensive variations of the autoregressive models available, with numerous degrees of complexity – for more details, see Bollerslev (2007). However, the three models mentioned in this study, namely the ARCH model, the GARCH model, and the EGARCH model, represent the most commonly used autoregressive models applied in producing VaR and ES forecasts.

While three autoregressive models are introduced below (namely the ARCH model, the GARCH model, and the EGARCH model), they are similar in the sense that they attempt to model the heteroscedasticity inherent in the underlying time series data, but distinct in the characterising equation of incorporating such heteroscedasticity and, therefore, volatility. Hence, it may be useful to address the three autoregressive models as a group when assessing their drawbacks in comparison to the BN approach proposed in this study and explored in the next chapter.

A drawback of the GARCH model is its mean reverting tendencies (Engle, 2001). This is not a characteristic of financial time series data that is always present and, hence, is not an accurate modelling assumption. Since the EGARCH model is a sub-model of the GARCH model, it too may suffer from this characteristic. The BN approach proposed in this study does not have any mean reverting tendencies built-in by default, suggesting its superior ability to model financial time series data.

Moreover, note that the autoregressive models introduced below (together with the RiskMetrics model, see Section 3.1.4) are iterative in some shape or form. The models take their own forecasts (i.e., output) for period t as inputs into their forecasts for period $t + 1$. Hence, even if the model employs some form of asymptotically diminishing weighting of contributions to the forecast (as is the case with the RiskMetrics model), it cannot exclude any past forecasts in its prediction process, even if such forecasts were statistically inaccurate. The BN approach, on the other hand, replaces forecasts with actual returns as time passes, thereby excluding any previous forecasts which may have been incorrect. Therefore, past errors are not necessarily built on further, thereby avoiding the exponentially increasing prediction error.

Last, all of the autoregressive models detailed below require the calibration of model (hyper) parameters which are sometimes taken to be some default values. This approach would not suffice under the BN approach which would calibrate any given parameters to optimise the process undertaken, or has extended models offered in the literature to assist with the calibrator of such a network to make more informed and theory-backed decisions, as no default values are usually available. Hence, even though this study recalibrates these model (hyper) parameters for the autoregressive models (which is not always done in the literature), the calibration of a BN model may be more precise when compared to the (default) calibrations of autoregressive models, as used in this study or as often used in the literature, even if simply

due to the availability of default values and the tendency of the calibrator of such models to opt for such defaults.

3.1.3.1. The Autoregressive Conditional Heteroscedasticity Model

Robert Engle first applied the ARCH model to financial data when assessing the uncertainty which revolved around inflation in the United Kingdom at the time (Engle, 1982). However, the use of ARCH models in the literature has been clear in many scenarios where econometric methods were needed to estimate changes in volatility in underlying time series data (Bollerslev, 2007).

The ARCH model focuses its attention on modelling the residuals of a time series, i.e., the ε_t terms of the time series. These residual terms are defined, in turn, by the product of the volatility of the time series, captured at time t by the standard deviation variable σ_t, and some random stochastic term, z_t, yielding the following equation.

$$\varepsilon_t = \sigma_t z_t \tag{11}$$

where the z_t is a series of independently and identically distributed standard normal random variables (Bollerslev, 2007).

Hence, the squared volatility (i.e., the variance) of the time series at time t can be modelled as follows.

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 , \qquad \alpha_0 > 0, \qquad \{\alpha_i\}_{i>0} \geq 0 \tag{12}$$

It is important to note that an ARCH model must be specified together with its order. Equation (11) and Equation (12) together define an ARCH model of order q, i.e., an ARCH(q) model. The order of the ARCH model represents the length of the lags built into the ARCH model to represent the lags believed to exist in the underlying time series data (Engle, 1982).

While the order of the ARCH model should be estimated every time from the data, financial data analyses make use of the ARCH model of order 1 almost exclusively. Hence, while Engle (1982) does provide a procedure to calibrate the order of an ARCH model, this study employs the ARCH(1) model for consistency with the literature and comparison purposes, and any references to the ARCH model in this study are references to the ARCH(1) model.

3.1.3.2. The Generalised Autoregressive Conditional Heteroscedasticity Model

Tim Bollerslev developed the GARCH model in 1986 as an alternative to its base, the ARCH model. The GARCH model, as does the ARCH model, makes use of squared residuals. However, the contribution of squared residuals at time τ diminishes asymptotically to zero as time τ moves further away from the current time t, i.e., as the contribution moves further into the past (Engle, 2001).

The variance prediction characteristics of the GARCH model involve a combination of previous volatility information (the long-run average volatility as calculated by the model) and any relevant current period information which is deemed to be new to the model.

The variance predicted by the GARCH model that is applied to financial time series data is that of the regression on return r_t, which can be calculated as a linear combination of averaged past returns, m_t, and the product of the standard deviation of past returns, σ_t, and the residual term, ε_t, i.e., $r_t = m_t + \sigma_t \varepsilon_t$. Hence, the GARCH model outputs a predicted variance of return r_{t+1}, i.e., σ_{t+1}^2, as follows.

$$\sigma_{t+1}^2 = \omega + \alpha \sigma_t^2 \varepsilon_t^2 + \beta \sigma_t^2 \tag{13}$$

where ω, α, and β are called the GARCH model parameters[14], and these must be estimated based on the following conditions.

First, each model parameter is strictly greater than zero, i.e., $\alpha, \beta, \omega > 0$. Second, the sum of α and β is strictly less than 1, i.e., $\alpha + \beta < 1$. This is required as the model's long-term average is defined to be $\sqrt{\omega/[1-(\alpha+\beta)]}$, which yields the second condition. The GARCH model parameters may be computed by maximising a log-likelihood parameterisation of the model. However, they are often outputted by most software which are ascribed to deal with GARCH forecasting, making this type of model relatively easy and cheap to implement (from a computational cost point of view).

The GARCH model's sensitivity to any market shocks is captured by α, which is the error parameter of the GARCH model. The parameter β is called the lag parameter. It captures the

[14] As stated previously, this study recalibrates the model parameters regularly with each new day that is added to the iterative calibration period, a methodology which is believed to be more accurate than that used in the literature, whereby the model parameters are estimated once from the data and are not updated.

volatility's persistence in the model. Hence, the sum of the two, i.e., $\alpha + \beta$, is a measure which captures the rate at which the GARCH model's predicted conditional variance converges to the model's long-term average conditional variance.

When a GARCH model is specified, two defining parameters must be specified to formulate the model. The parameter p (also known as the ARCH term) specifies the number of autoregressive lags of the model, while the parameter q (also known as the GARCH term) specifies the number of moving average lags of the model. Together, the two terms specify a GARCH(p, q)[15] model (Engle, 2001). Equation (13), above, defines a GARCH model with one autoregressive lag and one moving average lag, i.e., it specifies a GARCH(1,1) model.

Similar to the treatment of the order of the ARCH model, the ARCH term and the GARCH term of a GARCH model should be estimated from the data every time a model is fit to the underlying data. However, financial data analyses make use of the GARCH(1,1) model almost exclusively. Hence, this study employs the GARCH(1,1) model for consistency with the literature and comparison purposes, and any references to the GARCH model in this study are references to the GARCH(1,1) model.

3.1.3.3. The Exponential Generalised Autoregressive Conditional Heteroscedasticity Model

Modern portfolio theory often regards volatility to be a non-negative quantity, i.e., a measure that is at least zero. When considering the autoregressive equation defining the GARCH model, i.e., Equation (13), it is clear that the variance is a non-negative quantity taken as a linear combination of the three model parameters which are, by definition, positive (due to the conditions set out in the preceding section), some of which are multiplied by squared quantities. Hence, the variance calculated in Equation (13) is non-negative, i.e., at least zero (exclusive, due to the condition that $\omega > 0$). This, in turn, necessarily implies that the volatility forecasted by the GARCH model, as defined in Equation (13), is non-negative as it is the square root of a non-negative number.

This non-negative property of the volatility of the market can be made explicit even further in the construction of an autoregressive model, as done in 1991 by Daniel B. Nelson when he

[15] Bollerslev (1986) also specifies that q is to be assumed to be strictly greater than zero, while p is assumed to be greater than or equal to zero, i.e., p must take on a non-negative integer value.

developed the EGARCH model. The EGARCH model was proposed as a solution to some of the limitations of the GARCH model as pointed out by Nelson (1991). The model's primary objective is to facilitate the modelling of the asymmetry that exists when analysing the relationship between daily returns and the volatility attached to those. This objective is then met with an explicit non-negative condition imposed on the variance of returns (and, therefore, the volatility of returns), by incorporating a logarithmic treatment of the variance in the definition of the EGARCH model.

The EGARCH model is, therefore, defined as follows.

$$\log(\sigma_t^2) = \alpha_t + \sum_{\phi=1}^{\infty} \beta_\phi g\left(z_{t-\phi}\right), \qquad \beta_1 = 1 \tag{14}$$

where α_t is defined for $t \in (-\infty, \infty)$, β_ϕ is defined for $\phi \in [1, \infty)$, and the series $\{\alpha_t\}$ and $\{\beta_\phi\}$ are non-random series of scalars, i.e., $\alpha_t, \beta_\phi \in \mathbb{R}\ \forall\ t \in (-\infty, \infty), \forall\ \phi \in [1, \infty)$. Furthermore, z_t is again a series of independently and identically distributed standard normal random variables, and $g(z_t) \coloneqq \theta z_t + \gamma(|z_t| - E[|z_t|])$, where θ and γ are called the model coefficients associated with the GARCH model described by Equation (13) (Nelson, 1991).

Note that the model parameters (θ and γ) do not require any restrictions (as was the case for α, β, and ω in the GARCH model) as the value of the right-hand side can take on any value on the real number line due to the presence of the logarithmic argument on the left-hand side[16].

Similar to the GARCH model, an EGARCH model is also referenced as an EGARCH(p, q) model, where p and q play similar roles to the ones ascribed to their counterparts in the GARCH model (see the previous section for more detail). Hence, it is reasonable to conclude that an appropriate application of this model would require the estimation of the values of p and q from the underlying time series data. However, in line with the treatment of the ARCH model and the GARCH model, financial time series data analyses make use of the EGARCH(1,1) model almost exclusively. Therefore, this study employs the EGARCH(1,1) model for consistency with the literature and comparison purposes, and any references to the EGARCH model in this study are references to the EGARCH(1,1) model.

[16] Once again, this study recalibrates the model parameters regularly with each new day that is added to the iterative calibration period, a methodology which is believed to be more accurate than that used in the literature, whereby the model parameters are estimated once from the data and are not updated.

3.1.4. The RiskMetrics Model

The RiskMetrics model was introduced as a model to produce VaR forecasts in 1994 by the US bank J. P. Morgan & Co., through its subsidiary, the RiskMetrics Groups, Inc. (RiskMetrics Group, Inc., 2001).

The RiskMetrics model is built on the assumption that the logged daily return of the underlying financial time series data given the filtration system at time t, $\mathcal{F}_t$ (i.e., given all information available prior to period t), has a conditional normal distribution with a zero mean and a variance equal to σ_t^2. Mathematically, the RiskMetrics model is built on the assumption that $\log r_t \,|\mathcal{F}_t \sim N(0, \sigma_t^2)$.

The RiskMetrics model is, therefore, a predictive model for the variance at time t which can be defined as follows.

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)r_{t-1}^2 \tag{15}$$

where the smoothing parameter $\lambda \in (0,1)$ and the previous day's squared daily return serve as the market volatility proxy for the previous day (McMillan & Kambourourdis, 2009).

A manipulation of Equation (15) shows that the RiskMetrics model can be stated as an exponentially weighted moving average model (González-Rivera, Lee, & Yoldas, 2007). This means that the RiskMetrics model can be stated as follows, with all variables defined as above.

$$\sigma_t^2 = (1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau-1}r_{t-\tau}^2 \tag{16}$$

As with the autoregressive models, λ is the RiskMetrics model's parameter, and should be estimated from the underlying time series data. However, it is common industry practice to set λ to equal 0.94 (Pafka & Knodor, 2001) and, therefore, this is the value used in this study.

The very few inputs required to calibrate this model make the RiskMetrics model an easy-to-use option, which is one of the reasons why this model was so popular among practising firms (McMillan & Kambourourdis, 2009). The inputs required are a return for period t, r_t, the volatility for period t, σ_t, and the model's parameter, λ, which, as mentioned, can either be estimated from the data or taken to be equal to 0.94. These can then be used, together with Equation (16), to forecast the variance of period $t+1$.

Another advantage of this model over the more simplistic models (such as the historical simulation model and the delta-normal model) is its weighting property as displayed in

Equation (16). This allows the RiskMetrics model to forecast future variance by placing more emphasis on recent forecasts as opposed to forecasts that are in the relatively distant past. This property allows the RiskMetrics model to capture recent market volatility better and, therefore, calibrate forecasts to resemble the market more closely.

On the other hand, the RiskMetrics model's weights asymptotically approach zero. This means that any and all volatilities experienced by the model and the market (rather than just the market, see Equation (16)) contribute towards the next forecast. The model is incapable of excluding any inaccurate forecasts – a property that the BN approach proposed in this study and discussed in the next chapter does by replacing forecasts with actual values as time passes (see Chapter 4).

Last, one of the major criticisms of this model is its failure to predict the 2008 global financial crisis, considering its wide use (McMillan & Kambourourdis, 2009). Moreover, its ability to adapt to the new regime was inadequate and resulted in government intervention in the US market. Hence, even though the RiskMetrics model may have been popular in the past, its predictive and adaptive abilities are inadequate in dealing with the black swan events which banks fear so much, the same events that the BCBS regulations (see Chapter 2) attempt to prepare banks for and protect against. Hence, there is room for a new model to be introduced, a model whose predictive abilities allow for the discarding of irrelevant information coupled with the efficient ability to adapt to the regime-switching nature of crises. The BN model is theoretically suitable to do exactly that.

As introduced in Chapter 2, the lack of prescription offered by the BCBS means that the use of these models requires several choices to be made by the banks. First, the models themselves must be chosen. As stated, this study only considers the ARCH model, the GARCH model, the EGARCH model, and the RiskMetrics model, as these are commonly used in the literature. However, other autoregressive models exist, and a bank may pick any one of those to produce market risk forecasts. In addition, these autoregressive models require a distribution to be fitted to the profit and loss data of the bank (or its trading desk), to allow VaR forecasts and ES forecasts to be produced. The choice of the underlying statistical distribution is also left to the bank. In this study, the normal distribution and the skewed Student's t distribution are used as the underlying statistical distributions, and the rationales for these are discussed in Section 3.2. However, a bank may pick any distribution, due to the lack of prescription. Finally, these autoregressive models require the calibration of model parameters (as discussed under each

model). The frequency with which these are updated, if at all, is at the bank's discretion. This study updates these model parameters with each iteration of the model for all autoregressive models to facilitate the most accurate reflection of the data and, hence, to provide what are believed to be the most accurate forecasts. However, the bank may update these parameters less frequently or not at all, if these updating frequencies produce more 'favourable' results.

3.2. Literature Review

This section aims to bring together the sub-sections above and discuss the relevant academic literature surrounding the applications of various market risk forecasting models in the context of banks governed by the Basel Accords. Where available, discussions of the practical drawbacks of the various market risk forecasting models are provided to supplement the theoretical drawbacks discussed under each model's sub-section above.

The recent spree of bank failures in the US, starting in 2023 with Silicon Valley Bank through to the most recent 2024 collapse of Republic First Bank, highlighted that the continuing operations of US banks are crucial to the prospering operation of the local and global economies. Given how crucial banks' continuing operations are to both the US economy and to the global economy as a whole, it is unsurprising that the literature covering the performance of market risk models used by banks and the evaluation of such is vast. A notable early opinion piece is that by Daníelsson, et al., (2001). In their response to a request for comments by the BCBS, the authors praise the (then) proposed move to a risk-based capital framework using the VaR metric. However, in their response, the authors highlight that many of the models employed by banks to calculate VaR tend to fall short when reporting the combined VaR forecasts for different asset classes. This came in light of the introduction of operational risk and market risk by the Basel II Accord, in addition to credit risk, which was the sole focus of the original Basel Accord.

Daníelsson, et al., (2001) also claim that, since the market itself generates volatility, risk is an endogenous factor, and this is of fundamental importance, especially in times of crisis. The failures of models often employed to produce VaR forecasts to account for the endogeneity of risk (as opposed to its commonplace treatment as an exogenous factor) result in unreliable VaR forecasts (Daníelsson, et al., 2001).

A final important remark provided by Daníelsson, et al., (2001) highlights the relative theoretical ease with which banks can manipulate their VaR forecasts, should they wish to do so. The authors state that, since a VaR forecast is a point estimate, a bank can theoretically

simply shift its risks further into the tails of its risk distribution by using various readily available derivatives products. While the authors explicitly point out the use of options, the use of other over-the-counter instruments, such as swaps, may exacerbate this, to the extent that their added charges (due to being over-the-counter) counteract the charge that the bank is trying to avoid by implementing such a strategy. Until banks are forced to report more informative risk measures of their tail risks, such as ES forecasts, and reserve capital against such measures, there is no real reason for banks not to shift risks around to lower their VaR forecasts and, therefore, their capital reserves. This is further supported by Daníelsson (2002), as mentioned in Section 3.1, above.

Daníelsson (2002) further highlights the volatility of VaR forecasts when forecasted using commonplace models (such as the historical simulation model and the delta-normal model). In fact, Daníelsson shows that the volatility of VaR forecasts is similar in magnitude to the volatility of market returns themselves using the Standard and Poor's (S&P) 500 index over the study period and a 99% confidence level. Should a bank maintain its capital reserves at the level prescribed by the model, such volatility in forecasts will lead to breaches – an undesirable outcome for a bank. Daníelsson finds that the historical simulation model is among the least volatile models, necessarily due to its unconditional nature and slow adaptation of market volatility. Hence, by making use of the historical simulation model to produce VaR forecasts, banks can artificially 'smooth' their VaR forecasts and, therefore, reduce the number of breaches incurred.

Daníelsson (2002) makes an important remark, stating that any model designed to assess and predict risk is useless if applied within a regulatory framework. The study refers to the 1998 Russian sovereign debt default as an example, highlighting that, due to common regulatory frameworks applied, when breaches of VaR forecasts were encountered, market participants uniformly fled to holdings of safe assets, sharply lessening the liquidity of such instruments and, as a direct consequence, further feeding the crisis, in a positive feedback loop. This, in turn, suggests that the mere fact that risk is modelled (as required by regulatory bodies) affects risk itself. This might suggest that a constantly evolving approach to modelling risk, such as that proposed in this study, is of greater benefit to the regulator and market participants relative to the models currently employed.

Covering several developed markets, McAleer and da Veiga (2008) calculate VaR forecasts for portfolios of several indices and find that banks are encouraged to calculate VaR using

models which understate VaR while maintaining a number of breaches which is not considered so excessive as to alarm the regulator. This means that banks can benefit by employing VaR models that lead to multiple breaches and, therefore, a higher penalty (as captured by the penalty variable, k, in Table 1), as the penalty incurred is still cheaper than the additional capital expected to be carried by banks at the standardised regulatory capital level required of them (a level which is often higher than that calculated by the internal model). This allows a bank to balance its opportunity costs and the cost of additional capital held due to the breaches, resulting in a cost-benefit analysis which may suggest that the bank should hold less capital and experience more breaches.

Sharma (2012) summarises the historical simulation model by describing it as a model which backtests well and conforms to regulatory requirements, while displaying mixed performance under hypothesis testing scenarios, and failing when the independence of breaches is tested. Hence, the historical simulation model is a primary candidate for the aforementioned cost-benefit analysis exercise.

On the other hand, several studies find that VaR forecasts provided by banks, especially those forecasted using the historical simulation model, are often conservative and, therefore, inaccurate (Berkowitz & O'Brien, 2002). This contradicts the notion that VaR forecasts, especially volatile ones, may be underestimating the risk. Berkowitz and O'Brien (2002) find that, while few breaches are experienced by US banks during the study period, the breaches are not independent of each other, i.e., they are often clustered together. The authors claim that a possible explanation for this is that banks fail to grasp the concept of sub-additivity accurately and the effects of diversification across multiple business lines in an aggregated VaR forecast.

Pérignon, Deng, and Wang (2008) claim to be the first empirical study of non-anonymous banks' risk management frameworks. The authors use Canadian banks' profit and loss and VaR forecasts from 1 November 1999 to 31 October 2005 to determine whether Canadian banks underestimate or overestimate their VaR forecasts. While the authors recognise that banks do have an incentive to underestimate their VaR forecasts for market risk (to induce lower capital requirements), the Canadian banks examined seem to have exercised excessive conservatism when it comes to their VaR forecasts (Pérignon, et al., 2008) (and, therefore, their capital requirements). In fact, only two breaches were found by Pérignon, et al., relative to an expected number of breaches of 74 (for a 7,354 out-of-sample trading-day period). The authors find that banks overestimate their VaR forecasts by up to 79% and suggest that (Canadian) banks value

their reputations and recognise the potential harm that excessive breaches may have on such. This contradicts the findings presented by McAleer and da Veiga (2008), as discussed above. An alternative reason may be the banks' failure to account for diversification benefits between business lines, echoing the issue of the sub-additivity of VaR as a risk measure.

Pérignon and Smith (2010) build on the study conducted by Pérignon, et al., (2008). The authors find that the most common VaR forecasting model for international banks is the historical simulation model, and that the forecasts are often conservative, leading to few breaches (Pérignon & Smith, 2010). Moreover, Pérignon and Smith find that the historical simulation model suggests very little about future market volatility and, hence, provides VaR forecasts which are questionable and of little use.

On the other hand, several empirical studies (such as that conducted by Kuester, Mittnik, and Paolella, 2006) conclude that the historical simulation model falls within the 'Yellow zone' according to the Basel II Accord when producing one-day VaR forecasts. This means that at least 5 and at most 9 breaches are expected in any 250-trading-day period or, as reported by Kuester, et al., (2006) in their study, the historical simulation model exhibited between 80 and 98 violations over approximately four years using one-day VaR forecasts. This, in turn, suggests that the model is prone to underestimate market risk, leading to the increased likelihood of crises taking place, considering the systemic risk present in the (US) banking sector. This claim can be further emphasised by pointing out that VaR models in general, and the historical simulation model in particular (being the most commonly used VaR estimation model by banks, as found by Pérignon and Smith, 2010), failed to predict and prepare the financial industry for the 2008 global financial crisis (Laurens, 2012).

Berkowitz, Christoffersen, and Pelletier (2009) perform backtesting on the historically simulated VaR forecasts of four business lines of an international bank based on actual data provided by the bank. The authors conclude that, for three of the four business lines, the VaR forecasts are statistically inaccurate, and the number of breaches is lower than or equal to that expected. The authors further perform regression analysis of the actual profit and loss figures to the VaR forecasts as predicted by the historical simulation model and conclude that the bank's risk management operations could be improved if the bank incorporated some dynamic VaR forecasting model as opposed to the historical simulation model.

An issue highlighted by Berkowitz, et al., (2009) is that, at the time, there was no recommended and enforced model to perform backtesting, as the BCBS's proposed traffic light

test (see Table 1 and Section 3.3.3.1) is not actually a statistical test, but rather a penalty classification system. This suggests that banks (and other financial institutions) were able to manipulate the validity of their internal VaR models by applying whichever model led to the lowest number of breaches and/or whichever backtesting technique led to positive conclusions surrounding the VaR forecasting model chosen. In the case of the specific international bank examined by the authors, two of the four business lines' VaR forecasts were rejected by multiple backtesting techniques, while the third was rejected by the Kupiec test of unconditional convergence (see Section 3.3.3). This means that the bank could have employed any backtesting technique to validate the third business line's VaR forecasts, as long as it did not use the Kupiec test of unconditional convergence, leaving another major loophole in the regulation of VaR forecasts as calculated by the historical simulation model.

da Veiga, Chan, and McAleer (2012) point out that possible reasons for a bank to over-report risk include the possibility that a bank's investors may scrutinise the institution if it does not take enough risks, and that increased breaches produced by the model chosen as tested against under-reported risk may ignite increased regulatory attention. However, this may not be the case in jurisdictions where regulation is not as well enforced, and where investors are not as sophisticated (da Veiga, et al., 2012).

More recently, O'Brien and Szerszeń (2017) tested the performance of risk measures of US banks before, during, and after the 2008 global financial crisis. The authors find that the internal models used by banks produced conservative VaR forecasts, leading to few breaches in the periods preceding and following the global financial crisis. On the other hand, crisis period (June 2007) breaches were significantly higher than predicted and clustered often (O'Brien & Szerszeń, 2017). The authors highlight the banks' models' inability to adapt to changes in market conditions and the slow updating given new information, specifically during periods of market turmoil. While the banks' internal models produced conservative VaR forecasts during more stable periods (those preceding and following the crisis), supporting the findings of studies such as that of Berkowitz and O'Brien (2002), GARCH-based models experienced fewer breaches and increased independence among such breaches during the crisis. This shows the increased benefits of more accurate and timely volatility adjustments relative to models such as the historical simulation model.

Building on the work presented by O'Brien and Szerszeń (2017), Liu and Stentoft (2021) show that banks are internally motivated to use VaR models to project stable VaR forecasts

with an aim to minimise the volatility of such forecasts. Moreover, banks are also incentivised to use misspecified models, i.e., models that do not accurately or timely update VaR forecasts to account for changing market conditions, supporting the findings presented by O'Brien and Szerszeń (2017). The misspecifications of internal models support the general findings in the literature that VaR forecasts based on existing models are inaccurate, thereby leading to either conservative or aggressive forecasts.

Further building on the relatively recent finding that incorporating the heteroscedastic nature of volatility would enhance US banks' internal models' performance when it comes to producing VaR forecasts (O'Brien & Szerszeń, 2017), this study aims to fill a gap by developing a methodology that enhances the volatility updating nature of GARCH-based models. A BN model is presumed to be both implementable and more accurate than the historical simulation model while maintaining the lack of distributional assumptions as made by more complex models such as the autoregressive models. The BN model could potentially offer this lack of distributional assumption advantage, while also offering more accurate (and potentially frequent) volatility updating of returns for the profit and loss account, as continual incorporation and updating of market variables take place. This, in turn, is believed to offer a superior model of calculating VaR forecasts for banks, which will result in fewer breaches and increased confidence in the regulatory framework surrounding banks, should it be adopted and enforced by the national regulators and supranational organisations such as the BCBS.

Moreover, the model is less prone to manipulation. Since the output of the BN is a PDF, risks cannot be 'shifted' to the tail using derivative instruments, as the tails are included in the model's output. While the implementation of a BN alone as a VaR forecasting model cannot overcome the issue of choice of backtesting technique chosen to minimise breaches, the model is believed to offer more accurate and up-to-date VaR forecasts due to its inherent updating nature, which is explored further in Chapter 4.

Turning to examine the literature surrounding ES and its backtests, the literature's call to move from a VaR focus to one which quantifies tail risk was amplified at the turn of the century with several key studies surrounding market risk[17]. An early study considering the backtestability of ES is that of McNeil and Frey (2000). In their study, the authors use a

[17] See, for example, Agarwal and Naik (2004); Yamai and Yoshiba (2005); Kondor, Pafka, and Nagy (2007); and Lucas and Siegmann (2008).

combination of the historical simulation model, the RiskMetrics model, the ARCH model, the GARCH model, and extreme value theory (EVT) to model stationary return time series with stochastic volatility to estimate and use these to evaluate the performance of both VaR forecasts and ES forecasts. The study examines the returns over several indices, individual shares, and currencies, over periods ranging from 1960 to 1997, using various confidence levels. McNeil and Frey propose a bootstrap test (based on the work of Efron and Tibshirani, 1993) for the residuals and examine the discrepancy (i.e., magnitude) of any breach, using both a normal distribution and an EVT approach, and find that the assumption of normality to be significantly incorrect across all return series and confidence levels. The finding that the normal distribution as an underlying return assumption is statistically incorrect implies that distributions with higher moments different to those of the normal distribution, i.e., kurtosis and skew, may provide better ES forecasts and better backtesting results.

Building on the work presented by McNeil and Frey (2000), Wong, Fan, and Zeng (2012) use the saddlepoint backtesting technique proposed by Wong (2008) to backtest ES forecasts obtained over the period 1980 to 2008 for seven markets on both an annual basis and a full period basis. The authors backtest their ES forecasts by standardising the returns to achieve returns modelled by the standard normal distribution and compare the observed quantiles to those expected. The authors find that using the normal distribution as the underlying distribution produces inadequate risk forecasts. While the use of higher moments different to those of the standard normal distribution has produced improvements in the ES forecasts, these improvements were found to be marginal.

A discussion of the backtestability of ES would be incomplete without a discussion of the issue surrounding the measure's elicitability. As highlighted in a footnote in Chapter 1, ES's lack of elicitability means that it is perceived to be more challenging to backtest any forecasts of the measure due to the inexistence of a consistent scoring function to be used to rank such forecasts (Gneiting, 2011). However, as will be discussed in Section 3.3.3, some authors, such as Acerbi and Székely (2014), provide backtests which do not rely on the elicitability of a risk measure, or backtests that rely on the joint elicitability of VaR and ES – both approaches, in general, seem to have satisfied the debate on the matter.

One of the most commonly used approaches to backtest ES is the traffic light test of the BCBS (Chen, 2018), as depicted in Table 2. As discussed in Chapter 2, the traffic light test uses a cumulative probability approach to classify models into one of three 'zones', namely a

‘Green zone’, a ‘Yellow zone’, and a ‘Red zone’. Other notable ES backtests are described in Section 3.3.3 and are implemented in this study, namely the conditional backtest (see Section 3.3.3.4), and the unconditional backtest (see Section 3.3.3.5), introduced by Acerbi and Székely (2014), as well as the minimally biased backtest (see Section 3.3.3.6) introduced by Acerbi and Székely (2017), and, finally, the Du-Escanciano independence test, introduced by Du and Escanciano (2017) (see Section 3.3.3.7).

The backtests introduced by Acerbi and Székely (2014) assume that the independence of any ES breaches has been tested independently of the statistical backtests proposed (Acerbi & Székely, 2014). He, Kou, and Peng (2022) highlight that the three tests proposed by Acerbi and Székely are indirect backtests[18] of ES, meaning that the accuracy of the backtests deteriorates for larger banks, implying the larger bank’s increased ability to under-report ES relative to smaller banks (He, et al., 2022). This may have regulatory consequences in real-world applications, although only in economies that require banks of all sizes to comply with the Basel Accords. This is not the case in the US, where only first-tier banks are required to comply with the Basel Accords, while compliance for smaller banks is optional (Herring, 2007).

Du and Escanciano (2017) present a conditional backtest akin to Christoffersen’s test for the independence of VaR forecasts. In their test, Du and Escanciano employ Monte Carlo simulations to test the performance of their conditional cumulative violations test. The test is not without criticism. For example, Wang, Wang, and Ziegel (2023) note that the Du-Escanciano independence test requires distributional assumptions, is a two-sided test of misspecification, as opposed to an under-estimation test, and is not valid until a full calibration period of a fixed length has been observed. Moreover, it is highlighted that this test only examines the statistical accuracy of the distribution used to produce ES forecasts, as opposed to the use of the forecasts themselves, thereby resulting in no differences in the test’s results when examining ES forecasts and their stressed counterparts.

In assessing the criticisms presented by Wang, et al., (2023) of the Du-Escanciano independence test, some criticisms are more meaningful than others. For example, to criticise the test for being a two-sided test is rather benign, given that it was specifically specified to be

[18] A backtest is said to be indirect if it either: (a) Examines whether the distribution (in full or in part, e.g., the tail) or its properties correspond to the quantities of the true, yet unknown, distribution; or (b) Examines the backtestability of a collection of risk measures, which are elicitable collectively (He, et al., 2022).

a two-sided rather than a one-sided test. The ES backtests presented by Acerbi and Székely (2014, 2017), for example, are one-sided tests, and it can be argued that there is reason behind placing more weight on the risk of under-estimation versus either under- or over-estimation, especially in a regulatory environment. That being said, being able to determine when a model is misspecified, even if the directionality of such misspecification is not known, is still a useful statistical test. In the context of ES breaches and their independence, the misspecification of the model in terms of over-estimation may suggest that breaches are less independent than the model suggests. The relevance of this conclusion ought to be analysed in the context of the statistical accuracy of the number of breaches experienced (as established by other ES backtests), accompanied by a visual inspection of their independence, as is common practice (Acerbi & Székely, 2014). Last, the distributional assumption is not uncommon in the formulation of backtests, usually in the form of assuming that the (unknown) underlying return distribution is as observed. Hence, it seems to be a reasonable assumption for this backtest, too.

Overall, the shift from a sole VaR market risk universe to one of a more holistic view of tail risks is of great importance to the risk management of market risk by banks. The ability to statistically backtest the breaches of a forecasting model to assess its accuracy is vital in the implementation of market risk management. This study examines the models discussed in Section 3.1, in the context of the backtests detailed in the next section, to assess the market risk management of US banks.

3.3. Data and Methodology

This section opens by identifying the index which will serve as the market proxy for the US bank's equities trading desk. The data for the relevant market proxy are discussed, followed by a discussion surrounding the application of the various models discussed earlier to calculate 10-day 99% VaR forecasts, 10-day 97.5% ES forecasts, and their respective stressed counterparts. Then, a discussion of various backtesting techniques to test the statistical adequacy of forecasting models and the statistical independence of the breaches experienced is presented, followed by the cumulative forecasting error measures employed in this study. Finally, the Diebold-Mariano test is introduced, which will assist in determining the relative superiority of the forecasting models employed.

3.3.1. Market Proxy Identification

Before attempting to find an index to serve as a market proxy for the equities trading desk, it is worthwhile to consider what makes an index a good proxy. This study considers two

primary criteria: First, any index chosen to be used as a market proxy in this study must be investable. Second, it must be measurable. These two properties are crucial to allow for the calculating of relevant closing values daily, as these will feed into the market values of the equities desks, facilitating the calculation of daily 10-day 99% VaR forecasts, daily 10-day 97.5% ES forecasts, and their stressed counterparts.

Market indices such as the S&P 500 index for large capitalisation, the S&P MidCap 400 for medium capitalisation, and the Russell 2000 for small capitalisation stocks (Gastineau, 1994), all three of which are capitalisation-weighted indices, are well regarded as equity market proxies. All of these indices satisfy the criteria for what this study refers to as a good market proxy.

When it comes to choosing a suitable index for the equities trading desk, several studies consider the S&P 500 index to be representative of the US equities market (see, for example, Grinold, 1992). The S&P 500 index is used as the market proxy that captures movements in the US equities market as a whole in this study. Hence, the choice here mimics large capitalisation stocks, as those are, most often, the choice of the equities desk of a bank.

Further supporting the choice of the S&P 500 index as a representative of the equities market is the fact that the index represents the equities of the largest companies in the US based on market capitalisation. This suggests that the liquidity of such stocks is high, which is suitable for the analyses undertaken in this study. The readily available market prices for those equities (or, more precisely, index values for the index) allow the frequent calculation of the equity portfolio value. This, in turn, allows the frequent calculation of the profit and loss account of the bank and, therefore, its daily 10-day 99% VaR forecasts, daily 10-day 97.5% ES forecasts, and their stressed counterparts.

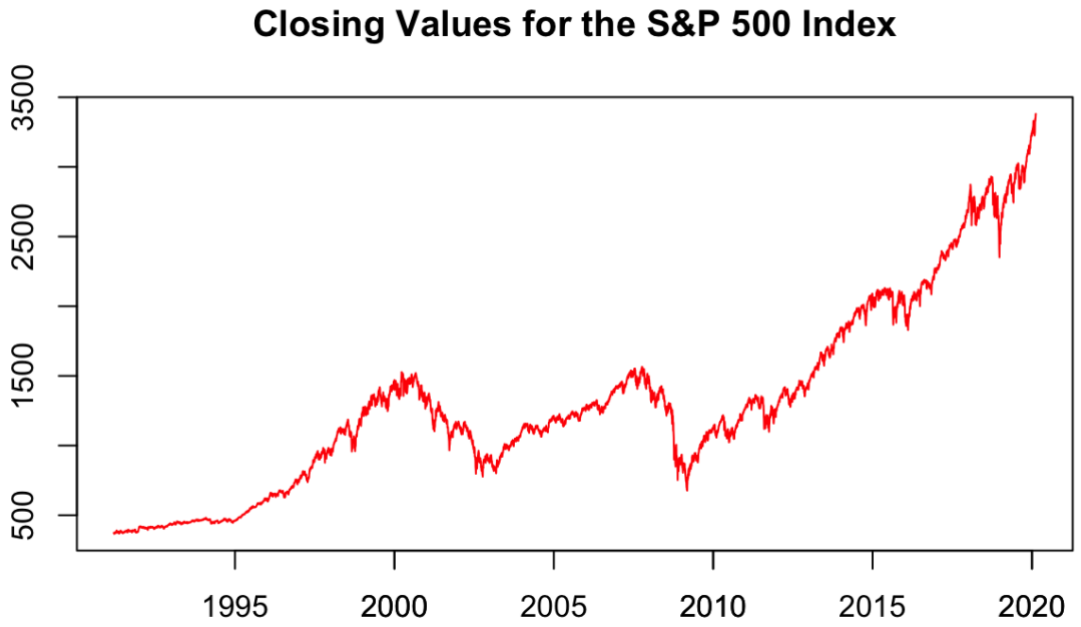
3.3.2. Data

The relevant data capturing the S&P 500 index values were gathered from the Bloomberg database for the period covering March 1991 to February 2020. While the period under examination is quite long, it covers approximately three business cycles (National Bureau of Economic Research, n.d.). While not necessary for the application of the traditional models described in this chapter, the BN model alternative explored in the next chapter requires a large period to facilitate training, together with the implementation of a rolling period methodology, as undertaken in this chapter. Hence, the long period is required for the comparison of the performances of the various traditional models relative to the BN model alternative proposed.

Since the US National Bureau of Economic Research (NBER) only provides the month in which a business cycle starts or ends, but not the specific day, this study uses the 15th of the respective month as the starting date of a cycle and, therefore, it also uses the 14th of the respective month as the ending date of a cycle. The period examined in this study, as mentioned, covers three business cycles over the period 15 March 1991 to 14 February 2020.

The daily closing levels of the S&P 500 index were collected for the entire period, together with the preceding five years, allowing for a calibration period for the various models, as is usual in the literature. The out-of-sample study period covers a total of 7,286 trading days.

Figure 1: Daily Closing Values of the S&P 500 Index



Note: This figure graphically depicts the daily closing values of the S&P 500 index for the period 15 March 1991 to 14 February 2020. The data used were obtained from the Bloomberg database.

The equities trading desk's daily profit and loss account is calculated in this study based on the return earned on the S&P 500 index, which represents the US equities market. The daily returns earned on the S&P 500 index were calculated using the following formula.

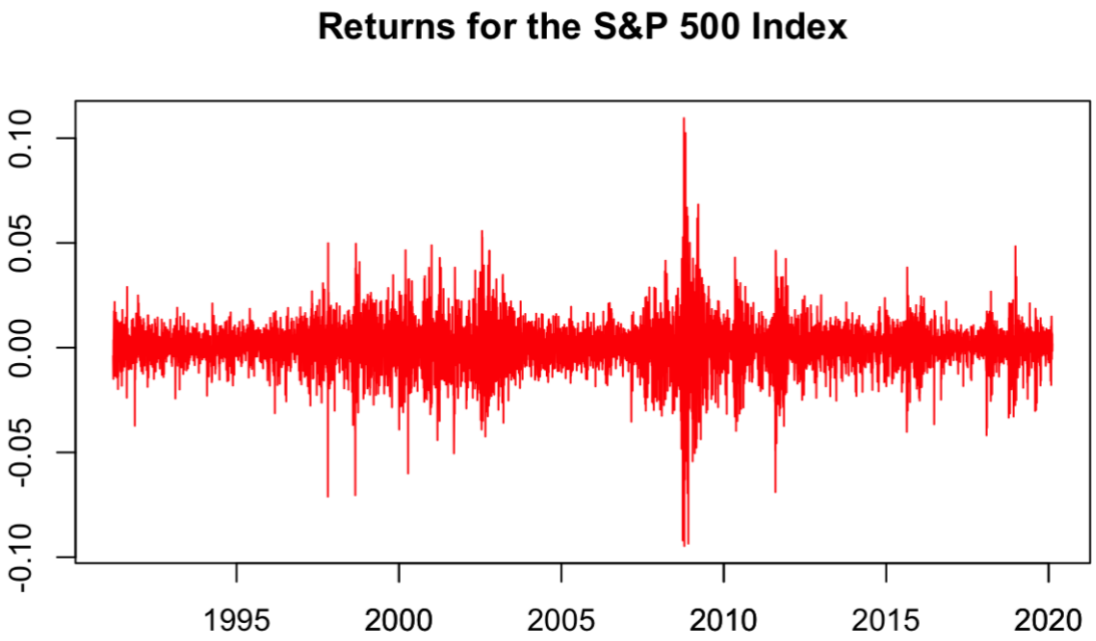
$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \tag{17}$$

where r_t is the daily return earned on the S&P 500 index at time t; P_t is the level of the S&P 500 index at time t; and P_{t-1} is the level of the S&P 500 index on the previous trading day. The time series r_t represents a series of daily (logged) returns on the index. Winsorisation was not applied to the daily return data as a bank's reserves must account for severe market moves

by definition, and winsorisation would have been counterproductive to this ultimate goal. The daily logged returns are displayed in Figure 2, below.

A total of 7,286 daily returns were calculated in the out-of-sample period, with an average daily return of 0.0302% and a standard deviation of 1.1002%. The minimum daily return experienced was -9.4695%, while the maximum daily return experienced was 10.9572%. The returns distribution had a skewness coefficient of -0.2778 and a kurtosis of 12.1012.

Figure 2: Daily Log Returns for the S&P 500 Index



Note: This figure graphically depicts the daily logged returns of the S&P 500 index for the period 15 March 1991 to 14 February 2020. The data used were obtained from the Bloomberg database.

The BCBS, as detailed in the Basel Accords and discussed in Chapter 2, requires banks to calculate 10-day 99% VaR forecasts, 10-day 97.5% ES forecasts, and their respective stressed versions. Hence, this study uses the BCBS's requirements to be consistent with the regulations governing many banks in the US. As per the Basel Accords, the 10-day 99% VaR forecasts are calculated by producing one-day 99% VaR forecasts and scaling them to 10-day 99% VaR forecasts using the square-root-of-time rule. When calculating the 10-day 97.5% ES forecasts, this study also follows the Basel Accords' requirements by calculating the 10-day 97.5% forecasts directly (i.e., without scaling) using a series of overlapping 10-day returns. As for the stressed counterparts of these metrics, the methodology applied is the same as per the non-stressed version, except that a stressed period is used to forecast the stressed metric. The stressed period is the most severe period in the period preceding the return's date, corresponding in length to the non-stressed calibration period, for consistency.

This study uses various models that require a distribution to be fit to the daily returns, i.e., the autoregressive models discussed in Section 3.1.3 and the RiskMetrics model discussed in Section 3.1.4. While this study follows many other studies by using the normal distribution as one of these underlying distributions, the returns of US equities are often found to not be normally distributed (Fama, 1965, being an early study finding this, followed by many other studies). Several authors[19] use the skewed Student's t distribution when modelling tail risk metrics such as VaR and ES as it exhibits skewness, which is believed to fit the data better. This study, therefore, also uses the skewed Student's t distribution when calibrating the various models used to produce risk forecasts due to the assumption that the fatter tails exhibited, coupled with the skew of the distribution, will provide a better fit to the daily profit and loss figures earned by a US bank.

It is worth noting that, while the skewed Student's t distribution is believed to offer a better fit to the profit and loss distribution of the equities trading desk of a US bank, as used in this study, this better fit is assumed to be of a general nature. This means that, while the overall shape, skew, and kurtosis of the profit and loss distribution are believed to be a better fit to the skewed Student's t distribution relative to the normal distribution, there is the possibility of a poorer fit elsewhere in the distribution, with specific reference to the tails. Therefore, it is possible that, while the skewed Student's t distribution is believed to fit the overall profit and loss distribution better, tail risk metrics such as VaR and ES may be less accurate due to the fatter tails exhibited by the skewed Student's t distribution relative to the normal distribution. Therefore, models generating tail risk measures using the normal distribution as the underlying statistical distribution may produce more accurate forecasts relative to those using the skewed Student's t distribution instead. This notion will be revisited in Section 3.4.

The 10-day 99% VaR forecasts and the 10-day 97.5% ES forecasts, and their stressed counterparts, are forecasted using the various models outlined earlier in this chapter, using either the normal distribution or the skewed Student's t distribution, where relevant. These 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts are calculated at the equities desk levels and then backtested, as banks are required to backtest their market risk metrics at a trading desk level (Basel Committee on Banking Supervision, 2019b).

[19] See, for example, Giot and Laurent (2003); Nieto and Ruiz (2016); and Lambert and Laurent (2016).

This offers two main advantages. First, the implementation of the BN model to calculate 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts in the next chapter and the benchmark testing on the equities trading desk level in this chapter allow for the simplification of the model (relative to the aggregation required for several trading desks). This, in turn, makes the model outlined in this study significantly more viable for implementation by banks. Second, the trading desk level analysis allows for the measurement of the suitability and robustness of the BN model to different trading desks. This will allow for conclusions to be made on the appropriateness of the BN model for the equities (and, potentially, other) trading desk(s), in future studies.

Using the daily 10-day 99% VaR forecasts and daily 10-day 97.5% ES forecasts, and their accompanying 10-day 99% SVaR forecasts and 10-day 97.5% SES forecasts, where the latter are calculated using the worst loss preceding period, a breach is recorded when the daily loss made on the S&P 500 index exceeds the daily market risk forecast. The daily forecast was calculated as the day's (S)VaR and (S)ES forecasts, calibrated using the various models outlined above and either the normal distribution or the skewed Student's t distribution as the underlying statistical distribution.

While backtesting is a requirement and a standard procedure when evaluating the statistical accuracy and suitability of the models used to produce market risk metric forecasts, it is important to ensure that even models that produce very few breaches are not inefficient from an excess reserve perspective. A model that spits out a capital requirement of $1 trillion would, indeed, lead to no breaches. However, such a model would yield excessive capital reserves relative to the actual risk borne by the bank. Hence, there needs to be a separate methodology to evaluate the efficiency of the models tested with respect to their level of capital projected (as determined by the risk metric forecast) relative to the actual profit and loss experienced (as determined by the actual return of the equities trading desk, i.e., the return on the S&P 500 index). This study employs four forecasting error measures to capture the potential inefficiency of the various models' forecasts relative to the daily returns observed during the out-of-sample test period. These are discussed further in Section 3.3.4.

These forecasting error measures are even more relevant where a model produces very few breaches (if any) and fails the various backtests. This combination often implies that the model employed produces excessive capital held (yielding few breaches, if any) and, therefore, the model is statistically inaccurate. Hence, the process of calculating the values of the various

forecasting error measures employed in this study is carried out for all models and all forecasts produced in this study.

Last, Section 3.3.4 also introduces the Diebold-Mariano test. This test is a statistical test for comparing the forecasting abilities of various models. This test is used to statistically test the superior forecasting ability of one model over another, providing a statistical test to further build on the results of the various forecasting error measures.

3.3.3. Backtesting

Using each of the models outlined above, the equities trading desk's profit and loss account is forecasted. From these, 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts are calculated for the 7,286-trading-day out-of-sample test period under investigation. The VaR forecasts calculated correspond to the regulatory required VaR forecast, i.e., one-day 99% VaR forecasts, which are then scaled to yield 10-day 99% VaR forecasts using the square-root-of-time rule. The ES forecasts are calculated at their regulated levels, i.e., 10-day 97.5% forecasts are calculated directly, rather than scaled one-day forecasts.

To evaluate the model's performance, the model's forecasts, forecasted over the period, are used and compared to actual profit or loss amounts using the equities trading desk data for the out-of-sample study period, as captured by the profit and loss made on the S&P 500 index. The number of breaches is then recorded and compared to the number of breaches achieved using commonplace VaR and ES forecasting models, as detailed in Section 3.1. Since the historical simulation model is the most popular model employed by banks to forecast VaR, it serves as an appropriate benchmark. This, together with the more evolved backtesting techniques discussed below, are used for comparison purposes, to establish the relative performances of the various models for the equities trading desk.

This study explores several backtesting models. First, the BCBS's traffic light test is explored, as the system is the regulatory framework to evaluate VaR and ES forecasting models. For VaR, Kupiec's proportion of failure coverage test and Christoffersen's test for independence are explored and employed. For ES, the conditional, the unconditional, and the minimally biased backtests are explored and employed, while the independence of breaches is backtested using the Du-Escanciano independence test. Finally, the efficiencies of the models are evaluated, where efficiency is defined to be the minimisation of forecasting errors (which would translate to excess reserves held) relative to the true profit and loss made by the equities trading desk. For example, a model may achieve relatively few breaches (if any) if it dictates

that reserves held cover decuple the VaR forecast obtained, although such a level of reserves would be excessive from the bank's (and shareholders') viewpoint.

As mentioned, a drawback of ES is the fact that it is not elicitable, which implies that it is not easily ranked and, therefore, backtested. Contrary to the implications of the lack of elicitability of ES, Acerbi and Székely (2014), building on the work presented by Kerkhof and Melenberg (2004), show that the backtesting of ES is not much more involved than that of VaR. Acerbi and Székely present several backtests for ES, some of which are discussed in this study. First, a definition for a backtesting function of ES, as formulated by Acerbi and Székely (2017), follows.

$$\bar{Z}(X) = \frac{1}{N}\sum_{t=1}^{N} Z\left(\widehat{ES}_{ht,\alpha}, \widehat{VaR}_{ht,\alpha}, X_t\right) \tag{18}$$

where X is a vector of profit and loss amounts; $\widehat{ES}_{ht,\alpha}$ is a vector of expected shortfall forecasts; and $\widehat{VaR}_{ht,\alpha}$ is a vector of value at risk forecasts.

Acerbi and Székely (2014) present two primary non-parametric backtesting techniques for ES, which are explored in this study. The conditional backtesting technique is explored, followed by the unconditional backtesting technique for ES. Finally, Acerbi and Székely (2017) present the minimally biased backtest, which is employed in this study.

While breaches may be experienced, whereby the daily loss incurred on the S&P 500 index exceeds the 10-day 97.5% ES forecast, their independence must also be tested, similarly to the independence testing performed for 10-day 99% VaR forecasts. The Du-Escanciano independence test for ES is employed in this study to complement the other backtests performed on the ES forecasts.

3.3.3.1. The Traffic Light Test

The process of comparing the daily 10-day 99% VaR forecast that is obtained by any model to that day's profit or loss amount as recorded in the bank's profit and loss account is known as backtesting. The BCBS discusses backtesting on both the trading desk level and the bank-wide level. In either case, the BCBS proposes that a bank monitors two accounts. First, the bank must monitor its actual profit and loss account, recording actual intraday transactions. These amounts, which the bank actually experiences, are used to compare actual trading outcomes. Second, the bank must also monitor its hypothetical profit and loss account for any given day, being the hypothetical state of the account at the end of that day, should no changes

in position sizes have taken place, i.e., no trades were performed, but the value of the positions may have changed. Each profit and loss account, i.e., the hypothetical and the actual, is then tested for breaches against the daily 10-day 99% VaR forecasts obtained via the bank's internal model. The BCBS then requires the bank to recognise a number of breaches equal to the higher number of breaches experienced by either of the accounts (Basel Committee on Banking Supervision, 2019b). This study adopts the latter account, i.e., the hypothetical profit and loss account, as it assumes that only the changes in the value of the market proxy, the S&P 500 index, are the drivers of market risk of the equities trading desk.

Table 3: The Basel Committee on Banking Supervision's Table of Breaches and Likelihoods of Statistical Errors for the Basel II Accord

	Model is accurate		**Model is inaccurate: Possible alternative levels of coverage**							
	Coverage = 99%		Coverage = 98%		Coverage = 97%		Coverage = 96%		Coverage = 95%	
	Exact	Type I	Exact	Type II	Exact	Type II	Exact	Type II	Exact	Type II
0	8.1%	100.0%	0.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
1	20.5%	91.9%	3.3%	0.6%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%
2	25.7%	71.4%	8.3%	3.9%	1.5%	0.4%	0.2%	0.0%	0.0%	0.0%
3	21.5%	45.7%	14.0%	12.2%	3.8%	1.9%	0.7%	0.2%	0.1%	0.0%
4	13.4%	24.2%	17.7%	26.2%	7.2%	5.7%	1.8%	0.9%	0.3%	0.1%
5	6.7%	10.8%	17.7%	43.9%	10.9%	12.8%	3.6%	2.7%	0.9%	0.5%
6	2.7%	4.1%	14.8%	61.6%	13.8%	23.7%	6.2%	6.3%	1.8%	1.3%
7	1.0%	1.4%	10.5%	76.4%	14.9%	37.5%	9.0%	12.5%	3.4%	3.1%
8	0.3%	0.4%	6.5%	86.9%	14.0%	52.4%	11.3%	21.5%	5.4%	6.5%
9	0.1%	0.1%	3.6%	93.4%	11.6%	66.3%	12.7%	32.8%	7.6%	11.9%
10	0.0%	0.0%	1.8%	97.0%	8.6%	77.9%	12.8%	45.5%	9.6%	19.5%
11	0.0%	0.0%	0.8%	98.7%	5.8%	86.6%	11.6%	58.3%	11.1%	29.1%
12	0.0%	0.0%	0.3%	99.5%	3.6%	92.4%	9.6%	69.9%	11.6%	40.2%
13	0.0%	0.0%	0.1%	99.8%	2.0%	96.0%	7.3%	79.5%	11.2%	51.8%
14	0.0%	0.0%	0.0%	99.9%	1.1%	98.0%	5.2%	86.9%	10.0%	62.9%
15	0.0%	0.0%	0.0%	100.0%	0.5%	99.1%	3.4%	92.1%	8.2%	72.9%

Note: This table depicts the probability of obtaining a specific number of breaches of a value at risk (VaR) forecast (given a level of confidence [coverage]) from a binomial distribution in the backtesting process as stipulated by the Basel Committee on Banking Supervision, together with the likelihood of obtaining one of two errors: A Type I error, which depicts the case in

which an accurate model is deemed inaccurate due to the result of the backtesting process, and a Type II error, which depicts the case in which an inaccurate model is not recognised due to the result of the backtesting process (Basel Committee on Banking Supervision, 2019b).

In conjunction with the traffic light test to backtest model breaches (see Table 1), the BCBS acknowledges that the statistical tests involved in undertaking backtesting may require some form of statistical consideration. The BCBS recognises two types of statistical errors. These are:

i. A Type I statistical error, where an accurate model may be rendered inaccurate due to its backtesting result; and
ii. A Type II statistical error, where an inaccurate model may fail to be rendered inaccurate due to its backtesting result.

The BCBS further provides Table 3, above, as a method to measure the likelihood of obtaining either type of statistical error, given the confidence level of the VaR forecast, as well as the exact number of breaches expected, based on a binomial distribution with a 250-trading-day period of independent observations. In the table, the column labelled 'Exact' depicts the exact number of breaches expected by the model, given the degree of confidence provided, and a sample of 250 independent observations. The column labelled 'Type I' represents the cumulative probability of falsely rejecting an accurate model, while the column labelled 'Type II' represents the cumulative probability of failing to reject an inaccurate model (Basel Committee on Banking Supervision, 2019b).

For example, if a bank sets a limit of seven breaches (so that the bank's model falls in the middle of the 'Yellow zone' range – see Table 1), then, for a model employing the desired confidence level (99%), Table 3 reports that the likelihood that an accurate model will be rejected is only 1.4%. Further, the bank can calculate that the probability of observing exactly 7 breaches in an accurate model is 1.0%. As the number of breaches that the bank is willing to absorb increases, the likelihood that the bank rejects an accurate model, i.e., performing a Type I error, is minimised. Conversely, as the number of breaches that the bank is willing to absorb decreases, the probability that the bank accepts an inaccurate model is maximised.

Table 3 further depicts some of the limitations involved in the backtesting process. The BCBS points out that there is no specific limit on the number of breaches which a bank may choose to tolerate so that the likelihoods of both types of statistical error are minimised (Basel Committee on Banking Supervision, 2019b).

Table 3 can now shed further light on the traffic light test depicted in Table 1. The 'Green zone' depicted in Table 1 represents a low likelihood of performing a Type II error, while the 'Red zone' depicted represents a low likelihood of performing a Type I error. A compromise zone, the 'Yellow zone', is one in which the likelihood of performing either statistical error is consistent with the number of breaches observed in the backtesting process.

As an alternative to the number of breaches application of the BCBS's traffic light test, which depends on the number of trading days used in the investigation, the BCBS also proposes a statistical approach to determining the zone classification used in the traffic light test. A model is determined to be in the 'Green zone' if the probability of the observed number of breaches, modelled using a binomial random variable at the desired confidence level, is less than or equal to 0.95. Alternatively, a model is determined to be in the 'Yellow zone' if the probability of the observed number of breaches, similarly modelled, is above 0.95 and less than or equal to 0.9999. Finally, a model is determined to be in the 'Red zone' if the probability of the observed number of breaches, similarly modelled, is above 0.9999 and less than or equal to 1.

3.3.3.2. Kupiec's Proportion of Failure Coverage Test of Value at Risk

While the BCBS's traffic light test is used by banks due to regulatory enforcement, it is used to determine the regulatory penalty variable, k, rather than to assess the internal model's predictive ability and suitability. However, there exist several backtesting techniques which are commonly used and are more intriguing from a theoretical standpoint. One such test is Kupiec's proportion of failure coverage test.

Under the proportion of failure test, just like under the BCBS's traffic light test, observations are assumed to be of a random variable X that is binomially distributed with $N + 1$ trading days and a confidence interval q, often being 0.99. Moreover, the null hypothesis of the test assumes that $q = \hat{q}$, where $\hat{q}$ is the observed likelihood of breaches in the model, i.e., $\hat{q} = \frac{x}{N}$. Hence, the proportion of failure test is a two-sided test. The proportion of failure test employs the binominal distribution assumption to calculate a likelihood ratio, Λ, as set out in Equation (19) below (Kupiec, 1995).

$$\Lambda = \frac{q^{N-x}(1-q)^{x}}{(1-\hat{q})^{N-x}(\hat{q})^{x}} \tag{19}$$

This ratio, however, is difficult to use in the process of inferring probabilities. Hence, a more common and more applicable version of the ratio is stated in Equation (20), below.

$$LR_{KPOF} = -2\log(\Lambda) = -2\log\left(\left(\frac{qN}{N-x}\right)^{N-x}\left(\frac{N(1-q)}{x}\right)^{x}\right)$$
$$= 2\log\left(\left(\frac{N-x}{qN}\right)^{N-x}\left(\frac{x}{N(1-q)}\right)^{x}\right) \quad (20)$$

The likelihood ratio test is a ratio which compares the maximum obtainable probabilities under the two available hypotheses, namely the null and the alternative hypotheses. The numerator in Equation (19) captures the maximum probability as calculated by the null hypothesis, while the denominator captures the maximum probability as calculated by the alternative hypothesis. A conclusion can then be reached based on the value of the ratio, as expressed by the test statistic in Equation (20), whereby a smaller ratio in Equation (19) implies a larger test statistic in Equation (20).

The likelihood ratio in Equation (20) is compared to the value of a χ^2 with 1 degree of freedom (Lehmann & Romano, 2005), i.e., 6.63490 for a test at the 99% confidence level. Hence, should Equation (20)'s LR_{KPOF} exceed 6.63490, then Kupiec's proportion of failure coverage test's null hypothesis is rejected, and it can concluded that the model in question is inaccurate or incorrectly specified at the stated level of confidence (Jorion, 2001).

Alternatively, the test may be carried out by calculating a confidence interval for the number of breaches expected for the model. Should the number of breaches observed not fall within the confidence interval, the null hypothesis ought to be rejected. Hence, should the number of breaches experienced fall below or above the confidence interval, it can be concluded that the model employed is inaccurate. This is the approach employed in this study, as detailed in the results section (see Section 3.4).

3.3.3.3. Christoffersen's Test for Independence of Value at Risk

The final step in backtesting a model producing VaR forecasts is to test the independence of breaches (or, alternatively, the clustering of such breaches), since a collection of successive breaches may put a bank under regulatory and market pressures (Christoffersen & Pelletier, 2004). Noting the heteroscedastic nature of financial data (and, by extension, market values), Peter Christoffersen (1998) introduced a test for the clustering of breaches, known as Christoffersen's test for independence.

Under Christoffersen's test, an indicator variable, $\mathbf{1}_{\{X_t \leq x^{(\alpha)}\}}$, takes on the value of one if a breach occurs or zero otherwise. Carrying through the notation used for Kupiec's proportion

of failure test, this indicator variable $\mathbf{1}_{\{X_t \leq x^{(\alpha)}\}}$ is distributed as a Bernoulli random variable with some probability of a breach taking place, q. Christoffersen's test for independence then tests the null hypothesis that the theoretical probability of a breach, q, is equal to the observed probability of a breach, $\hat{q}$.

By defining a transition probability $q_{ij} \coloneqq \Pr\left[\mathbf{1}_{\{X_t \leq x^{(\alpha)}\}} = j \middle| \mathbf{1}_{\{X_{t-1} \leq x^{(\alpha)}\}} = i\right] \; \forall \, i, j \in \{0,1\}, \forall \, t$, Christoffersen (1998) then defines a 2×2 transition matrix Π with its entries being the transition probabilities. By letting the sum of entries in the rows of the transition matrix add up to 1, the following transition matrix is derived.

$$\Pi = \begin{bmatrix} 1 - q_{01} & q_{01} \\ 1 - q_{11} & q_{11} \end{bmatrix} = \begin{bmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{bmatrix} \tag{21}$$

The observed transition matrix, $\widehat{\Pi}$, corresponding to Π can be captured by observing the numbers of consecutive forecasts (or, equivalently, transitions), corresponding to (i) a non-breach following a non-breach, denoted by n_{00}; (ii) a breach following a non-breach, denoted by n_{01}; (iii) a non-breach following a breach, denoted by n_{10}; and (iv) a breach following a breach, denoted by n_{11}. Hence, the observed transition probabilities $\hat{q}_{ij}$ can be calculated using the n_{ij}s to populate the observed transition matrix $\widehat{\Pi}$ as follows.

$$\widehat{\Pi} = \begin{bmatrix} \hat{q}_{00} & \hat{q}_{01} \\ \hat{q}_{10} & \hat{q}_{11} \end{bmatrix} = \begin{bmatrix} \dfrac{n_{00}}{n_{00} + n_{01}} & \dfrac{n_{01}}{n_{00} + n_{01}} \\ \dfrac{n_{10}}{n_{10} + n_{11}} & \dfrac{n_{11}}{n_{10} + n_{11}} \end{bmatrix} \tag{22}$$

where the transitions n_{ij} and observed transition probabilities $\hat{q}_{ij} \; \forall \, i, j \in \{0,1\}$, are as defined above[20].

Christoffersen (1998) then tests for the independence of the sequence of breaches by assuming that the probability of a breach following a non-breach is equal to that of a breach following a breach, i.e., $q_{01} = q_{11}$. Hence, this test is a two-sided test. By labelling such a probability q_2, Christoffersen derives the maximum likelihood estimate of this probability as follows.

$$\hat{q}_2 = \frac{n_{01} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}} \tag{23}$$

[20] Christoffersen (1998) derives this result by maximising the corresponding log-likelihood function.

where $\hat{q}_2$ is the observed value of q_2; and the transitions $n_{ij}, \forall\, i,j \in \{0,1\}$, are as defined above.

The test statistic for the independence of breaches, LR_{CIND}, according to Christoffersen (1998), is then as follows.

$$LR_{CIND} = -2\log\left(\frac{(1-\hat{q}_2)^{n_{00}+n_{10}} \cdot \hat{q}_2^{n_{01}+n_{11}}}{\hat{q}_{00}^{n_{00}} \cdot \hat{q}_{01}^{n_{01}} \cdot \hat{q}_{10}^{n_{10}} \cdot \hat{q}_{11}^{n_{11}}}\right) \tag{24}$$

where $\hat{q}_2$ is the observed value of q_2; and the transitions n_{ij} and observed transition probabilities $\hat{q}_{ij}, \forall\, i,j \in \{0,1\}$, are as defined above. Asymptotically, $LR_{CIND} \sim \chi_1^2$ (Christoffersen, 1998).

The likelihood ratio in Equation (24), similarly to that in Kupiec's proportion of failure test, is compared to the value of a χ^2 with 1 degree of freedom, i.e., 6.63490 for a test at the 99% confidence level. Hence, should Equation (24)'s LR_{CIND} exceed 6.63490, then Christoffersen's independence test's null hypothesis is rejected, and it can be concluded that the experienced frequency of breaches does not equal the assumed frequency at the desired confidence level, i.e., the model is incorrectly specified.

The VaR backtesting techniques described above are carried out in this chapter, using 10-day 99% VaR forecasts and 10-day 99% SVaR forecasts as obtained via the various models detailed in this study, as discussed in Section 3.1. The superior model will be the one which, for the equities trading desk in question, exhibits a number and independence of breaches which corresponds to a correctly specified forecasting model, coupled with minimising the forecasting error measures detailed in Section 3.3.4.

3.3.3.4. The Conditional Backtest of Expected Shortfall

The conditional backtest, a backtesting technique presented by Acerbi and Székely (2014), rests on the assumption that ES forecasts are produced and tested after their corresponding VaR forecasts have been produced and tested. Using the definition of ES as a starting point, as stated in Equation (10), the following expectation can be derived.

$$E\left[\frac{X}{ES_{ht,\alpha}} + 1 \middle| X + VaR_{ht,\alpha} < 0\right] = 0 \tag{25}$$

The null hypothesis states that the ES forecasts are equal to the true ES values given that VaR exhibits a breach, i.e., $ES_{ht,\alpha} = \widehat{ES}_{ht,\alpha}$ (Acerbi & Székely, 2014). Hence, the test statistic, using the notation of Equation (18), is as follows.

$$\bar{Z}_{CB}(X) = \frac{1}{N_{Breaches}} \sum_{t=1}^{N} \frac{X_t \mathbf{1}_{\{X \leq x^{(\alpha)}\}}}{\widehat{ES}_{ht,\alpha}} + 1 \tag{26}$$

where X and $\widehat{ES}_{ht,\alpha}$ are as defined for Equation (18); $\mathbf{1}_{\{X \leq x^{(\alpha)}\}}$ is as defined for Equation (9); and $N_{Breaches} = \sum_{t=1}^{N} \mathbf{1}_{\{X \leq x^{(\alpha)}\}}$ (Acerbi & Székely, 2014), i.e., $N_{Breaches}$ is the number of VaR breaches observed.

The alternative hypothesis for this test states that the true ES values experienced exceed those forecasted, i.e., the model underestimates ES. Acerbi and Székely (2014) note that, since the test assumes that VaR breaches are correctly captured, the conditional expectation of $\bar{Z}_{CB}$ is zero under the null hypothesis, where this expectation is conditional on the total number of VaR breaches being greater than zero, i.e., $E[\bar{Z}_{CB} | N_{Breaches} > 0] = 0$. On the other hand, under the alternative hypothesis, this expectation is negative (Acerbi & Székely, 2014). These expected values can be examined to evaluate the performance of the ES forecasting model as (any) negative values would indicate that the null hypothesis should be rejected for the model at hand, at the chosen confidence level.

3.3.3.5. The Unconditional Backtest of Expected Shortfall

Acerbi and Székely (2014) also present an unconditional backtesting technique, which uses the number of ES breaches observed as opposed to the number of VaR breaches observed. The test statistic developed, $\bar{Z}_{UB}(X)$, has the same null hypothesis as the conditional test, i.e., the ES forecasts are equal to the true ES values, while the alternative hypothesis states that the true ES values exceed the ES forecasts (Acerbi & Székely, 2014), i.e., the model in question underestimates the ES values for the given confidence level. Following the formulation of Equation (18), the test statistic for the unconditional ES backtest is as follows.

$$\bar{Z}_{UB}(X) = \frac{1}{N} \sum_{t=1}^{N} \frac{X_t \mathbf{1}_{\{X \leq x^{(\alpha)}\}}}{\alpha \widehat{ES}_{ht,\alpha}} + 1 \tag{27}$$

where X and $\widehat{ES}_{ht,\alpha}$ are, again, as defined for Equation (18); and $\mathbf{1}_{\{X \leq x^{(\alpha)}\}}$ is, again, as defined for Equation (9) (Acerbi & Székely, 2014). Note that while the test is unconditional in name, it can only be applied given at least one VaR breach, as captured by the indicator variable $\mathbf{1}_{\{X \leq x^{(\alpha)}\}}$ in Equation (27).

As with the conditional backtest, the expectation of the test statistic is zero under the null hypothesis and negative under the alternative hypothesis (Acerbi & Székely, 2014). Hence, any

negative values may indicate that the null hypothesis should be rejected for the model at hand, at the chosen confidence level, indicating that the ES forecasting model is incorrectly specified, as it underestimates the true ES figures experienced.

3.3.3.6. The Minimally Biased Backtest of Expected Shortfall

Acerbi and Székely (2017) present a minimally biased backtest, which is also be employed in this study. The null hypothesis is, once again, that the ES forecasts are equal to the true underlying ES values. The alternative hypothesis, on the other hand, states that (at least some of) the ES forecasts underestimate the true ES figures. The test statistic, $\bar{Z}_{MB}$, is as follows.

$$\bar{Z}_{MB}(X) = \frac{1}{N}\sum_{t=1}^{N} \widehat{ES}_{ht,\alpha} - \widehat{VaR}_{ht,\alpha} + \frac{\left(\widehat{VaR}_{ht,\alpha} + \widehat{ES}_{ht,\alpha}\right)\mathbf{1}_{\{X \leq x^{(\alpha)}\}}}{\alpha} \qquad (28)$$

where X and $\widehat{ES}_{ht,\alpha}$ are again as defined for Equation (18); $\mathbf{1}_{\{X \leq x^{(\alpha)}\}}$ is again as defined for Equation (9), and $\widehat{VaR}_{ht,\alpha}$ is as defined for Equation (1) (Acerbi & Székely, 2017). Note that this backtest is not conditioned on the existence of at least one VaR breach, as the previous two are.

Once again, under the null hypothesis, the expectation of the test statistic is zero, while it is negative under the alternative hypothesis (Acerbi & Székely, 2017). Hence, a negative expected value of the test statistic may indicate that the ES forecasts underestimate the true ones and the null hypothesis should be rejected.

This backtest is preferred to the unconditional backtest (formulated in Equation (27)) as the unconditional backtest displays a significant linear relationship to VaR forecasts (Acerbi & Székely, 2017), meaning that the incorrect calibration of the model to produce VaR forecasts may lead to an increased probability of producing either a Type I error or a Type II error.

3.3.3.7. The Du-Escanciano Independence Backtest of Expected Shortfall

The Du-Escanciano independence test for ES is employed in this study to complement any visual inspections of clustered ES breaches. It is the ES equivalent to Christoffersen's independence test for VaR forecasts, as discussed in Section 3.3.3.3. Du and Escanciano (2017) develop a Portmanteau Box-Pierce conditional test for the independence of ES breaches, i.e., a statistical test whose primary objective is to determine whether ES breaches cluster together.

Consider a cumulative breaches function $H_t(\alpha)$, which is the total count at time t of VaR breaches at the confidence level α, i.e., $H_t(\alpha) = \frac{1}{\alpha}\int_0^{\alpha} \mathbf{1}_{\{X_t \leq x^{(\alpha)}\}}\, dx$. Du and Escanciano (2017)

deduce that the expected value of the cumulative breaches function $H_t(\alpha)$ is equal to half the confidence level, i.e., $E[H_t(\alpha)] = \alpha/2$. Moreover, the authors show that the series $\{H_t(\alpha) - \alpha/2\}_{i=1}^{\infty}$ is a martingale difference sequence (Du & Escanciano, 2017), i.e., the expected value of the series, given the filtration system $\mathcal{F}_t$, is zero.

Defining u_t to be the conditional cumulative distribution function of the profit and loss account of a bank at time t evaluated at the level u, conditional on the filtration system $\mathcal{F}_t$, Du and Escanciano (2017) show that $H_t(\alpha) = \alpha^{-1}(\alpha - u_t)\mathbf{1}_{\{u_t \leq \alpha\}}$.

Further to the discussion above, consider the conditional parametric distribution of the profit and loss account, conditional on some parameter $\Theta \in \mathbb{R}^p$, that unknown parameter being θ_0. Du and Escanciano (2017) then show that the function of associated cumulative breaches, $H_t(\alpha, \theta_0)$, is as follows.

$$H_t(\alpha, \theta_0) = \frac{1}{\alpha}(\alpha - u_t(\theta_0))\mathbf{1}_{\{u_t(\theta_0) \leq \alpha\}} \tag{29}$$

where $u_t(\theta_0)$ is the conditional cumulative distribution function of the profit and loss account of a bank at time t, evaluated at the level u, conditional on the parameter θ_0 and the filtration system $\mathcal{F}_t$, the set of which is termed by Du and Escanciano as generalised errors (Du & Escanciano, 2017).

The conditional test presented in Du and Escanciano (2017) is based on such cumulative breaches, as captured in Equation (29). The conditional test essentially tests whether the terms of the series $\{H_t(\alpha) - \alpha/2\}_{i=1}^{\infty}$ are uncorrelated. Consider the null hypothesis stating that, given the filtration system $\mathcal{F}_t$, the expected value of each term of this series is zero, i.e., $E[H_t(\alpha) - \alpha/2|\mathcal{F}_t] = 0$.

Consider the autocovariance function and autocorrelation function of $H_t(\alpha)$ for lag j as $\gamma_{N,j}$ and $\rho_{N,j}$, where $\rho_j = \gamma_{N,j}/\gamma_{N,0}$ as usual. The Du-Escanciano conditional test statistic is, then, as follows.

$$\bar{Z}_{DE}(X, n) = N \sum_{j=1}^{n} \hat{\rho}_{T,j}^2 \tag{30}$$

where N is the number of out-of-sample trading days; n is the highest lag of autocorrelations, or 1 as default; and $\hat{\rho}_{N,j}$ is the estimate of $\rho_{N,j}$ as defined above (Du & Escanciano, 2017). Du

and Escanciano (2017) show that this test statistic is asymptotically distributed as a chi-squared random variable with n degree of freedom.

3.3.4. Forecasting Error Measures

Last, this study uses four forecasting error measures to aid in assessing the efficiency of the forecasts produced by the various models discussed in this study. Specifically, this study employs the mean absolute error (MAE, see Equation (31)), the root mean square error (RMSE, see Equation (32)), the mean absolute percentage error (MAPE, see Equation (33)), and the median absolute percentage error (MdAPE, see Equation (34)) as measures to capture the cumulative forecasting errors of each model. In the case of zero returns, the MAPE measure will produce invalid results. In such cases, the symmetric MAPE (SMAPE) is used instead of MAPE. As all models are wrong, it is reasonable to assume that their forecasts will differ from actual experience. Hence, these measures are used to capture such differences. The measures are as follows.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{31}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{32}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{33}$$

$$MdAPE = median\left(\left|\frac{y_i - \hat{y}_i}{y_i}\right|\right) \tag{34}$$

$$SMAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{\frac{1}{2}\times(|y_i| + |\hat{y}_i|)} \tag{35}$$

where N is the number of forecasts; $\hat{y}_i$ is the forecasted value at time i for $i \in \{1,2,\dots,N\}$; and y_i is the observed value at time i for $i \in \{1,2,\dots,N\}$. In this study, the y_is represent the returns observed, while the $\hat{y}_i$s represent the market risk metric forecasts.

Regardless of which of the above measures is used, a lower measure represents a lower difference between predicted (or forecasted) values and actual values when comparing values out-of-sample. Hence, a lower figure for any given model and a given measure is superior to a higher figure for that same model and measure.

Each of the four error measures has its respective advantages and disadvantages, and it is important to keep in mind that each one is useful in its own context. Moreover, all four may work in unison to provide information about the model used to produce forecasts. All four of the forecasting error measures provide standardised tools to compare the forecasts achieved by the various models used in this study. That being said, the measures are not without some disadvantages. The RMSE, for example, is more sensitive to large outliers relative to the other measures due to its squaring function. The other models, on the other hand, are still susceptible to the effects of outliers, but to a lesser extent, due to the weight assigned to each difference. For example, the MAE assigns equal weights to each difference between the forecast and the observed value. The four measures used offer measures that are in the same units as the original data (i.e., percentage returns). This is true for the RMSE too, due to its use of the square-root function (as can be seen in Equation (32)). The MAE, the MAPE, and the MdAPE do not take into account the direction of the errors (as the absolute value functions override the directions of the errors in each of the three measures). Last, the MdAPE is a median measure of error, which is also less sensitive to extreme outliers at either end of the difference distribution.

Since four distinct forecasting error measures are used in this study, it is possible for different models to prove more accurate (based on the forecasting error measures discussed above) using different forecasting error measures. Hence, a statistical test is to be employed to determine the relative superior predictive ability of the models (relative to each other). This test is the Diebold-Mariano test (Diebold & Mariano, 1995).

The Diebold-Mariano statistic's null hypothesis assumes no differences in the forecasting accuracy of the two models compared. The Diebold-Mariano statistic accounts for the relative rank of the two models compared. Hence, it is important to keep in mind that if the statistic indicates that there is sufficient evidence to reject the null hypothesis, then the first model considered when calculating the loss differential (see below) is less accurate than the second model considered (Diebold & Mariano, 1995).

The first step in this test is to calculate the loss differential, d, its average, $\bar{d}$, and the autocovariance of said loss differential at lag k, γ_k, defined as follows.

$$\gamma_k = \frac{1}{N} \sum_{i=k+1}^{N} \left(d_i - \bar{d}\right)\left(d_{i-k} - \bar{d}\right) \tag{36}$$

where N is number of (pairs of) forecasts, as before.

Using the autocovariance function detailed above to obtain an estimator of variance of the loss function, the Diebold-Mariano statistic, DM, is then as follows.

$$DM = \frac{\bar{d}}{\tilde{\sigma}_{\bar{d}}} \tag{37}$$

where DM is the Diebold-Mariano statistic; and $\tilde{\sigma}_{\bar{d}}$ is a (consistent) estimator of the average loss differential measure's standard deviation (Diebold & Mariano, 1995). Note that $DM \sim N(0,1)$ (asymptotically) using the central limit theorem.

The Diebold-Mariano statistic only contains information regarding the relative performance of the models compared (Diebold, 2015). It does not provide any information on whether any model examined in isolation is good or bad when it comes to assessing its relative forecasting abilities. Hence, the conclusions derived in any study using the Diebold-Mariano statistic must be used in conjunction with the cumulative error measures, such as those described in Equations (31) to (35), above.

3.4. Results

This section discusses the results using the data and methodology discussed in the preceding sections, and analyses of the results are provided. Specifically, this section examines the performance of each of the models detailed in Section 3.1 when producing 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts, and their stressed versions, using the normal distribution and the skewed Student's t distribution as the underlying statistical distributions. This chapter assesses which model and distribution combination yielded the most accurate market risk forecasts when setting reserve capital for a US bank. Hence, using each of the underlying distributions, each model's forecasts were backtested using the backtests detailed in Section 3.3.3. The forecasts were then further evaluated on their efficiency in forecasting by assessing their out-of-sample performances relative to actual profit and loss amounts experienced by the equities trading desk of the bank. The performance of these forecasting error measures was further statistically tested using the Diebold-Mariano test, as detailed in Section 3.3.4.

This study considers a total of ten models to calculate each market risk metric. The historical simulation model and the delta-normal model were used, with their respective assumptions. In addition, the normal distribution and the skewed Student's t distribution were each used as the underlying distribution for the ARCH model, the GARCH model, the EGARCH model, and RiskMetrics model. In total, this study evaluated the performance of 10-day 99% VaR

forecasts, 10-day 97.5% ES forecasts, 10-day 99% SVaR forecasts, and 10-day 97.5% SES forecasts using two distributions, a total of 40 models (ten models across four market risk metrics). Moreover, the performances of the 40 models were then evaluated using the MAE, the RMSE, the MAPE, and the MdAPE, yielding values for 160 different forecasting error measures.

The rest of this section discusses each risk metric individually, exploring and analysing the results of the different models employed to calculate said metric in its context. The number of breaches experienced was calculated, and said breaches were backtested using the relevant backtesting techniques. Forecasting error measures were then employed to calculate the efficiency of the models used to forecast the risk metric, before using the Diebold-Mariano test to statistically evaluate the different models' forecasting ability and relative superiority.

3.4.1. Value at Risk

This section discusses and analyses the results of the models employed in this study to forecast 10-day 99% VaR forecasts for a US bank using the S&P 500 index as the underlying, over the period 15 March 1991 to 14 February 2020. First, the number of breaches experienced for each of the various models is presented and discussed, followed by an analysis of the backtesting results achieved using the BCBS's traffic light test, Kupiec's proportion of failure test, and Christoffersen's test for independence. Finally, the forecasting errors are measured using a variety of error measures, as detailed in Section 3.3.4, and these results are statistically tested for their relative predictive ability using the Diebold-Mariano test.

Table 4, below, depicts the number of breaches observed for the various models during the out-of-sample test period. Surprisingly, the historical simulation model does not achieve the lowest number of breaches, ranking third with three breaches achieved over the 7,286-day out-of-sample test period. Several models achieved no breaches during the test period, namely the GARCH model and the EGARCH model when calibrated using the normal distribution and the ARCH model when calibrated using the skewed Student's t distribution. The worst performing models, based on the number of breaches observed during the test period, were the ARCH model when calibrated using the normal distribution and the RiskMetrics model when calibrated using the skewed Student's distribution.

While the historical simulation did not achieve the fewest breaches, it is notable that it only achieved three breaches over 7,286 trading days. One of these breaches was, unsurprisingly, on 29 September 2008, indicating the model's failure to adequately react to changing market

conditions, as experienced in the period leading up to the 2008 global financial crisis, although this finding is not unique to the historical simulation model.

Table 4: Breaches Observed for the 10-day 99% Value at Risk Metric using Traditional Models and the Normal and Skewed Student's t Underlying Distributions

	Normal Distribution	**Skewed Student's t Distribution**
	VaR Breaches	VaR Breaches
Historical Simulation	3	
Delta-Normal	5	
ARCH(1)	21	0
GARCH(1,1)	0	4
EGARCH(1,1)	0	6
RiskMetrics	2	21

Note: This table reports the number of breaches experienced for an application of various 10-day value at risk (VaR) forecasting models at the 99% confidence level using either a normal distribution or a skewed Student's t distribution as the underlying distribution, using the daily logged returns of the Standard & Poor's (S&P) 500 index from 15 March 1991 to 14 February 2020. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the forecasted VaR forecast obtained via one of the models detailed in this table. The total number of VaR forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

It is worth highlighting that the low number of breaches achieved by each of the models tested to produce 10-day 99% VaR forecasts, relative to the much larger out-of-sample test period, is supported by the literature. As discussed in Section 3.2, McAleer and da Veiga (2008) find that achieving few breaches is preferable to achieving no breaches, as a bank would rather use an inaccurate model that produces few breaches and incur the additional penalty by a higher penalty variable, k, over using an accurate model which leads to higher capital requirements. The relatively lower number of breaches achieved by the historical simulation model is indicative of its poor performance, as highlighted recently by Taylor (2020). Moreover, this result highlights the historical simulation's lacking volatility-updating capabilities, leading to smoother forecasts relative to other models, as highlighted by Daníelsson (2002).

Table 5 shows that each of the models used in this study to produce 10-day 99% VaR forecasts passes the BCBS's traffic light test with a 'Green zone' outcome, indicating that the models tested and their resulting forecasts are acceptable based on this simple regulatory backtest. Given the few breaches achieved by each of the models, as shown in Table 4, relative to the number of trading days in the out-of-sample test period, even by the worst performing models, it is unsurprising that the BCBS's traffic light test results in 'Green zone' all around.

Table 5: Results of the Basel Committee on Banking Supervision's Traffic Light Test for the 10-day 99% Value at Risk Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Green zone	
Delta-Normal	Green zone	
ARCH(1)	Green zone	Green zone
GARCH(1,1)	Green zone	Green zone
EGARCH(1,1)	Green zone	Green zone
RiskMetrics	Green zone	Green zone

Note: This table reports the results of the Basel Committee on Banking Supervision (BCBS) Traffic Light test for banks' internal models based on the number of breaches of various 10-day value at risk (VaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the models detailed in this table. The total number of VaR forecasts for the study period was 7,286 per model. The number of breaches relative to the period length was then analysed using the BCBS's Traffic Light test to achieve one of three classifications. The classifications are 'Green zone' (if the binomial probability of the number of breaches relative to the number of trading days is greater than 99.99%), 'Yellow zone' (if the corresponding probability is lesser than 99.99% but greater than 99.99%), or 'Red zone' (if the corresponding probability is lesser than 95%). Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

When applying Kupiec's proportion of failure test, it is clear that the very few breaches achieved by each of the models yield the conclusion that the models used are inaccurate at the 1% significance level, as shown in Table 6, below. As shown in the table, given the 99% confidence level of the backtest and the 7,286-trading-day out-of-sample test period, the expected number of breaches for each of the tests was between 50 and 94 breaches, inclusive. Since even the models producing the highest number of breaches only produced 21 breaches, it is clear from the backtest that the forecasting models are inaccurate at the 1% significance level.

As a final backtest of 10-day 99% VaR breaches, the independence of breaches was tested using Christoffersen's test for independence. The universal result is shown in Table 7, below. With so few breaches over a relatively much larger out-of-sample test period, it is unsurprising that whatever few breaches observed are deemed independent of each other.

Table 6: Results of the Kupiec Proportion of Failure Test for the 10-day 99% Value at Risk Metric using Traditional Models

	Normal Distribution		**Skewed Student's t Distribution**	
	Breaches and Range	Conclusion	Breaches and Range	Conclusion
Historical Simulation	$3 \notin [50,94]$	Reject		
Delta-Normal	$5 \notin [50,94]$	Reject		
ARCH(1)	$21 \notin [50,94]$	Reject	$0 \notin [50,94]$	Reject
GARCH(1,1)	$0 \notin [50,94]$	Reject	$4 \notin [50,94]$	Reject
EGARCH(1,1)	$0 \notin [50,94]$	Reject	$6 \notin [50,94]$	Reject
RiskMetrics	$2 \notin [50,94]$	Reject	$21 \notin [50,94]$	Reject

Note: This table reports the results of the Kupiec Proportion of Failure (PoF) backtest at the 1% significance level based on the number of breaches of various 10-day value at risk (VaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the number of true breaches of the model is equal to the observed number of breaches. The results are based on the breaches of the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the models detailed in this table. The total number of VaR forecasts for the study period was 7,286 per model. The left column for each distribution shows the number of breaches experienced under the model and distribution, together with the range predicted using the PoF test. Since the period has 7,286 trading days, the range of expected breaches is $[50,94]$ at the 99% significance level. The critical value of the χ_1^2 at the 99% confidence level is 6.64897. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Table 7: Results of the Christoffersen Test for Independence for the 10-day 99% Value at Risk Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Fail to reject	
Delta-Normal	Fail to reject	
ARCH(1)	Fail to reject	Fail to reject
GARCH(1,1)	Fail to reject	Fail to reject
EGARCH(1,1)	Fail to reject	Fail to reject
RiskMetrics	Fail to reject	Fail to reject

Note: This table reports the results of the Christoffersen test of independence backtest for banks' internal models based on the number of breaches of various 10-day value at risk (VaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the probability of a breach following a non-breach is equal to the probability of a breach following a breach, i.e., breaches are independent. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the models detailed in this table. The total number of VaR forecasts for the study period was 7,286 per model. The critical value of the χ_1^2 at the 99% confidence level is 6.63490. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

As the historical simulation model is widely used by banks, this study focuses on analysing its performance. The backtesting results of the historical simulation model are consistent with the literature. For example, the model conforms to regulatory requirements such as the BCBS's traffic light test, as highlighted in Sharma (2012). Moreover, as per the finding presented by Sharma, the model's performance is mixed when using hypothesis testing, as indicated by its results for Kupiec's backtest and Christoffersen's backtest. The few breaches exhibited also indicate the conservatism of the historical simulation model, a result which is in agreement with the findings of Berkowitz and O'Brien (2002).

Table 8, below, depicts the MAE, the RMSE, the MAPE, and the MdAPE values for the various 10-day 99% VaR models using the normal distribution, while Table 9 does the same for the skewed Student's t distribution. As discussed in Section 3.3.4, the lower each of these error measures is, the better. The model experiencing the lowest measure implies the lowest cumulative difference between the forecasts made and the returns experienced, therefore being the most accurate model. The one whose forecasts are furthest from the experienced returns would be considered the least accurate model, based on the application of forecasting error measures.

When it comes to the normal distribution, as can be seen from the relevant table, the EGARCH model produced the most accurate 10-day 99% VaR forecasts based on all four forecasting error measures. The second most accurate model, also across all measures, is the RiskMetrics model. Interestingly, three out of four measures (i.e., all but the MdAPE) indicate that the historical simulation model is the second-least accurate forecasting model tested, with the ARCH model being the least accurate model, further highlighting the model's low updating abilities and its tendency to produce conservative forecasts, as highlighted by Berkowitz and O'Brien (2002).

When it comes to the skewed Student's t distribution, the forecasting error measures shown in Table 9 tell a relatively similar story. While the forecasting error measures show that, depending on which measure is chosen, either the EGARCH model or the RiskMetrics model is the most accurate forecasting model, there is no question that the historical simulation model is the least accurate forecasting model when using the skewed Student's t distribution as the underlying distribution. The results suggest that the models that achieved the most breaches (i.e., the RiskMetrics model with 21 breaches and the EGARCH model with six breaches) are

the most accurate with respect to the forecasting error measures used in this study. This result does not necessarily hold for the normal distribution in the case of the 10-day 99% VaR models.

Table 8: Forecasting Error Measures for the 10-day 99% Value at Risk Metric using Traditional Models (Normal Distribution)

	MAE	**RMSE**	**MAPE**	**MdAPE**
Historical Simulation	0.0944	0.1003	98.1347%	5.7790%
Delta-Normal	0.0793	0.0833	82.4741%	5.0418%
ARCH(1)	0.1115	0.1378	113.0929%	5.7685%
GARCH(1,1)	0.0715	0.0814	60.3629%	5.5495%
EGARCH(1,1)	0.0694	0.0776	59.7189%	4.8893%
RiskMetrics	0.0698	0.0813	59.7930%	4.9021%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day value at risk (VaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the normal distribution as the underlying distribution and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Table 9: Forecasting Error Measures for the 10-day 99% Value at Risk Metric using Traditional Models (Skewed Student's t Distribution)

	MAE	**RMSE**	**MAPE**	**MdAPE**
Historical Simulation	0.0944	0.1003	98.1347%	5.7790%
Delta-Normal	0.0793	0.0833	82.4741%	5.0418%
ARCH(1)	0.0822	0.0876	80.7287%	5.6866%
GARCH(1,1)	0.0728	0.0839	63.5373%	4.9744%
EGARCH(1,1)	0.0713	0.0812	63.3376%	4.3513%
RiskMetrics	0.0708	0.0826	62.8939%	4.4593%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day value at risk (VaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the skewed Student's t distribution as the underlying distribution and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

The differences between the results of Table 8 and Table 9 are worth discussing. As a reminder, note that the historical simulation model and delta-normal model are distribution agnostic, meaning that they did not use either the normal distribution or the skewed Student's t distribution to produce forecasts. Hence, of the remaining models, it is interesting that the

GARCH model, the EGARCH model, and the RiskMetrics model all yield less accurate forecasts when the skewed Student's t distribution was used to produce forecasts as opposed to the normal distribution. This is evident by the three forecasting error measures which do not depend on ordered forecasts (i.e., all but the MdAPE). On the other hand, the ARCH model experienced improved forecasting ability when calibrated using the skewed Student's t distribution as the underlying statistical distribution relative to the forecasts achieved when using the normal distribution instead.

Across both distributions, it is clear that the historical simulation model performs relatively poorly, while the EGARCH model and the RiskMetrics model both produced more accurate forecasts relative to all other models and across all error measures and distributions. However, it is worth noting that no model of those tested yielded particularly accurate forecasts, regardless of the distributional assumption used, as shown by the relatively high error measures in Table 8 and Table 9. Hence, it can be concluded that the general methodology used in practice today to calculate 10-day 99% VaR forecasts is inadequate and inefficient, as shown by the large forecasting error measures.

Finally, consider the results of the Diebold-Mariano tests performed for the various models at the 99% confidence level using the normal distribution (Table 10) and the skewed Student's t distribution (Table 11). First, recall that the null hypothesis of the test states that the two models tested have equal forecasting abilities, while the alternative hypothesis states that the model stated second has superior forecasting abilities relative to the first model stated. In Table 10 and Table 11, positive Diebold-Mariano statistics lead to lower p-values, ultimately leading to the rejection of the null hypothesis, while negative Diebold-Mariano statistics lead to higher p-values, ultimately leading to the failure to reject the null hypothesis.

As can be seen from Table 10, when it comes to forecasting 10-day 99% VaR forecasts using the normal distribution, the following can be concluded.

i. Every model has superior forecasting abilities relative to the historical simulation model, apart from the ARCH model, at the 1% significance level.
ii. The delta-normal model has superior forecasting abilities relative to the historical simulation model, but has equal forecasting abilities to the ARCH model, at the 1% significance level. Every other model has superior forecasting abilities relative to the delta-normal model at the 1% significance level.

Table 10: Results of the Diebold-Mariano Test for the 10-day 99% Value at Risk Metric using Traditional Models (Normal Distribution)

	Diebold-Mariano Statistic	**p-Value**
Historical Simulation versus Delta-Normal	69.71	<2.2e-16
Historical Simulation versus ARCH(1)	-23.68	1
Historical Simulation versus GARCH(1,1)	25.36	<2.2e-16
Historical Simulation versus EGARCH(1,1)	36.13	<2.2e-16
Historical Simulation versus RiskMetrics	24.66	<2.2e-16
Delta-Normal versus ARCH(1)	-32.67	1
Delta-Normal versus GARCH(1,1)	2.43	0.0075
Delta-Normal versus EGARCH(1,1)	9.49	<2.2e-16
Delta-Normal versus RiskMetrics	2.50	0.0063
ARCH(1) versus GARCH(1,1)	33.92	<2.2e-16
ARCH(1) versus EGARCH(1,1)	35.97	<2.2e-16
ARCH(1) versus RiskMetrics	33.43	<2.2e-16
GARCH(1,1) versus EGARCH(1,1)	12.17	<2.2e-16
GARCH(1,1) versus RiskMetrics	0.72	0.2357
EGARCH(1,1) versus RiskMetrics	-9.18	1

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day value at risk (VaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the normal distribution as the underlying distribution. The total number of VaR forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

iii. The ARCH model has equal forecasting abilities to the historical simulation model and delta-normal model at the 1% significance level. Every other model has superior forecasting abilities to the ARCH model at the 1% significance level.

iv. The GARCH model has superior forecasting abilities to the historical simulation model, the delta-normal model, and the ARCH model and has equal forecasting abilities to the RiskMetrics model at the 1% significance level. The GARCH model does not have superior forecasting abilities to the EGARCH model at the 1% significance level.

v. The EGARCH model has superior forecasting abilities to every other model at the 1% significance level, except for the RiskMetrics model, which has equal forecasting abilities.

vi. The RiskMetrics model has superior forecasting abilities to the historical simulation model, the delta-normal model, and the ARCH model at the 1% significance level, and has equal forecasting abilities to the GARCH model and EGARCH model.

The results of the Diebold-Mariano test presented in Table 10 correspond to the findings presented in Table 8. Both tables show that the EGARCH model produced the most accurate forecasts when forecasting 10-day 99% VaR forecasts using the normal distribution as the underlying statistical distribution. The two tables further show that, using the various forecasting error measures and the Diebold-Mariano test, several models may produce more accurate forecasts relative to the historical simulation model for this specific calibration of the VaR metric.

Similarly, as can be seen from Table 11, when it comes to forecasting 10-day 99% VaR forecasts using the skewed Student's t distribution, the following can be concluded.

i. Every model has superior forecasting abilities relative to the historical simulation model at the 1% significance level.

ii. The delta-normal model has superior forecasting abilities relative to the historical simulation model, and has equal forecasting abilities to the GARCH model and the RiskMetrics model at the 1% significance level. It does not, however, have superior forecasting abilities to the EGARCH model at the 1% significance level.

iii. The ARCH model has superior forecasting abilities to the historical simulation model and has equal forecasting abilities to the delta-normal model at the 1% significance level. The GARCH model, EGARCH model, and the RiskMetrics model all have equal forecasting abilities to the ARCH model at the 1% significance level.

iv. The GARCH model has superior forecasting abilities to the historical simulation model and the ARCH model, but has equal forecasting abilities to the delta-normal model, at the 1% significance level. However, the EGARCH model and the RiskMetrics model have superior forecasting abilities to the GARCH model at the 1% significance level.

v. The EGARCH model has superior forecasting abilities to every other model, except for the RiskMetrics model, where these have equal forecasting abilities at the 1% significance level.

vi. The RiskMetrics model has equal forecasting abilities to the EGARCH model and the delta-normal model, but has superior forecasting abilities to all other models at the 1% significance level.

Table 11: Results of the Diebold-Mariano Test for the 10-day 99% Value at Risk Metric using Traditional Models (Skewed Student's t Distribution)

	Diebold-Mariano Statistic	**p-Value**
Historical Simulation versus Delta-Normal	69.71	<2.2e-16
Historical Simulation versus ARCH(1)	21.05	<2.2e-16
Historical Simulation versus GARCH(1,1)	17.88	<2.2e-16
Historical Simulation versus EGARCH(1,1)	24.07	<2.2e-16
Historical Simulation versus RiskMetrics	19.82	<2.2e-16
Delta-Normal versus ARCH(1)	-8.08	1
Delta-Normal versus GARCH(1,1)	-0.65	0.7416
Delta-Normal versus EGARCH(1,1)	2.87	0.0021
Delta-Normal versus RiskMetrics	0.80	0.2108
ARCH(1) versus GARCH(1,1)	5.61	1.027e-08
ARCH(1) versus EGARCH(1,1)	11.60	<2.2e-16
ARCH(1) versus RiskMetrics	7.08	7.746e-13
GARCH(1,1) versus EGARCH(1,1)	7.67	9.49e-15
GARCH(1,1) versus RiskMetrics	6.62	1.943e-11
EGARCH(1,1) versus RiskMetrics	-3.69	0.9999

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day value at risk (VaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the skewed Student's t distribution as the underlying distribution. The total number of VaR forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Finally, the results of the Diebold-Mariano test for the skewed Student's t distribution presented in Table 11 also correspond to the results of the same models when evaluating the

various forecasting error measures, as presented in Table 9. While the ARCH model produced more accurate forecasts when using the skewed Student's t distribution rather than the normal distribution, the various forecasting error measures and the Diebold-Mariano test concur that the EGARCH model and the RiskMetrics model produced the most accurate forecasts when it comes to 10-day 99% VaR forecasts when using the skewed Student's t distribution as the underlying statistical distribution. Moreover, the results presented in the two tables further concur that the historical simulation model produces the least accurate forecasts across the board.

3.4.2. Stressed Value at Risk

This section discusses and analyses the results of the models employed in this study to forecast 10-day 99% SVaR forecasts for a US bank using the S&P 500 index as the underlying, over the period 15 March 1991 to 14 February 2020. First, the number of breaches experienced for each of the various models is presented and discussed, followed by an analysis of the backtesting results achieved using the BCBS's traffic light test, Kupiec's proportion of failure test, and Christoffersen's test for independence. Finally, the forecasting errors are measured using a variety of error measures, as detailed in Section 3.3.4, and these results are statistically tested for their relative predictive ability using the Diebold-Mariano test.

The breaches observed by the various models used to forecast the 10-day 99% SVaR forecasts are summarised in Table 12, below. Unsurprisingly, there were very few breaches to report, if any. This is expected as this metric was calculated over a stressed period, i.e., the most severe daily profit and loss figures were used to calibrate the models and obtain forecasts. Intriguingly, the EGARCH model and the GARCH model exhibited the highest number of breaches out of all models using the normal distribution as the underlying statistical distribution, even if this high number of breaches is translated into only two breaches over the 7,286-day out-of-sample period.

The very few breaches experienced across the models and the distributions used unsurprisingly led to 'Green zone' outcomes across the board when applying the BCBS's traffic light test to the 10-day 99% SVaR forecasts, as shown in Table 13, below. It is evident that the BCBS's traffic light test is not a suitable test to test the statistical accuracy of a forecasting model, but, rather, it is a test to validate that a forecasting model's observed breaches are not statistically excessive, rendering the test to be of little use when banks' arsenals of forecasting models produce so few breaches by design.

Table 12: Breaches Observed for the 10-day 99% Stressed Value at Risk Metric using Traditional Models and the Normal and Skewed Student's t Underlying Distributions

	Normal Distribution	**Skewed Student's t Distribution**
	SVaR Breaches	SVaR Breaches
Historical Simulation	0	
Delta-Normal	0	
ARCH(1)	1	0
GARCH(1,1)	2	0
EGARCH(1,1)	2	0
RiskMetrics	0	0

Note: This table reports the number of breaches experienced for an application of various 10-day stressed value at risk (SVaR) forecasting models at the 99% confidence level using either a normal distribution or a skewed Student's t distribution as the underlying distribution, using the daily logged returns of the Standard & Poor's (S&P) 500 index from 15 March 1991 to 14 February 2020. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the forecasted SVaR forecast obtained via one of the models detailed in this table, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SVaR forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Table 13: Results of the Basel Committee on Banking Supervision's Traffic Light Test for the 10-day 99% Stressed Value at Risk Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Green zone	
Delta-Normal	Green zone	
ARCH(1)	Green zone	Green zone
GARCH(1,1)	Green zone	Green zone
EGARCH(1,1)	Green zone	Green zone
RiskMetrics	Green zone	Green zone

Note: This table reports the results of the Basel Committee on Banking Supervision (BCBS) Traffic Light test for banks' internal models based on the number of breaches of various 10-day stressed value at risk (SVaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the forecasted SVaR forecast obtained via one of the models detailed in this table, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SVaR forecasts for the study period was 7,286 per model. The number of breaches relative to the period length was then analysed using the BCBS's Traffic Light test to achieve one of three classifications. The classifications are 'Green zone' (if the binomial probability of the number of breaches relative to the number of trading days is greater than 99.99%), 'Yellow zone' (if the corresponding probability is lesser than 99.99% but greater than 99.99%), or 'Red zone' (if the corresponding probability is lesser than 95%). Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Table 14 depicts the results of the Kupiec proportion of failure test for the 10-day 99% SVaR forecasts obtained using the various models employed in this study. As discussed in the previous section, the number of breaches expected for the 7,286-trading-day out-of-sample period is the inclusive range from 50 to 94 breaches, at the 99% confidence level. Hence, given the few breaches observed for each of the models employed in this study, as shown in Table 12, it is unsurprising that all models were deemed statistically inaccurate from a forecasting adequacy perspective at the 99% confidence level.

Table 14: Results of the Kupiec Proportion of Failure Test for the 10-day 99% Stressed Value at Risk Metric using Traditional Models

	Normal Distribution		**Skewed Student's t Distribution**	
	Breaches and Range	Conclusion	Breaches and Range	Conclusion
Historical Simulation	$0 \notin [50,94]$	Reject		
Delta-Normal	$0 \notin [50,94]$	Reject		
ARCH(1)	$1 \notin [50,94]$	Reject	$0 \notin [50,94]$	Reject
GARCH(1,1)	$2 \notin [50,94]$	Reject	$0 \notin [50,94]$	Reject
EGARCH(1,1)	$2 \notin [50,94]$	Reject	$0 \notin [50,94]$	Reject
RiskMetrics	$0 \notin [50,94]$	Reject	$0 \notin [50,94]$	Reject

Note: This table reports the results of the Kupiec Proportion of Failure (PoF) backtest at the 1% significance level based on the number of breaches of various 10-day stressed value at risk (SVaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the number of true breaches of the model is equal to the observed number of breaches. The results are based on the breaches of the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the SVaR forecast obtained via one of the models detailed in this table, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SVaR forecasts for the study period was 7,286 per model. The left column for each distribution shows the number of breaches experienced under the model and distribution, together with the range predicted using the PoF test. Since the period has 7,286 trading days, the range of expected breaches is $[50,94]$ at the 99% significance level. The critical value of the χ_1^2 at the 99% confidence level is 6.64897. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

The final 10-day 99% SVaR forecasts backtest employed is Christoffersen's test for independence, the results of which are shown in Table 15. Once again, since the majority of the models employed produced zero breaches in the 7,286-trading-day out-of-sample period, it is unsurprising that the backtest's results lead to the conclusion that the null hypothesis cannot be rejected, i.e., all of the models employed exhibit independent breaches. This result holds regardless of which model and which statistical distribution were used to produce the 10-day 99% SVaR forecasts tested.

Table 15: Results of the Christoffersen Test for Independence for the 10-day 99% Stressed Value at Risk Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Fail to reject	
Delta-Normal	Fail to reject	
ARCH(1)	Fail to reject	Fail to reject
GARCH(1,1)	Fail to reject	Fail to reject
EGARCH(1,1)	Fail to reject	Fail to reject
RiskMetrics	Fail to reject	Fail to reject

Note: This table reports the results of the Christoffersen test of independence backtest for banks' internal models based on the number of breaches of various 10-day stressed value at risk (SVaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the probability of a breach following a non-breach is equal to the probability of a breach following a breach, i.e., breaches are independent. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the SVaR forecast obtained via one of the models detailed in this table, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SVaR forecasts for the study period was 7,286 per model. The critical value of the χ_1^2 at the 99% confidence level is 6.63490. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Given the little new insight obtained by employing the various backtests above, it is of greater interest to examine the relative efficiency of the models employed to produce the 10-day 99% SVaR forecasts examined in this section of the study. Table 16, below, depicts the results of the various forecasting error measures employed in this study, namely the MAE, the RMSE, the MAPE, and the MdAPE, as calculated for the various 10-day 99% SVaR forecasting models using the normal distribution. Table 17 depicts the same for the skewed Student's t distribution.

The results of the various forecasting error measures displayed in Table 16 show that, for the normal distribution, the GARCH model produced the most accurate forecasts based on all forecasting error measures, apart from the MdAPE measure, followed by the EGARCH model (which produced the most accurate forecasts when evaluating the 10-day 99% VaR forecasts in the previous section). It is also concluded that the RiskMetrics model produced the least accurate forecasts when forecasting 10-day 99% SVaR forecasts, while the historical simulation model and the ARCH model produced the second-least accurate and third-least accurate forecasts, depending on which forecasting error measure was used.

Table 16: Forecasting Error Measures for the 10-day 99% Stressed Value at Risk Metric using Traditional Models (Normal Distribution)

	MAE	**RMSE**	**MAPE**	**MdAPE**
Historical Simulation	0.1482	0.1544	156.9730%	8.2343%
Delta-Normal	0.1149	0.1172	120.3341%	7.0125%
ARCH(1)	0.1270	0.1606	125.5626%	6.2191%
GARCH(1,1)	0.0931	0.0965	97.2833%	5.6640%
EGARCH(1,1)	0.0937	0.0977	97.9407%	5.6394%
RiskMetrics	0.3746	0.3929	394.8635%	18.5501%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day stressed value at risk (SVaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the normal distribution as the underlying distribution and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Interestingly, the RiskMetrics model produced surprisingly poor forecasts for the 10-day 99% SVaR metric, ranking last of the forecasting models when using the normal distribution. This is in contrast to its rank as being the model producing the second most accurate forecasts when generating 10-day 99% VaR forecasts using the normal distribution, as shown in Table 8 in the previous section. In fact, it also produces the least accurate forecasts when using the skewed Student's t distribution, as shown in Table 17, although the model's performance is much closer to that of the other models tested in the case of the skewed Student's t distribution serving as the underlying statistical distribution. This is in line with the expectation that, while the skewed Student's t distribution fits the return distribution better in general, its fit in the tail is worse relative to the fit of the normal distribution as the underlying statistical distribution.

The two measures (VaR and SVaR) do, however, concur on the poor performance of the ARCH model and the historical simulation model, placing them either as the least accurate and the second-least accurate or the second-least accurate and the third-least accurate, respectively, using the 10-day 99% VaR forecasting models and 10-day 99% SVaR forecasting models, and the normal distribution for both. The GARCH model produced the most accurate forecasts, while the EGARCH model produced the second-most accurate forecasts, using the 10-day 99% SVaR metric, scoring close to their respective positions when evaluating the 10-day 99% VaR forecasts, i.e., third-most accurate and most accurate, respectively.

Returning to the results of the various 10-day 99% SVaR forecasting models with reference to their performances using the various forecasting error measures employed in this study, as depicted in Table 17, the results are quite different to those obtained using the normal distribution. Note, once again, that the historical simulation model and the delta-normal model produce the exact same values for the various forecasting error measures due to their distribution agnostic nature. Hence, their values are repeated in Table 17 for ease of reference.

Table 17: Forecasting Error Measures for the 10-day 99% Stressed Value at Risk Metric using Traditional Models (Skewed Student's t Distribution)

	MAE	**RMSE**	**MAPE**	**MdAPE**
Historical Simulation	0.1482	0.1544	156.9730%	8.2343%
Delta-Normal	0.1149	0.1172	120.3341%	7.0125%
ARCH(1)	0.1117	0.1154	117.2018%	6.6152%
GARCH(1,1)	0.1400	0.1449	148.5899%	7.9436%
EGARCH(1,1)	0.1524	0.1564	159.3564%	8.7720%
RiskMetrics	0.1789	0.1922	190.9900%	8.6353%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day stressed value at risk (SVaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the skewed Student's t distribution as the underlying distribution and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution and, therefore, are not included in this graph.

Using the skewed Student's t distribution, Table 17 shows that the model that produces the most accurate forecasts is now the ARCH model – a result that is in contrast to the model's performance when using the normal distribution as the underlying statistical distribution, where it ranked as the third-least accurate model. While the RiskMetrics model is still the least accurate model, as mentioned, its performance is now much closer to those of the other models, indicating that the skewed Student's t distribution produces more accurate forecasts when employing the RiskMetrics model relative to the normal distribution. The RiskMetrics model is followed by the EGARCH model, which produces the second-least accurate 10-day 99% SVaR forecasts using the normal distribution as the underlying statistical distribution. Using the skewed Student's t distribution as the underlying statistical distribution, the historical simulation model is now only the third-least accurate model, trailing the RiskMetrics model and the EGARCH model.

Interestingly, the ARCH model's performance has not improved much when using the skewed Student's t distribution instead of the normal distribution as the underlying statistical distribution, meaning that it climbs to be the model producing the most accurate forecasts among the models employing the skewed Student's t distribution was driven primarily by the marked deterioration in the performances of the GARCH model and the EGARCH model when employing the skewed Student's t distribution instead of the normal distribution. Specifically, the two measures deteriorated between 40% and 65% across the various forecasting error measures, indicating that the fit of the skewed Student's t distribution is inferior to that of the normal distribution when used to produce 10-day 99% SVaR forecasts. This corroborates the observation made in the previous section surrounding the fit of the skewed Student's t distribution's tail to the underlying returns data.

Once again, as seen with respect to the skewed Student's t distribution in the previous section, it can be observed that the models that achieve the most breaches are the models that produce the most accurate forecasts, per their respective performances with respect to the forecasting error measures. In the case of the 10-day 99% SVaR forecasts, this result is now true when using the normal distribution as the underlying statistical distribution, while this result was true for the skewed Student's t distribution when producing 10-day 99% VaR forecasts. This result's merit cannot be assessed when using the skewed Student's t distribution to produce 10-day 99% SVaR forecasts as the models produced no breaches.

Moreover, it is clear that the historical simulation model, which is often used by banks as the go-to model when forecasting VaR, performed inadequately across the two distributions used to forecast 10-day 99% SVaR forecasts. The model produces the second-least accurate forecasts when employing the normal distribution as the underlying distribution assumption, while it produces the third-least accurate forecasts when employing the skewed Student's t distribution as the underlying distribution assumption.

From the results of the forecasting error measures detailed above, it is clear, still, that the various models employed are not efficient, regardless of which error measure or distribution is used, as evident by the high magnitudes of the error measures' values, suggesting highly conservative practices. Hence, it can be concluded that the general methodology used in practice today to produce 10-day 99% SVaR forecasts is inadequate and inefficient, as shown by the large forecasting error measures.

Finally, the statistical superiority in forecasting 10-day 99% SVaR forecasts using the various models employed in this study is tested using the Diebold-Mariano test at the 99% confidence level. Once again, recall that the null hypothesis of the test states that the models have equal forecasting abilities, while the alternative hypothesis states that the model stated second has superior forecasting abilities relative to the first model stated. Positive Diebold-Mariano statistics in Table 18 and Table 19 are accompanied by lower p-values, and suggest that the null hypothesis should be rejected, while the opposite is true for negative Diebold-Mariano statistics.

The results depicted in Table 18 can be summarised as follows, as they pertain to 10-day 99% SVaR forecasting models using the normal distribution as the underlying statistical distribution.

i. The historical simulation model has equal forecasting abilities to the ARCH model and the RiskMetrics model, and worse forecasting abilities relative to the delta-normal model, the GARCH model, and the EGARCH model at the 1% significance level.
ii. The delta-normal model has superior forecasting abilities relative to the historical simulation model and has equal forecasting abilities to the ARCH model and the RiskMetrics model at the 1% significance level. It also has worse forecasting abilities relative to the GARCH model and EGARCH model at the 1% significance level.
iii. The ARCH model has equal forecasting abilities relative to the historical simulation model, the delta-normal model, and the RiskMetrics models at the 1% significance level, but worse forecasting abilities relative to the GARCH model and the EGARCH model.
iv. The GARCH model has equal forecasting abilities relative to the EGARCH model and the RiskMetrics model, but superior forecasting abilities relative to the historical simulation model, the delta-normal model, and the ARCH model at the 1% significance level.
v. The EGARCH model has equal forecasting abilities relative to the GARCH model and the RiskMetrics model, but superior forecasting abilities relative to the historical simulation model, the delta-normal model, and the ARCH model at the 1% significance level.

vi. The RiskMetrics model has equal forecasting abilities relative to all other models at the 1% significance level.

Table 18: Results of the Diebold-Mariano Test for the 10-day 99% Stressed Value at Risk Metric using Traditional Models (Normal Distribution)

	Diebold-Mariano Statistic	**p-Value**
Historical Simulation versus Delta-Normal	104.05	<2.2e-16
Historical Simulation versus ARCH(1)	-2.53	0.9944
Historical Simulation versus GARCH(1,1)	142.01	<2.2e-16
Historical Simulation versus EGARCH(1,1)	145.06	<2.2e-16
Historical Simulation versus RiskMetrics	-156.26	1
Delta-Normal versus ARCH(1)	-16.21	1
Delta-Normal versus GARCH(1,1)	365.06	<2.2e-16
Delta-Normal versus EGARCH(1,1)	185.10	<2.2e-16
Delta-Normal versus RiskMetrics	-152.35	1
ARCH(1) versus GARCH(1,1)	22.18	<2.2e-16
ARCH(1) versus EGARCH(1,1)	21.87	<2.2e-16
ARCH(1) versus RiskMetrics	-102.36	1
GARCH(1,1) versus EGARCH(1,1)	-16.77	1
GARCH(1,1) versus RiskMetrics	-156.17	1
EGARCH(1,1) versus RiskMetrics	-156.44	1

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day stressed value at risk (SVaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the normal distribution as the underlying distribution. The total number of SVaR forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

The forecasting error measures, as shown in Table 16, and the Diebold-Mariano test, the results of which are shown in Table 18, conclude similarly that the GARCH model and the EGARCH model produced the most accurate forecasts when using the normal distribution as the underlying statistical distribution. The two tables' results further concur that the RiskMetrics model had the poorest fit, captured by its performance being the worst out of the various models tested for the normal distribution. Last, the results presented in the two tables

further concur that the historical simulation model provides poor forecasts, ranking most of the models tested ahead of this model.

Table 19: Results of the Diebold-Mariano Test for the 10-day 99% Stressed Value at Risk Metric using Traditional Models (Skewed Student's t Distribution)

	Diebold-Mariano Statistic	**p-Value**
Historical Simulation versus Delta-Normal	104.05	<2.2e-16
Historical Simulation versus ARCH(1)	89.66	<2.2e-16
Historical Simulation versus GARCH(1,1)	59.44	<2.2e-16
Historical Simulation versus EGARCH(1,1)	-11.28	1
Historical Simulation versus RiskMetrics	-92.39	1
Delta-Normal versus ARCH(1)	5.89	2.037e-09
Delta-Normal versus GARCH(1,1)	-106.58	1
Delta-Normal versus EGARCH(1,1)	-152.93	1
Delta-Normal versus RiskMetrics	-99.39	1
ARCH(1) versus GARCH(1,1)	-79.90	1
ARCH(1) versus EGARCH(1,1)	-111.99	1
ARCH(1) versus RiskMetrics	-97.94	1
GARCH(1,1) versus EGARCH(1,1)	-71.56	1
GARCH(1,1) versus RiskMetrics	-90.64	1
EGARCH(1,1) versus RiskMetrics	-69.26	1

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day stressed value at risk (SVaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the skewed Student's t distribution as the underlying distribution. The total number of SVaR forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

The results depicted in Table 19 are summarised as follows, as they pertain to 10-day 99% SVaR forecasting models using the skewed Student's t distribution as the underlying statistical distribution.

i. The historical simulation model has equal forecasting abilities to the EGARCH model and the RiskMetrics model at the 1% significance level. However, the

delta-normal model, the ARCH model, and the GARCH model all have superior forecasting abilities relative to the historical simulation model at the 1% significance level.

ii. The delta-normal model has equal forecasting abilities to the GARCH model, the EGARCH model, and the RiskMetrics model at the 1% significance level, but superior forecasting abilities relative to the historical simulation model and inferior forecasting abilities to the ARCH model, also at the 1% significance level.

iii. The ARCH model has equal forecasting abilities relative to the GARCH model, the EGARCH model, and the RiskMetrics model at the 1% significance level, but superior forecasting abilities relative to the historical simulation model and the delta-normal model at the 1% significance level.

iv. The GARCH model has equal forecasting abilities relative to the delta-normal model, the ARCH model, the EGARCH model, and the RiskMetrics model, but superior forecasting abilities relative to the historical simulation model at the 1% significance level.

v. The EGARCH model has equal forecasting abilities relative to all other models at the 1% significance level.

vi. The RiskMetrics model has equal forecasting abilities relative to all other models at the 1% significance level.

The results of the Diebold-Mariano test for the skewed Student's t distribution presented in Table 19 also correspond to the results of the same models when evaluating the various forecasting error measures, as presented in Table 17. Both tables show that the ARCH model produces the most accurate 10-day 99% SVaR forecasts relative to the other models tested when using the skewed Student's t distribution as the underlying distribution. Moreover, they also concur on the poor performance of the RiskMetrics model, showing that it ranks last among the models tested (or equal in forecasting ability to the worst ranking model). The forecasting error measure and the Diebold-Mariano tests further support each other's conclusions when concluding that the performance of the historical simulation model is ranked somewhere in the lower half of the ranks, as depicted by the various models that produce both lower forecasting error values and superior forecasting abilities at the 1% significance level.

3.4.3. Expected Shortfall

This section discusses and analyses the results of the models employed in this study to forecast 10-day 97.5% ES forecasts for a US bank using the S&P 500 index as the underlying,

over the period 15 March 1991 to 14 February 2020. First, the number of breaches experienced for each of the various models is presented and discussed, followed by an analysis of the backtesting results achieved using the BCBS's traffic light test, the conditional, the unconditional, the minimally biased, as well as the Du-Escanciano backtest. Finally, the forecasting errors are measured using a variety of error measures, as detailed in Section 3.3.4, and these results are statistically tested for their relative predictive ability using the Diebold-Mariano test.

Table 20: Breaches Observed for the 10-day 97.5% Expected Shortfall Metric using Traditional Models and the Normal and Skewed Student's t Underlying Distributions

	Normal Distribution	**Skewed Student's t Distribution**
	ES Breaches	ES Breaches
Historical Simulation	3	
Delta-Normal	7	
ARCH(1)	19	19
GARCH(1,1)	10	21
EGARCH(1,1)	8	2
RiskMetrics	11	18

Note: This table reports the number of breaches experienced for an application of various 10-day expected shortfall (ES) forecasting models at the 97.5% confidence level using either a normal distribution or a skewed Student's t distribution as the underlying distribution, using the daily logged returns of the Standard & Poor's (S&P) 500 index from 15 March 1991 to 14 February 2020. An ES breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the ES forecast obtained via one of the models detailed in this table. The total number of ES forecasts for the study period was 7,286 per model. Note also that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Table 20, above, shows the breaches observed for each of the models used to obtain 10-day 97.5% ES forecasts during the 7,286-trading-day out-of-sample test period. The historical simulation model, the primary model used by banks in the US, achieves three breaches, placing it as the model to yield the second-fewest breaches, trailing only the EGARCH model using the skewed Student's t distribution as the underlying distribution. The performances of the various models showed mixed results when swapping the normal distribution for the skewed Student's t distribution as the underlying statistical distribution. For example, the number of breaches observed for the GARCH model and the RiskMetrics model increased (i.e., the models showed deteriorating performance in producing 10-day 97.5% ES forecasts) when using the skewed Student's t distribution instead of the normal distribution, while the number of breaches observed for the EGARCH model decreased (i.e., the performance improved). The ARCH model achieved 19 breaches regardless of which underlying distribution was used.

The results of the BCBS's traffic light test for the 10-day 97.5% ES forecasts are shown in Table 21, below. As per the 10-day 99% VaR forecasts and 10-day 99% SVaR forecasts, and breaches thereof, discussed in the sections preceding this section, the very few breaches experienced over the 7,286-trading-day out-of-sample test period lead to a 'Green zone' conclusion across the board for the various models and underlying distributions used to forecast 10-day 97.5% ES forecasts. Once again, it seems that the models used to produce 10-day 97.5% ES forecasts produced conservative forecasts which underestimated risk, produced few breaches, and backtested well using the BCBS's traffic light test.

Table 21: Results of the Basel Committee on Banking Supervision's Traffic Light Test for the 10-day 97.5% Expected Shortfall Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Green zone	
Delta-Normal	Green zone	
ARCH(1)	Green zone	Green zone
GARCH(1,1)	Green zone	Green zone
EGARCH(1,1)	Green zone	Green zone
RiskMetrics	Green zone	Green zone

Note: This table reports the results of the Basel Committee on Banking Supervision (BCBS) Traffic Light test for banks' internal models based on the number of breaches of various 10-day expected shortfall (ES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. An ES breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the ES forecast obtained via one of the models detailed in this table. The total number of ES forecasts for the study period was 7,286 per model. The number of breaches relative to the period length was then analysed using the BCBS's Traffic Light test to achieve one of three classifications. The classifications are 'Green zone' (if the binomial probability of the number of breaches relative to the number of trading days is greater than 99.99%), 'Yellow zone' (if the corresponding probability is lesser than 99.99% but greater than 99.99%), or 'Red zone' (if the corresponding probability is lesser than 95%). Note also that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Turning to more statistically important backtests, the first ES-specific backtest employed is the conditional backtest, whereby the 10-day 97.5% ES forecasts are evaluated together with their respective 10-day VaR forecasts. The backtest's results are shown in Table 22, below.

Note that the conditional backtest can only be performed should at least one VaR breach be observed, rendering the backtest impossible to employ in the cases where no VaR breaches are observed. Considering the 10-day 99% VaR breaches displayed in Table 4, this is the case for three of the models employed in this study, namely the ARCH model using the skewed Student's t distribution, and the GARCH model and the EGARCH model using the normal

distribution. Hence, the results for these three models are captured as 'Cannot perform backtest' in Table 22, below.

Table 22: Results of the Conditional Backtest for the 10-day 97.5% Expected Shortfall Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Reject	
Delta-Normal	Reject	
ARCH(1)	Reject	Cannot perform backtest
GARCH(1,1)	Cannot perform backtest	Reject
EGARCH(1,1)	Cannot perform backtest	Reject
RiskMetrics	Reject	Reject

Note: This table reports the results of the conditional backtest for banks' internal models based on the number of breaches of various 10-day expected shortfall (ES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the ES forecasts observed are the true ES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. The test can only be carried out if there is at least one value at risk (VaR) breach. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the models detailed in this table. The total number of ES forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

For the remaining models, the conditional backtest is employed and the null hypothesis is tested, where the null hypothesis states that the observed 10-day 97.5% ES forecasts are the true 10-day 97.5% ES forecasts. For all of the models that can be tested (i.e., those with at least one corresponding 10-day VaR breach), this null hypothesis is rejected at the 97.5% confidence level.

Similarly to the conditional backtest, the unconditional backtest's test statistic is also dependent on at least one observed corresponding 10-day VaR breach. Hence, as for the conditional backtest, the results for the ARCH model using the skewed Student's t distribution, and the GARCH model and the EGARCH model using the normal distribution are captured as 'Cannot perform backtest' in Table 23, below.

For the remaining models, the unconditional backtest is employed and the null hypothesis is tested, where the null hypothesis, once again, states that the observed 10-day 97.5% ES forecasts are the true 10-day 97.5% ES forecasts. For all of the models that can be tested (i.e., those with at least one corresponding 10-day VaR breach), this null hypothesis is rejected at the 97.5% confidence level, as shown in Table 23, below.

Table 23: Results of the Unconditional Backtest for the 10-day 97.5% Expected Shortfall Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Reject	
Delta-Normal	Reject	
ARCH(1)	Reject	Cannot perform backtest
GARCH(1,1)	Cannot perform backtest	Reject
EGARCH(1,1)	Cannot perform backtest	Reject
RiskMetrics	Reject	Reject

Note: This table reports the results of the unconditional backtest for banks' internal models based on the number of breaches of various 10-day expected shortfall (ES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the ES forecasts observed are the true ES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. The test can only be carried out if there is at least one value at risk (VaR) breach. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the models detailed in this table. The total number of ES forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Table 24: Results of the Minimally Biased Backtest for the 10-day 97.5% Expected Shortfall Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Reject	
Delta-Normal	Reject	
ARCH(1)	Reject	Reject
GARCH(1,1)	Reject	Reject
EGARCH(1,1)	Reject	Reject
RiskMetrics	Reject	Reject

Note: This table reports the results of the minimally biased backtest for banks' internal models based on the number of breaches of various 10-day expected shortfall (ES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the ES forecasts observed are the true ES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. The total number of ES forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Next, the minimally biased backtest is applied to the 10-day 97.5% ES forecasts. As for the previous two backtests, the null hypothesis states that the observed 10-day 97.5% ES forecasts are the true 10-day 97.5% ES forecasts. In contrast to the preceding two backtests, however, this backtest is not conditional on the existence of at least one corresponding VaR breach and,

therefore, results are presented in Table 24, above, for all models and across both distributions. The null hypothesis is rejected across the board at the 97.5% confidence level, leading to the conclusion that all models employed do not produce statistically accurate 10-day 97.5% ES forecasts.

Finally, the last backtest employed is the Du-Escanciano backtest. This final backtest statistically tests the null hypothesis at the 97.5% confidence level, where the null hypothesis states that the underlying profit and loss distribution observed is the true profit and loss distribution. This backtest is not dependent on the forecasts produced by the various models, but, rather, it depends solely on the distributional assumption employed. Hence, the backtest's results are summarised in Table 25, where it can be seen that the backtest leads to the rejection of the null hypothesis for each of the distributions employed in this study.

Table 25: Results of the Du-Escanciano Backtest for the 10-day 97.5% Expected Shortfall Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Du-Escanciano	Reject	Reject

Note: This table reports the results of the Du-Escanciano backtest for banks' internal models based on the number of breaches of various 10-day expected shortfall (ES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The null hypothesis for the test states that the underlying profit and loss distribution observed is the true distribution. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. The total number of ES forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

The various backtests employed yielded conclusions rejecting the adequacy of the various distributions and models used to produce the 10-day 97.5% ES forecasts in this study. Table *26*, below, summarises the results of the forecasting error measures for the various models and distributions used to produce 10-day 97.5% ES forecasts using the normal distribution as the underlying statistical distribution. Table 27 then summarises the results of the same error measures for the various models employed in this study using the skewed Student's t distribution as the underlying statistical distribution.

Table 26's results show that the EGARCH model produced the most accurate 10-day 97.5% ES forecasts when using the normal distribution, apart from the MdAPE (that depends on ordered returns, and is expected to possibly conclude differently from the other three forecasting error measures). It is followed by the GARCH model, which is concluded to be the second-most accurate model by three of the four measures used (the RMSE being the fourth measure, ranking the GARCH model as the third-most accurate model). The table also shows

that the historical simulation model produced the least accurate forecasts based on all four measures when it comes to forecasting 10-day 97.5% ES forecasts.

Table 26: Forecasting Error Measures for the 10-day 97.5% Expected Shortfall Metric using Traditional Models (Normal Distribution)

	MAE	**RMSE**	**MAPE**	**MdAPE**
Historical Simulation	0.0917	0.0987	96.9729%	5.2595%
Delta-Normal	0.0683	0.0720	71.5508%	4.3951%
ARCH(1)	0.0589	0.0766	53.7785%	3.1535%
GARCH(1,1)	0.0559	0.0717	46.8418%	3.2586%
EGARCH(1,1)	0.0546	0.0669	45.6498%	3.3802%
RiskMetrics	0.0600	0.0705	51.3835%	4.1860%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day expected shortfall (ES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the normal distribution as the underlying distribution and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Table 27: Forecasting Error Measures for the 10-day 97.5% Expected Shortfall Metric using Traditional Models (Skewed Student's t Distribution)

	MAE	**RMSE**	**MAPE**	**MdAPE**
Historical Simulation	0.0917	0.0987	96.9729%	5.2595%
Delta-Normal	0.0683	0.0720	71.5508%	4.3951%
ARCH(1)	0.0704	0.0894	62.7869%	3.4777%
GARCH(1,1)	0.0696	0.0886	61.7423%	3.5731%
EGARCH(1,1)	0.2369	0.4394	263.9214%	4.9487%
RiskMetrics	0.0687	0.0798	61.6661%	4.3730%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day expected shortfall (ES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the skewed Student's t distribution as the underlying distribution and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Note that the values obtained for the various forecasting error measures for the historical simulation model and the delta-normal model are equal in both Table 26 and Table 27 due to their distribution agnostic nature. These figures are repeated for ease of reference.

When using the skewed Student's t distribution as the underlying statistical distribution, Table 27, surprisingly, shows that the delta-normal model produced the most accurate 10-day 97.5% ES forecasts using two of the four forecasting error measures, while the other two measures rank it as the fourth-most accurate model. Surprisingly still, the EGARCH model seems to offer incredibly poor 10-day 97.5% ES forecasts when using the skewed Student's t distribution, as it produces the least accurate forecasts using said distribution, in contrast to its most accurate status when using the normal distribution as the underlying statistical distribution. The historical simulation model, however, consistently ranks poorly across the models, as is the case here, where it is the second-least accurate model using three of the four forecasting error measures, and the least accurate model when considering the fourth measure.

Interestingly, and noteworthy, the rise in forecasting accuracy of the delta-normal model is not due to its improving forecasting abilities due to a change in distribution, since the model is distribution agnostic. In fact, it is the deterioration in the performances of the ARCH model, the GARCH model, and the EGARCH model when using the skewed Student's t distribution relative to the normal distribution that pushed the delta-normal model up the ranks when employing the latter distribution. This, once more, is possibly due to the distribution's worse tail fit relative to the normal distribution.

Last, attention is turned to the performances of the various models when applying the Diebold-Mariano test at the 99% confidence level to determine which models produce more accurate 10-day 97.5% ES forecasts. Table 28 shows the results of the Diebold-Mariano test for the normal distribution, while Table 29 shows the results of the test for the skewed Student's t distribution. As stated in the preceding sections, positive Diebold-Mariano statistics in Table 28 and Table 29 are accompanied by lower p-values, and suggest that the null hypothesis should be rejected, while the opposite is true for negative Diebold-Mariano statistics.

The results depicted in Table 28 are summarised as follows, as they pertain to 10-day 97.5% ES forecasting models using the normal distribution as the underlying statistical distribution.

i. Every model has superior forecasting abilities to the historical simulation model at the 1% significance level.

Table 28: Results of the Diebold-Mariano Test for the 10-day 97.5% Expected Shortfall Metric using Traditional Models (Normal Distribution)

	Diebold-Mariano Statistic	**p-Value**
Historical Simulation versus Delta-Normal	84.04	<2.2e-16
Historical Simulation versus ARCH(1)	22.61	<2.2e-16
Historical Simulation versus GARCH(1,1)	28.14	<2.2e-16
Historical Simulation versus EGARCH(1,1)	37.93	<2.2e-16
Historical Simulation versus RiskMetrics	42.58	<2.2e-16
Delta-Normal versus ARCH(1)	-4.23	1
Delta-Normal versus GARCH(1,1)	0.30	0.3818
Delta-Normal versus EGARCH(1,1)	5.75	4.691e-09
Delta-Normal versus RiskMetrics	2.24	0.0125
ARCH(1) versus GARCH(1,1)	7.89	1.774e-15
ARCH(1) versus EGARCH(1,1)	11.13	<2.2e-16
ARCH(1) versus RiskMetrics	6.00	1.069e-09
GARCH(1,1) versus EGARCH(1,1)	6.62	1.88e-11
GARCH(1,1) versus RiskMetrics	1.33	0.0919
EGARCH(1,1) versus RiskMetrics	-4.28	1

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day expected shortfall (ES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the normal distribution as the underlying distribution. The total number of ES forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

ii. The delta-normal model has equal forecasting abilities to the ARCH model, the GARCH model, and the RiskMetrics model[21], and has superior forecasting abilities to the historical simulation model, all at the 1% significance level. It has inferior forecasting abilities to the EGARCH model at the 1% significance level.

iii. The ARCH model has superior forecasting abilities to the historical simulation model, and has equal forecasting abilities to the delta-normal model, both at the 1%

[21] The p-value of this test is 0.0125.

significance level. It has inferior forecasting abilities relative to the GARCH model, the EGARCH model, and the RiskMetrics model.

iv. The GARCH model has superior forecasting abilities to the historical simulation model and the ARCH model, and has equal forecasting abilities to the delta-normal model and the RiskMetrics model, at the 1% significance level. It has inferior forecasting abilities relative to the EGARCH model at the 1% significance level.

v. The EGARCH model has superior forecasting abilities to the historical simulation model, the delta-normal model, the ARCH model, and the GARCH model, but has equal forecasting abilities to the RiskMetrics model, all at the 1% significance level.

vi. The RiskMetrics model has superior forecasting abilities to the historical simulation model and the ARCH model, and has equal forecasting abilities to the delta-normal model, the GARCH model, and the EGARCH model, all at the 1% significance level.

The forecasting error measures, as shown in Table 26, and the Diebold-Mariano test, the results of which are shown in Table 28, concur on the rankings of the various models used to produce 10-day 97.5% ES forecasts, when using the normal distribution as the underlying statistical distribution. Ranked as most accurate is the EGARCH model, while ranked as least accurate is the historical simulation model. The results depicted in the two tables, i.e., those showing the values of the various forecasting error measures used in this study and the results of the Diebold-Mariano test at the 1% significance level, show that the distribution agnostic models, i.e., the historical simulation model and the delta-normal model, rank as least accurate and second-least accurate, while the distributional-dependent models rank higher and, therefore, can be concluded to produce more accurate forecasts relative to the distribution agnostic models.

Last, the results depicted in Table 29 are summarised as follows, as they pertain to 10-day 97.5% ES forecasting models using the skewed Student's t distribution as the underlying statistical distribution.

i. The historical simulation model has equal forecasting abilities to the EGARCH model, and inferior forecasting abilities to the delta-normal model, the ARCH model, the GARCH model, and the RiskMetrics model, all at the 1% significance level.

Table 29: Results of the Diebold-Mariano Test for the 10-day 97.5% Expected Shortfall Metric using Traditional Models (Skewed Student's t Distribution)

	Diebold-Mariano Statistic	**p-Value**
Historical Simulation versus Delta-Normal	84.04	<2.2e-16
Historical Simulation versus ARCH(1)	6.36	1.062e-10
Historical Simulation versus GARCH(1,1)	6.99	1.488e-12
Historical Simulation versus EGARCH(1,1)	-10.13	1
Historical Simulation versus RiskMetrics	21.71	<2.2e-16
Delta-Normal versus ARCH(1)	-10.69	1
Delta-Normal versus GARCH(1,1)	-10.44	1
Delta-Normal versus EGARCH(1,1)	-10.38	1
Delta-Normal versus RiskMetrics	-8.93	1
ARCH(1) versus GARCH(1,1)	2.98	0.0014
ARCH(1) versus EGARCH(1,1)	-10.285	1
ARCH(1) versus RiskMetrics	7.01	1.297e-12
GARCH(1,1) versus EGARCH(1,1)	-10.29	1
GARCH(1,1) versus RiskMetrics	6.86	3.763e-12
EGARCH(1,1) versus RiskMetrics	10.33	<2.2e-16

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day expected shortfall (ES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the skewed Student's t distribution as the underlying distribution. The total number of ES forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

ii. The delta-normal model has equal forecasting abilities to all other models, except for the historical simulation model, which has equal inferior forecasting abilities relative to the delta-normal model, at the 1% significance level.

iii. The ARCH model has equal forecasting abilities to the delta-normal model and the EGARCH model at the 1% significance level, while it has inferior forecasting abilities when compared to the historical simulation model, the GARCH model, and the RiskMetrics model, also at the 1% significance level.

iv. The GARCH model has superior forecasting abilities to the historical simulation model and the ARCH model at the 1% significance level. It has equal forecasting abilities to the delta-normal model and the EGARCH model, but inferior forecasting abilities to the RiskMetrics model, all at the 1% significance level.

v. The EGARCH model has equal forecasting abilities to all models but the RiskMetrics model, where the RiskMetrics model has superior forecasting abilities relative to the EGARCH model, at the 1% significance level.

vi. The RiskMetrics model has superior forecasting abilities relative to the historical simulation model, the ARCH model, the GARCH model, and the EGARCH model at the 1% significance level, while it has equal forecasting abilities to the delta-normal model at that same significance level.

The results of the Diebold-Mariano test for the skewed Student's t distribution presented in Table 29 also correspond to the results of the same models when evaluating the various forecasting error measures, as presented in Table 27. The results presented in the two tables concur that the EGARCH model is the worst-performing model when it comes to producing 10-day 97.5% ES forecasts using the skewed Student's t distribution, while the delta-normal model produces the most accurate forecasts. Interestingly, in the case of using the skewed Student's t distribution as the underlying statistical distribution, the conclusion made regarding the relative performances of the distribution agnostic models versus the distribution-dependent models when using the normal distribution no longer applies, as evident by the preceding comment.

3.4.4. Stressed Expected Shortfall

This section discusses and analyses the results of the models employed in this study to forecast 10-day 97.5% SES forecasts for a US bank using the S&P 500 index as the underlying, over the period 15 March 1991 to 14 February 2020. First, the number of breaches experienced for each of the various models is presented and discussed, followed by an analysis of the backtesting results achieved using the BCBS's traffic light test, the conditional, unconditional, and minimally biased backtests. Finally, the forecasting errors are measured using a variety of error measures, as detailed in Section 3.3.4, and these results are statistically tested for their relative predictive ability using the Diebold-Mariano test.

The breaches observed for the various models used to forecast the 10-day 97.5% SES are summarised in Table 30, below. Given the very few breaches observed for the 10-day 97.5%

ES forecasts using the same models, as discussed in the preceding section, it is unsurprising that even fewer breaches were observed for the 10-day 97.5% SES metric, given that this metric is calculated over a stressed period, i.e., the most severe daily profit and loss figures were used to calibrate the models and obtain forecasts. Interestingly, almost all models employed using the skewed Student's t distribution as the underlying statistical distribution recorded no breaches at all, while the models experiencing the highest number of breaches were the GARCH model and the EGARCH model, both calibrated using the normal distribution as the underlying statistical distribution, achieving three breaches using either model.

Table 30: Breaches Observed for the 10-day 97.5% Stressed Expected Shortfall Metric using Traditional Models and the Normal and Skewed Student's t Underlying Distributions

	Normal Distribution	**Skewed Student's t Distribution**
	SES Breaches	SES Breaches
Historical Simulation	0	
Delta-Normal	0	
ARCH(1)	2	1
GARCH(1,1)	3	0
EGARCH(1,1)	3	0
RiskMetrics	2	0

Note: This table reports the number of breaches experienced for an application of various 10-day stressed expected shortfall (SES) forecasting models at the 97.5% confidence level using either a normal distribution or a skewed Student's t distribution as the underlying distribution, using the daily logged returns of the Standard & Poor's (S&P) 500 index from 15 March 1991 to 14 February 2020. A SES breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the SES forecast obtained via one of the models detailed in this table, where the SES forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SES forecasts for the study period was 7,286 per model. Note also that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Considering the very few breaches observed when using the normal distribution and the zero breaches observed for three of the models when using the skewed Student's t distribution and the distribution agnostic models relative to the 7,286-trading-day out-of-sample test period, it is not surprising that the results of the BCBS's traffic light test, as shown in Table 31, below, all result fall into the 'Green zone' for the 10-day 97.5% SES forecasts. Once again, it is possible to conclude that the BCBS's traffic light test's results convey little meaning when it comes to evaluating a model's performance in forecasting a relevant risk metric, as it is only useful in deciding whether a high number of breaches observed (if such a number is observed at all) is 'too high', or as a penalty system, as pointed out by Berkowitz, et al., (2009).

Table 31: Results of the Basel Committee on Banking Supervision's Traffic Light Test for the 10-day 97.5% Stressed Expected Shortfall Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Green zone	
Delta-Normal	Green zone	
ARCH(1)	Green zone	Green zone
GARCH(1,1)	Green zone	Green zone
EGARCH(1,1)	Green zone	Green zone
RiskMetrics	Green zone	Green zone

Note: This table reports the results of the Basel Committee on Banking Supervision (BCBS) Traffic Light test for banks' internal models based on the number of breaches of various 10-day stressed expected shortfall (SES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. A SES breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the SES forecast obtained via one of the models detailed in this table, where the SES forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SES forecasts for the study period was 7,286 per model. The number of breaches relative to the period length was then analysed using the BCBS's Traffic Light test to achieve one of three classifications. The classifications are 'Green zone' (if the binomial probability of the number of breaches relative to the number of trading days is greater than 99.99%), 'Yellow zone' (if the corresponding probability is lesser than 99.99% but greater than 99.99%), or 'Red zone' (if the corresponding probability is lesser than 95%). Note also that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

The conditional backtest, being the first of the three[22] backtests applied to the 10-day 97.5% SES forecasts, is only applicable when there is at least one corresponding 10-day SVaR breach observed. Recalling the number of observed 10-day 99% SVaR breaches observed for the various models, as shown in Table 12, it is the case that all but three of the models produce zero observed 10-day 99% SVaR breaches. Those models exhibiting at least one corresponding breach are the ARCH model, the GARCH model, and the EGARCH model, all of which used the normal distribution as the underlying statistical distribution. Hence, for all other models, the result of 'Cannot perform backtest' is captured in Table 32, below.

For the three models to which the conditional backtest can be applied, the null hypothesis, stating that the 10-day 97.5% SES forecasts are the true 10-day 97.5% SES forecasts, is tested

[22] The three backtests employed to test the 10-day 97.5% SES forecasts are the conditional backtest, the unconditional backtest, and the minimally biased backtest. The Du-Escanciano backtest is not dependent on the actual forecasts obtained and, therefore, its results are unchanged between the 10-day 97.5% ES forecasts and their stressed counterparts.

at the 97.5% confidence level. All three models lead to the rejection of the null hypothesis, implying that the forecasts produced are inaccurate.

Table 32: Results of the Conditional Backtest for the 10-day 97.5% Stressed Expected Shortfall Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Cannot perform test	
Delta-Normal	Cannot perform test	
ARCH(1)	Reject	Cannot perform test
GARCH(1,1)	Reject	Cannot perform test
EGARCH(1,1)	Reject	Cannot perform test
RiskMetrics	Cannot perform test	Cannot perform test

Note: This table reports the results of the conditional backtest for banks' internal models based on the number of breaches of various 10-day stressed expected shortfall (SES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the SES forecasts observed are the true SES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. The test can only be carried out if there is at least one stressed value at risk (SVaR) breach. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the models detailed in this table, where the SVaR forecast is calculated over the most period preceding the return's date, i.e., over a stressed period. The total number of SES forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

As for the conditional backtest, the unconditional backtest also requires at least one 10-day SVaR breach to be observed. Hence, once again, only the ARCH model, the GARCH model, and the EGARCH model using the normal distribution produce any result. The remaining models yield the result 'Cannot perform backtest' in Table 33, below.

For the three models to which the unconditional backtest can be applied, the null hypothesis, stating that the 10-day 97.5% SES forecasts are the true 10-day 97.5% SES forecasts, is tested at the 97.5% confidence level. As per the results of the conditional backtest, all three models lead to the rejection of the null hypothesis, implying that the forecasts produced are inaccurate.

Table 33: Results of the Unconditional Backtest for the 10-day 97.5% Stressed Expected Shortfall Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Cannot perform test	
Delta-Normal	Cannot perform test	
ARCH(1)	Reject	Cannot perform test
GARCH(1,1)	Reject	Cannot perform test
EGARCH(1,1)	Reject	Cannot perform test
RiskMetrics	Cannot perform test	Cannot perform test

Note: This table reports the results of the unconditional backtest for banks' internal models based on the number of breaches of various 10-day stressed expected shortfall (SES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the SES forecasts observed are the true SES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. The test can only be carried out if there is at least one stressed value at risk (SVaR) breach. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the models detailed in this table, where the SVaR forecast is calculated over the most period preceding the return's date, i.e., over a stressed period. The total number of SES forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Table 34: Results of the Minimally Biased Backtest for the 10-day 97.5% Stressed Expected Shortfall Metric using Traditional Models

	Normal Distribution	**Skewed Student's t Distribution**
	Backtest Result	Backtest Result
Historical Simulation	Reject	
Delta-Normal	Reject	
ARCH(1)	Reject	Reject
GARCH(1,1)	Reject	Reject
EGARCH(1,1)	Reject	Reject
RiskMetrics	Reject	Reject

Note: This table reports the results of the minimally biased backtest for banks' internal models based on the number of breaches of various 10-day stressed expected shortfall (SES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the SES forecasts observed are the true SES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using either a normal distribution or a skewed Student's t distribution as the underlying distribution. A SES breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the ES forecast obtained via one of the models detailed in this table, where the SES forecast is calculated over the most period preceding the return's date, i.e., over a stressed period. The total number of SES forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the number of breaches experienced does not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Finally, the minimally biased backtest is applied to the 10-day 97.5% SES forecasts. Once again, the null hypothesis here states that the observed 10-day 97.5% SES forecasts are the true

10-day 97.5% SES forecasts. Unlike the previous two backtests, this backtest is not conditional on the observation of corresponding 10-day 99% SVaR breaches and, therefore, can be applied to all models using both distributions. Table 34 summarises the results of the backtest, showing the universal rejection of the null hypothesis for all models using either distribution. Hence, it is concluded that no model of those employed produces statistically accurate 10-day 97.5% SES forecasts.

Throughout this section, the backtests employed produced little comfort surrounding the application of the various models to produce 10-day 97.5% SES forecasts, regardless of the underlying statistical distribution used. The next step is to assess the models' performances with respect to their forecasting errors, as captured by the MAE, the RMSE, the MAPE, and the MdAPE, as summarised in Table 35 for the normal distribution, and in Table 36 for the skewed Student's t distribution.

Note that the values obtained for the various forecasting error measures for the historical simulation model and the delta-normal model are equal in both Table 35 and Table 36 due to their distribution agnostic nature. These figures are repeated for ease of reference.

The results summarised in Table 35, below, show that the GARCH model produced the most accurate forecasts when using the normal distribution as the underlying statistical distribution, with three of the four forecasting error measures being the smallest for this model when compared with all other models (the fourth one being the MdAPE, for which the EGARCH model produced the smallest value). The GARCH model is then followed by the EGARCH model as the model that produced the second-most accurate forecasts, and the ARCH model as the model that produced the third-most accurate forecasts. The historical simulation produced the least accurate forecasts based on each of the four measures.

When using the skewed Student's t distribution as the underlying statistical distribution, the delta-normal model produced the most accurate forecasts, followed by the ARCH model (with the MdAPE being the only forecasting error measure that swaps the two models' ranks), as can be seen in Table 36. The historical simulation model produced the least accurate forecasts in this case, too.

Table 35: Forecasting Error Measures for the 10-day 97.5% Stressed Expected Shortfall Metric using Traditional Models (Normal Distribution)

	MAE	**RMSE**	**MAPE**	**MdAPE**
Historical Simulation	0.1700	0.1728	178.3905%	9.9945%
Delta-Normal	0.1056	0.1076	110.6901%	6.5699%
ARCH(1)	0.0973	0.1197	107.9476%	4.7967%
GARCH(1,1)	0.0834	0.0892	88.8212%	4.6616%
EGARCH(1,1)	0.0864	0.0935	92.7174%	4.5582%
RiskMetrics	0.1363	0.1533	147.1597%	6.0524%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day stressed expected shortfall (SES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020, where the SES forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the normal distribution as the underlying distribution and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

Table 36: Forecasting Error Measures for the 10-day 97.5% Stressed Expected Shortfall Metric using Traditional Models (Skewed Student's t Distribution)

	MAE	**RMSE**	**MAPE**	**MdAPE**
Historical Simulation	0.1700	0.1728	178.3905%	9.9945%
Delta-Normal	0.1056	0.1076	110.6901%	6.5699%
ARCH(1)	0.1115	0.1169	117.3426%	6.3889%
GARCH(1,1)	0.1337	0.1415	141.7565%	7.1026%
EGARCH(1,1)	0.1447	0.1534	150.6881%	7.5058%
RiskMetrics	0.1933	0.2105	206.2580%	8.1665%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day stressed expected shortfall (SES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020, where the SES forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the skewed Student's t distribution as the underlying distribution and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

When examining the various models with a focus on the change in underlying distribution, it is noteworthy that the performances of the GARCH model and the EGARCH model deteriorated significantly when using the skewed Student's t distribution rather than the normal distribution. As stated previously, this may be indicative of the poor tail fit offered by the

skewed Student's t distribution relative to the normal distribution as the underlying statistical distribution when producing 10-day 97.5% SES forecasts.

Finally, the Diebold-Mariano test was performed for the various models used to produce 10-day 97.5% SES forecasts, with the result of these tests, tested at the 99% confidence level, summarised in Table 37 for the normal distribution, and Table 38 for the skewed Student's t distribution. As previously stated, positive Diebold-Mariano statistics in Table 37 and Table 38 are accompanied by lower p-values, and suggest that the null hypothesis should be rejected, while the opposite is true for negative Diebold-Mariano statistics.

The results depicted in Table 37 are summarised as follows, as they pertain to the 10-day 97.5% SES forecasting models using the normal distribution as the underlying statistical distribution.

i. Every model has superior forecasting abilities relative to the historical simulation model at the 1% significance level.
ii. The delta-normal model has superior forecasting abilities relative to the historical simulation model, equal forecasting abilities relative to the ARCH model and the RiskMetrics model, and inferior forecasting abilities relative to the GARCH model and the EGARCH model, all at the 1% significance level.
iii. The ARCH model has superior forecasting abilities relative to the historical simulation model, equal forecasting abilities relative to the ARCH model and the RiskMetrics model, and inferior forecasting abilities relative to the GARCH model and the EGARCH model, all at the 1% significance level.
iv. The GARCH model has superior forecasting abilities relative to all other models at the 1% significance level.
v. The EGARCH model has inferior forecasting abilities relative to the GARCH model and equal forecasting abilities to the RiskMetrics model, both at the 1% significance level. It has superior forecasting abilities relative to all other models at the 1% significance level.
vi. The RiskMetrics model has superior forecasting abilities relative to the historical simulation model and inferior forecasting abilities relative to the EGARCH model, both at the 1% significance level. It has equal forecasting abilities relative to all other models at the 1% significance level.

Table 37: Results of the Diebold-Mariano Test for the 10-day 97.5% Stressed Expected Shortfall Metric using Traditional Models (Normal Distribution)

	Diebold-Mariano Statistic	**p-Value**
Historical Simulation versus Delta-Normal	245.86	<2.2e-16
Historical Simulation versus ARCH(1)	44.32	<2.2e-16
Historical Simulation versus GARCH(1,1)	333.40	<2.2e-16
Historical Simulation versus EGARCH(1,1)	392.05	<2.2e-16
Historical Simulation versus RiskMetrics	52.47	<2.2e-16
Delta-Normal versus ARCH(1)	-7.90	1
Delta-Normal versus GARCH(1,1)	253.52	<2.2e-16
Delta-Normal versus EGARCH(1,1)	109.22	<2.2e-16
Delta-Normal versus RiskMetrics	-62.44	1
ARCH(1) versus GARCH(1,1)	18.42	<2.2e-16
ARCH(1) versus EGARCH(1,1)	16.175	<2.2e-16
ARCH(1) versus RiskMetrics	-23.81	1
GARCH(1,1) versus EGARCH(1,1)	-51.21	1
GARCH(1,1) versus RiskMetrics	-85.76	1
EGARCH(1,1) versus RiskMetrics	-88.27	1

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day stressed expected shortfall (SES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020, where the SES forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the normal distribution as the underlying distribution. The total number of SES forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

The results of the Diebold-Mariano test for the normal distribution presented in Table 37 correspond to the results of the same models when evaluating the various forecasting error measures, as presented in Table 35. The results depicted in the two tables concur that the model that produced the least accurate 10-day 97.5% SES forecasts is the historical simulation model, followed by the RiskMetrics model as the second-least accurate. On the other hand, the various forecasting error measures and the Diebold-Mariano statistical test concur that the model that produced the most accurate forecasts is the GARCH model.

Table 38: Results of the Diebold-Mariano Test for the 10-day 97.5% Stressed Expected Shortfall Metric using Traditional Models (Skewed Student's t Distribution)

	Diebold-Mariano Statistic	**p-Value**
Historical Simulation versus Delta-Normal	245.86	<2.2e-16
Historical Simulation versus ARCH(1)	226.96	<2.2e-16
Historical Simulation versus GARCH(1,1)	256.28	<2.2e-16
Historical Simulation versus EGARCH(1,1)	44.31	<2.2e-16
Historical Simulation versus RiskMetrics	-62.62	1
Delta-Normal versus ARCH(1)	-33.55	1
Delta-Normal versus GARCH(1,1)	-93.86	1
Delta-Normal versus EGARCH(1,1)	-74.23	1
Delta-Normal versus RiskMetrics	-110.15	1
ARCH(1) versus GARCH(1,1)	-82.61	1
ARCH(1) versus EGARCH(1,1)	-62.02	1
ARCH(1) versus RiskMetrics	-113.84	1
GARCH(1,1) versus EGARCH(1,1)	-24.40	1
GARCH(1,1) versus RiskMetrics	-114.09	1
EGARCH(1,1) versus RiskMetrics	-80.47	1

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day stressed expected shortfall (SES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020, where the SES forecast is calculated over the most period preceding the return's date, i.e., over a stressed period. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the skewed Student's t distribution as the underlying distribution. The total number of SES forecasts for the study period was 7,286 per model. Note that the historical simulation model and the delta-normal model are distribution agnostic, meaning that the forecasted values do not vary with the use of either the normal distribution or the skewed Student's t distribution as the underlying distribution.

The results summarised in Table 38 are summarised as follows, as they pertain to 10-day 97.5% SES forecasting models using the skewed Student's t distribution as the underlying statistical distribution.

i. The historical simulation model has forecasting abilities equal to those of the RiskMetrics model at the 1% significance level, while all other models have superior forecasting abilities relative to it at the 1% significance level.

ii. All model other than the RiskMetrics model have superior forecasting abilities relative to the historical simulation model and equal forecasting abilities relative to all other models at the 1% significance level.

iii. The RiskMetrics model has equal forecasting abilities relative to all other models at the 1% significance level.

Interestingly, the Diebold-Mariano tests' results for these forecasts are not as useful as all other Diebold-Mariano test results, as they either indicate a clear superiority of forecasting or no superiority at all, and almost focus to highlight the poor performances of the historical simulation model and the RiskMetrics model in producing 10-day 97.5% SES forecasts. These results concur with the conclusions reached using the forecasting error measures, as shown in Table 36.

3.5. Conclusion

This chapter provided the basis for this study by examining and evaluating the performances of various traditional market risk models when producing 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts, and their stressed counterparts. These forecasts were produced using historical return data for the S&P 500 index from 15 March 1991 to 14 February 2020, mimicking the market risk management of a US bank's equities desk. This basis is later used to assess the performances of BNs in performing the same task.

All models assessed in this chapter yielded very few breaches, and all models were classified as 'Green zone' models using the BCBS's traffic light test. Hence, it is prudent to conclude that the BCBS's traffic light test is of little use in the practical testing of the accuracy of either VaR forecasts, ES forecasts, or their stressed versions. This chapter's results seem to concur with the literature stating that, for banks, few breaches are preferred to no breaches (see, for example, McAleer and da Veiga, 2008).

The other statistical backtests used to evaluate the performances of the various traditional models when producing all four market risk metrics yielded little additional insight into the suitability of the models in producing the necessary market risk forecasts. This study finds that the forecasts produced are generally conservative, with few breaches observed for the various models considered across the four market risk measures. Hence, this chapter's findings support those of studies such as Berkowitz and O'Brien (2002), Pérignon, et al., (2008), Berkowitz, et al., (2009), Pérignon and Smith (2010), and O'Brien and Szerszeń (2017), among others, by concluding that the models employed by (US) banks to calculate 10-day 99% VaR forecasts

produce conservative forecasts. Moreover, this finding is extended to 10-day 99% SVaR forecasts, 10-day 97.5% ES forecasts, and 10-day 97.5% SES forecasts.

Across the models used to produce 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts, the EGARCH model using the normal distribution as the underlying statistical distribution produced the most accurate forecasts. This result supports the findings of studies suggesting that the use of autoregressive models may produce more accurate 10-day 99% VaR forecasts, such as that of O'Brien and Szerszeń (2017). This may be explained by the tendency of autoregressive models to better capture volatility clustering and leptokurtosis (Angelidis, Benos, & Degiannakis, 2004), even when the normal distribution is used as the underlying statistical distribution, although the level of leptokurtosis induced by the autoregressive model may not capture that present in the return data (Angelidis, et al., 2004). Due to the general tendency of US banks to use the historical simulation model, it is noted that it ranked either the least accurate model or the second-least accurate model across the ten models considered for each market risk metric.

When it comes to the models used to produce 10-day 99% SVaR forecasts and 10-day 97.5% SES forecasts, the GARCH model using the normal distribution as the underlying statistical distribution produced the most accurate forecasts. Once again, this echoes the results above, as well as the literature stating that the use of autoregressive models may produce more accurate market risk forecasts. This historical simulation model consistently ranked in the bottom half of models in terms of accuracy for these two market risk metrics.

This study contributes to the literature by providing a comprehensive assessment of the performances of traditional models to produce market risk forecasts, and the conclusions above, especially with respect to the stressed metrics, are a novel contribution to the literature. Another novel contribution is the assessment of said models, especially the stressed metric models, with respect to the two statistical distributions used as underlying statistical distributions when producing the market risk metric forecasts.

With respect to the performances of the two distributions in producing 10-day 97.5% ES forecasts and 10-day 97.5% SES forecasts, the normal distribution and the skewed Student's t distribution were statistically backtested to assess their fit to the underlying profit and loss distribution using the Du-Escanciano backtest. This backtest rejected its null hypotheses at the 2.5% significance levels and concluded that the distributions did not provide good fits to the underlying profit and loss data.

This chapter's results highlight the poor performance of the skewed Student's t distribution when it comes to producing the market risk metrics' forecasts in this study. The results above highlight the poor fit offered by the skewed Student's t distribution to the profit and loss account of a generic equities trading desk of a US bank, in contrast to studies suggesting that distributions with higher skew relative to the normal distribution may produce better results, such as those of McNeil and Frey (2000), and Wong, et al., (2012). While the literature indicates that the distribution itself may fit the return data better, due to the different values of skew and kurtosis relative to the normal distribution, it is the tail that is of interest in this study for the purposes of calculating tail-based market risk metrics. Using all four market risk metrics considered in this study, the forecasts produced using the normal distribution proved to be more accurate than their counterparts produced using the skewed Student's t distribution as the underlying statistical distribution. This suggests that the tail fit of the skewed Student's t distribution may be poorer than that of the normal distribution. Specifically, in the case of this study, the fatter tail of the skewed Student's t distribution relative to that of the normal distribution provides less accurate market risk forecasts.

The primary conclusion of this chapter relates to the superior performance of autoregressive models when used to produce market risk metrics. For both 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts, the EGARCH model produced more accurate forecasts relative to other models. For the stressed metrics, both exhibited more accurate forecasts when employing GARCH models. Both specifications use past return data to produce forecasts of return volatility, using some weighting mechanisms.

This finding suggests that models that incorporate forward-looking forecasts perform better in the market risk setting than those that use backwards-looking data exclusively, as is the case for the historical simulation model, for example. Taking this a step further, there is a strong case to be made to incorporate forward-looking forecasts beyond simply the volatility of returns, and actually examine the performances of models that incorporate the forecasts of the returns themselves. BNs have been increasingly used in financial settings, with a recent relevant application being that of Apps (2020). Apps uses a simplified methodology to forecast VaR using a BN, the details of which are discussed further in Section 4.2. Hence, the incorporation of a forward-looking methodology to the market risk field can be achieved using BNs to forecast the closing values of the S&P 500 index. These forward-looking network predictions can further be used to produce market risk metrics, as is discussed in the next chapter.

4. Market Risk Management using Bayesian Networks

This chapter introduces Bayesian networks (BNs) to the market risk framework introduced by the Basel Committee on Banking Supervision (BCBS). 10-day 99% value at risk (VaR) forecasts and 10-day 97.5% expected shortfall (ES) forecasts, and their stressed counterparts, were forecasted using BNs applied to the equities trading desk of a theoretical bank based in the United States (US), highlighting the usefulness and advantages of applying a BN to forecast market risk metrics such as VaR and ES.

The previous chapter focused solely on traditional models when producing 10-day 99% VaR forecasts, 10-day 97.5% ES forecasts, and their stressed counterparts. It was shown that the models that produced the most accurate market risk metric forecasts incorporated some element of forecasting, even if said forecasting was limited to the volatility of returns, as displayed for the autoregressive models used. Moreover, recent studies, such as that by Apps (2020), show that the use of BNs to produce VaR forecasts has become feasible with advances in computing power and improved network learning algorithms. This chapter builds on the work presented by Apps and other authors in the field by developing an extensive methodology for the construction of BNs in the context of producing market risk metric forecasts.

The use of BNs to produce market risk forecasts across both VaR and ES, and their stressed counterparts, is a novel contribution to the literature, given that BN applications in the context of market risk have been limited to VaR (see, for example, Apps, 2020). Hence, the introduction of BNs to produce 10-day 99% VaR forecasts, 10-day 97.5% ES forecasts, and their stressed counterparts, is a novel contribution to the literature.

Moreover, studies such as those of Apps (2020) considered a simplified methodology when applying BNs to VaR. Hence, another novel contribution of this study is a comprehensive specification of BNs, using a variety of learning algorithms, to establish the suitability and usefulness of BNs when producing market risk metrics in the literature.

Therefore, this chapter establishes the performances of different learning algorithms when producing market risk metrics, and the performances of said forecasts when backtested in the context of quantifying and managing the market risk of the equities desk of a US bank. The performances of these network learning algorithms are used in the following chapter to introduce a novel methodology combining the results of this and the previous chapter.

The remainder of this chapter is structured as follows. First, this chapter provides a brief overview of the probability theory that is the foundation of Bayesian statistics. The field of BNs is a specific application of the broader field of machine learning. Hence, a general overview of machine learning is provided to establish a basic understanding of machine learning concepts and techniques before addressing BNs in detail. Then, this chapter introduces the theory behind BN construction, including data preparation, node selection, model evaluation, learning, and inference. In addition, some key algorithms are discussed in detail, as these were used to either construct or train the network. A literature review of BNs as used in asset pricing is also provided, highlighting the increased importance and use of machine learning and BNs in finance. The data and methodology employed in Chapter 3, with specific reference to the backtests employed to statistically test the accuracy of the traditional market risk models, are examined in this chapter in the context of the BN used to produce the relevant market risk metrics. The forecasting error measures and the Diebold-Mariano test, also introduced in Chapter 3, are also applied to the application of the BN to produce the relevant market risk metrics. The results of the applications of BNs to produce the various market risk metrics are discussed and analysed. Finally, a conclusion surrounding the performances of the various models is provided.

4.1. Construction of Bayesian Networks

This section opens with an introduction to Bayesian statistics and machine learning. It then introduces the fundamental notions surrounding the construction of BNs and discusses the various network-learning algorithms employed in this study.

It is often asked whether information that relates to a specific event affects our perception of other events. This uncertainty regarding the effects of certain events on the outcomes of others is a valuable insight when analysing various scenarios. While the relation between events and the update of our belief system as new information becomes available is a natural task for the human brain, it is not always so simple for a computer. BNs, however, are a tool used to mimic this process.

4.1.1. Probability Theory and Bayesian Statistics

The necessary foundational probability theory, as it relates to conditional probabilities and distributions, is introduced in this section. The construction of BNs is discussed in detail later in this section.

The probability of an event A taking place is defined as the number of occurrences of A divided by the total number of occurrences of all outcomes in an experiment. This probability is denoted by $\Pr[A]$.

As mentioned earlier, it is not uncommon for some events to be dependent on other events, i.e., for the probability of some event taking place to depend on the outcomes of other related events. Hence, a conditional probability can be measured as follows: Given that event B has taken place, what is the probability of event A as influenced by event B? If this probability is defined to be equal to ν, then $\Pr[A|B] = \nu$, i.e., the probability of event A taking place given that event B has occurred is equal to ν.

The intersection of two events, A and B, is be defined as the event in which both events A and B have occurred. This is denoted by $A \cap B$ (or A, B). The probability of such an event is defined as follows: $\Pr[A \cap B] \coloneqq \Pr[A|B] \times \Pr[B]$.

It is clear that the intersection of event A with event B is equal to the intersection of event B with event A, i.e., $A \cap B = B \cap A$. This implies that the probabilities of the two intersections occurring are also equal, i.e., $\Pr[A \cap B] = \Pr[B \cap A]$.

Therefore, by applying a simple substitution, the following is found.

$$\begin{aligned} & \Pr[A \cap B] = \Pr[B \cap A] \\ \Leftrightarrow\ & \Pr[A|B] \times \Pr[B] = \Pr[B|A] \times \Pr[A] \\ \Leftrightarrow\ & \Pr[B|A] = \frac{\Pr[A|B] \times \Pr[B]}{\Pr[A]} \end{aligned} \tag{38}$$

This is Bayes' Theorem for conditional probability.

It is concluded that event A and event C are independent if and only if $\Pr[A|C] = \Pr[A]$. Moreover, it is concluded that event A is independent of event C given event B if $\Pr[A|B, C] = \Pr[A|B]$.

4.1.2. Overview of Machine Learning

Machine learning is the scientific field of developing computerised systems that automatically learn from their own experiences and improve their own processes (Mitchell, 2006). It can be thought of as a combination of the fields of statistics and computer science. The former's attention is directed at the inference of results from data and distribution assumptions, while the latter's attention is directed at manual problem-solving techniques

through the use of code. Machine learning is the product of the marriage of the two fields and the goal of machine learning combines those of its parents (Mitchell, 2006).

Put simply, machine learning is the study of data-driven methods used to make automatic and self-improving decisions based on the processing of data. It is often used in the solving of data mining tasks and the enhancement of decisions made by the humans employing these methods.

A major advantage of machine learning is its ability to handle the complexity involved in some tasks – an ability that is not inherently possessed by humans. Machine learning allows its user to make both explicit and implicit decisions – these are available to the user even if some data points are missing, depending on the machine learning algorithm employed. If it were not for machine learning, a complicated procedure to handle the missing data would have had to be developed due to the incapability of the human brain to deal with missing data, an issue which a computer can handle quite well.

Generally speaking, a machine learning application makes use of labelled datasets, a training dataset and a validation dataset, as learning tools for the extraction of useful information from the data (Suthaharan, 2014). The former set is used for training the machine learning application to recognise patterns, while the latter is used to validate that the patterns recognised are accurate (or, at least, mostly correct).

Consider a set $\mathcal{S} = \{(x^n, y^n), n = 1,2, \dots, N\}$ of data two-ple (2-tuple) ordered pairs, where x is an input and y is the desired output. Supervised learning is the process of learning the relationship between the input x and output y, given the input, x. The model learned by the computer then receives input $x^* \notin \mathcal{S}$ and output $y^* \notin \mathcal{S}$. The relationship between x^* and y^* is assumed to follow the relationship learned from the generated set $\mathcal{S}$. Of interest is, therefore, the conditional probability distribution $p(x|y, \mathcal{S})$ (Barber, 2012).

As an example of supervised learning, consider a classroom of 20 students. Each student achieves a certain grade for a task. This combination of students and grades is the dataset $\mathcal{S}$, where each student's characteristics are the input x (which is a vector, in this case) and each student's grade is the output y. Should a new student be introduced to the class with characteristics x^*, supervised learning is the study of the relationship in the set $\mathcal{S}$ in order to conclude a likely grade y^* for the new student.

In supervised learning, the difference between classification problems and regression problems is usually considered. If the output of the supervised learning task is one of several classes, i.e., discrete, then the problem is said to be one of classification. If, on the other hand, the output is one of a continuous nature, then the problem is said to be one of regression (Barber, 2012).

Unsupervised learning, as opposed to supervised learning, does not feed the set of outputs y into the learning method. As an example, consider a set $\mathcal{U} = \{x^n, n = 1,2, \dots, N\}$. Since there are no output values provided, the issue of modelling the distribution of x, $p(x)$, is now at hand, rather than the conditional distribution $p(x|y, \mathcal{S})$ as with supervised learning (Barber, 2012). The goal of unsupervised learning, put simply, is to group various data points into groups of similar characteristics.

As an example of unsupervised learning, consider the classroom example above. Each student has a vector of characteristics, x, just as before. The unsupervised learning algorithm can now group students of similar characteristics, for example, into groups of tall students and short students. Another grouping may be based on gender. The learning algorithm can obtain many different groupings based on what characteristics are available to it.

A commonly applied branch of machine learning is that of artificial neural networks (ANNs). Since the early 1940s, humans have been exploring the inner workings of the brain and its neurons. As early as 1959, Bernard Widrow and Marcian Hoff of Stanford applied the first neural network to a real-world problem, sparking an interest in the application of such networks to various problems.

An ANN is a directed graph whose nodes act like artificial neurons. Nodes are connected as layers to other layers utilising directed edges (Jain & Mao, 1996). The first layer is called the input layer while the last layer is called the output layer. ANNs mostly produce discrete outputs.

Today, common applications of ANNs include pattern recognition (used often by the United States Postal Services to redirect mail), forecasting weather patterns, phrase detection and completion in search engines, classification of emails as spam, and many more. When it comes to financial applications, ANNs can be used to forecast and predict financial data (for an early example of the application of ANNs to the pricing and hedging of options, see Hutchinson, Lo, and Poggio, 1994; for a more recent example, see Nunes, Gerding, McGroarty, and Niranjan, 2018).

While ANNs and BNs share some similarities (both are directed graphs), they are different. The most fundamental difference is that, when using BNs, the graphical representation created is crucial to the construction of the network and has a meaning when it comes to the conditional dependencies of the various nodes on other nodes. An ANN's structure does not make conditional dependencies explicit. Moreover, an ANN's output is often a point estimate, while the output of a BN is a probability density function (PDF), which is better suited to produce tail forecasts or values at different levels of confidence (e.g., 95% and 99%).

The conditional dependencies between the different nodes are of the utmost importance when the effects of different events on financial returns are to be modelled, given the degree of dependency between the two. Hence, it is a simple choice to employ BNs as opposed to ANNs in the pricing of portfolios and the calculation of market risk metrics for said portfolios, as BNs allow for the calculation of both VaR and ES at various levels of confidence given the PDF outputs.

4.1.3. Defining a Bayesian Network, the Markov Condition, and Markov Equivalence

A BN, in its basic mathematical structure, is a causal network. A causal network is a collection of variables (or nodes) which are linked to each other by a set of directed links (edges or arcs). Using mathematical terminology, this is referred to as a directed graph (Jensen, 1996). As mentioned earlier, a BN is a directed acyclic graph (DAG). The graph is made up of a series of nodes and edges, where each node represents a variable with a set of mutually exclusive propositions referred to as states. These nodes may have values that are either observed (in which case they are referred to as evidence nodes) or not observed (in which case they are referred to as latent or hidden nodes). The values of hidden nodes are not directly observed but, rather, are inferred. Conditional dependencies between nodes are represented by the use of edges (the directional lines connecting nodes) such that the edge between nodes A and B, directed from A to B, indicates that the value or state of B is dependent on that of A. The characteristic that makes a BN a DAG is the fact that there is not a single edge that starts and ends at the same node, and the process does not allow the return to any node once the process leaves that node.

The causal relationships depicted by the structures of BNs allow for various types of reasoning to be performed based on said structure. One type of reasoning that can be performed is diagnostic reasoning, whereby the symptoms are examined to infer the cause. Diagnostic

reasoning follows the reverse direction of the edge between two nodes. The other type of reasoning that can be performed is deductive (or predictive) reasoning, where the cause is first examined, and then the likely symptoms are deduced (Korb & Nicholson, 2004).

When examining a directed graph, the use of descriptive terms is often made to describe the family of nodes. If there exists a directed edge from node P to node C, then node P is known as the 'parent' of node C, and node C is the 'child' of node P (Stephenson, 2000). This relationship can be extended further: If there exists a relationship that can be traced between two nodes, an originating node and a terminating node, the originating node is referred to as the 'ancestor' of the terminating node, while the terminating node is referred to as the 'descendant' of the originating node.

The topology of a BN makes use of conditional probability tables (CPTs). A CPT is used for each node to describe the node's local probability distribution. Each node's CPT is used to describe the node's possible states given the range(s) of states of the node's parent(s). Nodes can be either continuous or discrete. For discrete nodes, multinomial distributions parameterised by a set of probability vectors are employed. On the other hand, if the node is continuous, a normal distribution with a mean calculated as a linear combination of the node's parents' states can be used (Lauritzen, 1996). A node that has no causal relationships to any ancestry nodes is referred to as parentless and has an unconditional local probability distribution.

Since the structure of the network is crucial, the probability distributions of each ancestor node must first be specified before proceeding to the various descendant nodes. The global joint distribution function of the whole network can be obtained by taking the product of the individual local distributions over all nodes in the belief network.

Some notation is now introduced. A BN, B, is defined to be of the form $B(V, E)$, where V represents the nodes of the graph and E represents its edges. For each node $\{v_i \in V\}$, a set of states for its parent nodes can be defined to be π_i. Hence, $P(v_i|\pi_i)$ can be used to denote the joint probability for a given value occurring at node v_i, conditional on the value of its parent nodes, π_i. This, in turn, adequately captures the causal relationships found in the BN.

To find the global joint distribution of a network of n nodes as a whole, the product of the individual joint distribution can be calculated as follows.

$$P(v_1, v_2, \dots, v_n) = \prod_{i=1}^{n} P(v_i|\pi_i) \tag{39}$$

While BNs can be described and characterised by their conditional dependencies, they can also be described by their conditional independencies as those simplify the computation of the network's global joint probability distribution. The independencies that exist within the belief network are captured by local relations within the network (Pearl & Russell, 2001) as well as by the use of dependence-separation (commonly known as d-separation[23]). d-separation is a formal graphical property that is used as a criterion to determine the independence of a set of variables X from another set of variables Y, given a third set of variables Z. The idea behind d-separation is that d-separation allows one to associate 'independence' with 'separation' in a causal graph. Hence, d-separation is used as a more formal method of determining independence in the topology of the network. The Markov condition[24] states that every node in a BN is conditionally independent of all of its non-descendants, given its parents (Cowell, Dawid, Lauritzen, & Spiegelhalter, 1999). This property is also described by the term 'local semantics' (Pearl & Russell, 2001). The term 'causal Markov condition' is used to describe the situation where a BN accurately models the causality between the network's nodes.

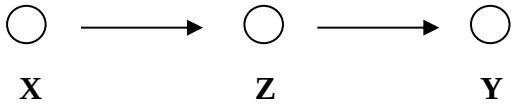
There is much debate in the literature surrounding the strict definition of causality and its use. Suppes (1970) suggests that a relationship can be classified as causal if (i) the variables are correlated; (ii) there exists some kind of temporal asymmetry in the variables; and (iii) no hidden variable capable of explaining the correlation of the two variables exists. The debate builds on to take into account the sometimes 'loose' use of the word 'causality' in the field of BNs as it relates to the causal Markov condition being an assumed property of the networks (Lemmer, 1996). While a BN makes use of relationships and statistical properties, it is dependent on the application of the expertise of its creator. Hence, a BN cannot be thought of as an absolute structure when considering its comprehensibility. The existence of at least one hidden layer of explanatory nodes that the model fails to capture cannot be ruled out, thereby missing out on being a causal structure from a statistical standpoint. Hence, it is worthwhile to

[23] Also known as 'the directed global Markov condition' (Lauritzen, 1996).

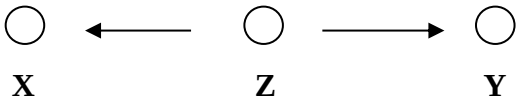
[24] Also known as 'the local directed Markov condition' or 'the paternal Markov condition'.

point out that the analysis that follows in this study, while based on sound statistical and mathematical applications, might not be causal in the strictest sense of the word.

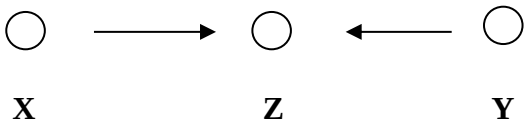
Figure 3: d-Separation



A chain. X is conditionally independent of Y given Z.



A fork. X is conditionally independent of Y given Z.



A collider. X and Y are marginally independent but become conditionally dependent once Z is known.

Note: This figure depicts the principle of d-separation. d-separation is a concept borrowed from graph theory. It depicts the independence of (sets of) variables graphically in a directed acyclic graph (DAG). Two variables, X and Y, are said to be d-separated by a third variable, Z, in a DAG G if the variables form a chain, a fork, or a collider, as depicted in this figure.

d-separation is often defined as the negation of dependence-connection, or d-connection. d-connection is defined as follows: Let G be a DAG with three disjoint sets of variables, X, Y, and Z. The sets X and Y are said to be d-connected by the set Z in G if and only if there exists an edge (which need not be directed) between the set X and the set Y. The sets X and Y are said to be d-separated by the set Z if and only if they are not d-connected by Z in the graph G. Figure 3, above, illustrates a graphical depiction of d-separation.

The notation $\langle X|Y|Z\rangle_D$ is used to represent the principle of d-separation. The principle is one of a graphical description of conditional independence between (sets of) variables. As mentioned earlier, the d-separation principle allows for the independence of a set of variables X from another set of variables Y to be determined, given a third set Z. This is true if:

i. a chain is formed between the sets X and Y, with the set Z in between them, i.e., there is a directed edge from X to Z and another from Z to Y;
ii. a fork is formed with set Z being the parent node of sets X and Y, i.e., there exist directed edge from Z to each X and Y; and

iii. a collider is formed with set Z being the child node of sets X and Y, i.e., there exists a directed edge from X to Z and another from Y to Z.

A collider depicts the conditional dependence of two variables once the collider variable's value is known. For example, in Figure 3, X and Y are independent of one another, when the value of Z is unknown. However, should Z's value become known, then X and Y are conditionally dependent on one another.

By making use of the d-separation principle and the Markov condition, all of a BN's conditional independencies can be identified. However, this does not rule out the possibility of the existence of a hidden node not captured by the d-separation principle, i.e., this node may explain some of the conditional dependencies that exist in the BN. Hence, it is further assumed that BNs satisfy what is called the faithfulness condition.

The faithfulness condition states that in order for an acyclic causal structure (in this case, a DAG) to be considered faithful, the probabilistic independencies as explained by the causal Markov condition are the only probabilistic independencies present in an acyclic causal structure (Tsamardinos & Aliferis, 2003; Steel, 2006).

A more mathematical definition is as follows: A graph G of a BN is considered faithful to a joint probability P over a set of variables X if every dependence entailed by G is also present in P. Conversely, a distribution P is said to be faithful over a set of variables X if there exists a DAG G satisfying the faithfulness condition.

The DAG of a belief network can be associated with the joint global probability distribution of the network if both the concept of faithfulness and the Markov condition can be employed.

The last principle discussed in this section is that of Markov equivalence. Regardless of whether two graphs A and B are both directed or undirected, if they have the same set of conditional independencies within their structures, then graph A and graph B are referred to as Markov equivalent.

The principle of Markov equivalence is crucial to the construction of BNs as it eliminates the dilution of the validity of the BN used, should there exist other Markov equivalent BNs, i.e., there exists a theoretical limit on structure learning from the data.

4.1.4. Learning the Network Structure

The process of learning the structure of a BN is often undertaken by employing at least one of the two main approaches – a search-and-score approach or a constraint-based approach. Both approaches are used to learn both the structure of the network, i.e., the DAG, as well as the network's parameters, i.e., the causal relationships between the nodes. There exists a trade-off between the computational cost of learning the structure of the belief network and the structure's accuracy. While a brute-force search for the structure may be comprehensive, it is computationally expensive. This means that there is room for the application of different algorithms to determine an approach that is both computationally feasible and, at the same time, accurate.

The search-and-score approach is employed to maximise some scoring function which, in turn, indicates how well the network constructed fits the data. The constraint-based approach is employed to determine the presence of conditional independencies between the nodes of the belief network. These are often assumed to exist before being tested and removed by the algorithm employed. Finally, a combination of the two approaches may be employed in what is called a hybrid approach. The algorithms explored in this study are hybrid algorithms, employing a constraint-based approach while attempting to maximise the chosen network score.

Robinson (1977) derives a recursive function to compute the number of possible n-node belief network structures. The formula is as follows.

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i) \tag{40}$$

Hence, it can be shown that the number of 5-node belief network structures is 29,000, while the number of 10-node belief network structures is around 4.2×10^{18}. The issue of a large number of available n-node belief structures is mitigated by the application of the Markov equivalence principle, as discussed earlier (see Section 4.1.3).

Once a network is fitted to the data, its goodness-of-fit must be evaluated using some scoring metric. The Akaike information criterion (AIC) is a relative estimator depicting the quality of the fit of a statistical model to a dataset. It is relative as it compares the goodness-of-fit of a model relative to those of other models available. This makes this estimator particularly useful

for the comparison of different models and, hence, model selection. The AIC does, in fact, offer a metric that discourages overfitting of the data.

The AIC is calculated as follows.

$$AIC = 2v - 2\ln(\hat{L}) \tag{41}$$

where $\hat{L}$ represents the maximised likelihood function value for the model under consideration and v is the number of estimated parameters in the model.

The AIC values of the various learned network models can then be compared. The ideal is to minimise the AIC values and then rank the different models. The model with the lowest minimised AIC score is then chosen, given the learning algorithm employed, as it suggests that the model fits the data best.

The four learning algorithms explored in this study are the genetic algorithm, the Peter and Clark (stable) algorithm, the max-min hill-climbing algorithm, and the semi-interleaved HITON parents and children algorithm. A brief overview of each is given below.

4.1.4.1. The Genetic Algorithm

Genetic algorithms were developed by John Holland in the 1960s and are often employed to solve problems whose goals are to find the best attainable solution out of a large domain of possible solutions (Mitchell, 1996). The original concept of a genetic algorithm as presented by Holland (1975) tries to mimic the biological processes often encountered in nature; specifically crossover, mutation, and inversion. The algorithm was designed as a method to mimic a 'survival of the fittest' environment whereby the transfer from one population of chromosomes (strings of information, in computer terms, discussed below) to another involves the aforementioned biological processes along with heuristic search strategies (Mitchell, 1996).

Each chromosome in the solution space is a candidate solution and consists of genes (bits of information). Each gene, in turn, is an instance of what is called an allele (either a 0 or a 1) (Mitchell, 1996).

As mentioned earlier, the algorithm can be said to apply the principle of survival of the fittest. This means that the algorithm's goal is to select the chromosome (candidate solution) or the combination of chromosomes which maximises some selected scoring metric along with a goodness-of-fit function.

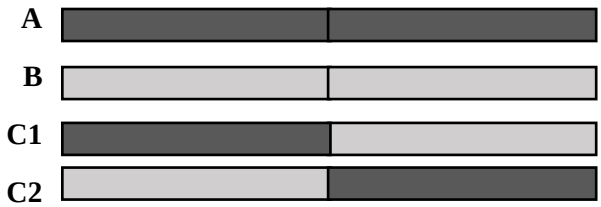
Standard genetic algorithm applications begin with a random generation of initial candidate solutions using a uniform distribution (Correa & Goodacre, 2011). The fitness function is then applied to this initial population to determine how well each candidate solution solves the problem. Each fitness score associated with each initial candidate solution is then used to stochastically select which chromosomes will be used as inputs to the crossover, mutation, or inversion processes.

The crossover process involves swapping subparts of two strong chromosomes (chromosomes with high fitness scores) to create a combined chromosome with an even higher fitness score, i.e., the process takes parts of well-suited candidate solutions to a problem and creates a third candidate solution that is superior to the two. The mutation process involves the random change of the allele values within each chromosome and tests if the product of such a change makes for a better candidate solution. Last, inversion is the process of reversing the order of the genes making up the chromosome, again with the intention of finding a better candidate solution (Mitchell, 1996).

The application of the crossover process can further be broken down into one-point crossovers and two-point crossovers. In the case of a one-point crossover, a crossover point is selected in each of the two candidate solutions chosen, and data beyond the crossover point are exchanged between the two original chromosomes, thus creating two new chromosomes. Figure 4, below, depicts this process graphically. In the case of a two-point crossover, two crossover points are selected in each of the two candidate solutions chosen, and data between the two crossover points are swapped, creating, again, two new chromosomes. Figure 5, below, depicts this process graphically.

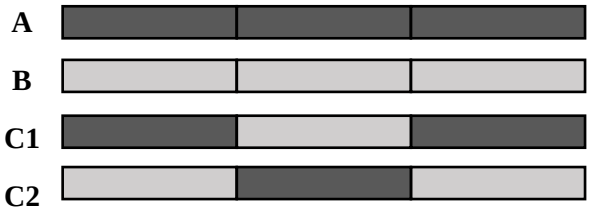
The use of both the crossover process and the mutation process ensures that the genetic algorithm employed would not converge to sub-optimal local minima (as a result of the crossover process) or end up in a random walk (as a result of the mutation process). Hence, once crossover is performed, mutation is employed and the worst-fitting chromosomes are removed from the solution space. This process is repeated until it converges to the optimal solution, or until a pre-specified number of iterations during which no improvement was observed is completed.

Figure 4: One-Point Crossover



Note: This figure depicts the process of crossover applying one-point crossover in a genetic algorithm. The process involves two chromosomes (A and B). A crossover point is selected in the two chromosomes (depicted by the straight line dissecting each chromosome). Data beyond the crossover point in each chromosome are then swapped with the data beyond the crossover point of the other chromosome to create two new chromosomes, C1 and C2.

Figure 5: Two-Point Crossover



Note: This figure depicts the process of crossover applying two-point crossover in a genetic algorithm. The process involves two chromosomes (A and B). Two crossover points are selected in the two chromosomes (depicted by the straight lines dissecting each chromosome). Data between the crossover points in each chromosome are then swapped with the data between the crossover points of the other chromosome to create two new chromosomes, C1 and C2.

4.1.4.2. The Peter and Clark (Stable) Algorithm

The Peter and Clark (PC) algorithm was introduced by Peter Spirtes and Clark Glymour in 1991 in an attempt to improve some of the existing algorithms at the time. Being a constraint-based algorithm, the PC algorithm uses conditional independence tests to determine the causal relationships existing in the network (Colombo & Maathuis, 2014). The PC algorithm employs a limited adjacency search to identify which edges can be removed. This limited adjacency search attempts to identify the edge to be removed between two nodes using conditional independence tests on any of the nodes in the network (Spirtes & Glymour, 1991).

The limited adjacency search employed by Spirtes and Glymour (1991) is, however, applied in some order (Colombo & Maathuis, 2014). This dependency can result in different structures being learned by the PC algorithm for sets of the same nodes ordered differently (Dash & Druzdzel, 1999; Cano, Gómez-Olmedo, & Moral, 2008; Spirtes, Glymour, & Scheines, 2000). The significance of the algorithm's order dependence increases in importance for larger datasets (Colombo & Maathuis, 2014), warranting an algorithmic adjustment to remove the dependence.

The resulting adjustment is the PC (Stable) algorithm presented by Colombo and Maathuis (2014). In this algorithm, the authors remove the order dependence in three different

stages of the implementation of the PC algorithm: The skeleton stage, the collider[25] stage, and the orientation stage.

The skeleton stage examines the adjacency step of the PC algorithm and results in the removal of edges based on conditional independence tests after all adjacencies are examined, as opposed to as soon as a conditional independence test suggests the removal of an edge (Colombo & Maathuis, 2014). This adjustment removes the possibility of edges being removed prematurely, affecting the network structure, and, potentially, resulting in false positive edges remaining part of the structure once other edges were removed prematurely.

In the collider stage, Colombo and Maathuis (2014) determine conditional dependence yielding a collider by determining a triple (X, Y, Z) to be unambiguous if there exists at least one separating set S that d-separates X and Z that contains Y and Y is in strictly less than half of such a set S. This is a majority rule adjustment to the conservative PC (CPC) algorithm developed by Ramsey, Zhang, and Spirtes (2006) to learn the network's colliders. As in the CPC algorithm, this treatment of colliders removes the order-dependent identification of the separating sets S.

Finally, the orientation stage examines the colliders identified in the collider stage and deals with the potential of conflicting directionalities within the colliders. The PC (Stable) algorithm combines the conflicting directionalities into bidirectional edges, while the PC algorithm chooses the true collider directionality arbitrarily (Colombo & Maathuis, 2014).

4.1.4.3. The Max-Min Hill-Climbing Algorithm

The max-min hill-climbing (MMHC) algorithm was introduced by Tsamardinos, Brown, and Aliferis in 2006 as a new structure-learning algorithm to be applied to BNs. The algorithm makes use of the max-min parents and children (MMPC) algorithm to first learn the skeleton of the network before calculating the conditional independencies of the belief network.

In principle, the MMHC algorithm is meant to allow for the learning of and inference from the belief network to be scaled to the point where these processes can be applied to thousands of data points and nodes (Tsamardinos, et al., 2006). It aims to reduce computational demands, something that earlier 'state-of-the-art' algorithms at the time struggled with before the introduction of the algorithm (Silverstein, Brin, & Ullman, 2000).

[25] This is also known as a v-structure in BNs, and is referred to as a v-structure by Colombo and Maathuis (2014).

As mentioned earlier, there are two main approaches to learning the structure of the BN. The MMHC is a combination of the two, i.e., a hybrid of both methods (Tsamardinos, et al., 2006). The algorithm first employs the MMPC algorithm to learn the skeleton of the network and then proceeds to apply a greedy scoring algorithm using a hill-climbing search (Tsamardinos, et al., 2006). Therefore, the algorithm begins with the second method, i.e., the conditional independence method, and then proceeds to employ the first method, i.e., a search-and-score method.

Tsamardinos, et al., (2006) point out in their study that the MMHC algorithm is an instantiation of the sparse candidate (SC) algorithm originally developed by Friedman, Nachman, and Pe'er (1999) as a learning algorithm designed to be employed with databases containing hundreds of data points. The SC algorithm is a search-and-score algorithm that is constrained by the total number of parents of a node being capped, being at most κ, where κ is chosen by the user of the algorithm. The SC algorithm then applies the hill-climbing algorithm to maximise the local score metric (Friedman, et al., 1999). The process is then reiterated while updating the parent sets of each node until the changes in the parent sets are eliminated or, alternatively, until a certain number of iterations lapses.

As mentioned, the MMHC algorithm starts by developing a skeleton of the network, thereby identifying all possible parents. This mitigates a drawback of the SC algorithm: The MMHC algorithm does not assume that there only κ parents, avoiding the situation where the actual number of parents may be $m > \kappa$. The application of the MMPC algorithm ensures that the number of parent nodes m is established for each node (Tsamardinos, et al., 2006). This makes the algorithm an exhaustive algorithm.

The MMPC algorithm is employed to build the skeleton of the network by establishing the relationships between children nodes and parent nodes. The notation used in Tsamardinos, et al., (2006) is adopted. $PC_T^{\mathcal{G}}$ is used to denote the set of both parents and children of a node T in a DAG $\mathcal{G}$. If $(\mathcal{G}, P)$ and $(\mathcal{G}', P)$ are two faithful BNs (see Section 4.1.3), then, for any node T, $PC_T^{\mathcal{G}} = PC_T^{\mathcal{G}'}$. Hence, the superscript can be dropped and the set of parents and children of a node T can be denoted by PC_T, due to the uniqueness of the set among all BNs that comply with the faithfulness condition as applied to the distribution P. The set PC_T is the output of the MMPC algorithm (Tsamardinos, et al., 2006). Note that the set PC_T identifies both parent nodes and children nodes, i.e., it does not identify the direction of the relationship (edge), but

simply identifies the existence of the relationship (edge). This is the skeleton identification process of the belief network.

The first step in the MMPC algorithm just described adds all possible edges to and from a node T. This leaves a skeleton that could be burdened by too many edges related to node T, some of which are likely to be redundant. The second step in the algorithm is then to remove some of the edges constructed in the first step, thereby ensuring that the resulting skeleton does not contain any false negative relationships. However, some false positive relationship might still be present. The final step removes any false positive relationships, leaving node T with a set of parents and children that are indeed related to it (Koski & Noble, 2009).

The MMPC algorithm attempts to find the minimum association between two variables, where association is a measure of dependence. Tsamardinos, et al., (2006) define a minimum association function between variables X and T relative to another set Z (or any subset S of Z) as follows.

$$MinAssoc(X,T|Z) = \min_{S \subseteq Z} Assoc(X,T|S) \tag{42}$$

This is then used to establish whether the variable X belongs to the parents-children set PC_T of T.

The processes of establishing the parents-children set PC_T of T is iterative. It begins by identifying a candidate node for either a parent or child role, denoted by CPC, which is to be included in the set PC_T. The CPC variable is established using the max-min heuristic that stipulates that the minimum association function, defined in Equation (42), should be maximised relative to the CPC. This initial phase of the process is terminated when all the variables that remain are independent of node T given some subset of the CPC.

The next phase of the algorithm targets false positive edges by determining whether a node X is independent of node T for some subset S of the CPC. Should this condition prevail, node X is then removed from the CPC.

The MMHC algorithm is employed to identify the nature of the relationships (i.e., the directions of the edges) between the parents and children of each node, as identified by the MMPC algorithm. This is performed using a greedy hill-climbing search (Tsamardinos, et al., 2006).

The hill-climbing is initialised with a graph (one that is either full, empty, or randomly generated), or, in the case of the MMHC algorithm, with the graph determined using the MMPC algorithm, and a scoring metric. The algorithm then amends the graph by either adding edges, deleting edges, or changing the direction of existing edges in order to maximise the selected scoring metric. As mentioned, edges may only be added if a relationship was established between the variables by the implementation of the MMPC algorithm. Else, the relationship is ignored.

The algorithm's time complexity is reduced when compared to the SC algorithm. This is due to the prior identification of possible relationships as performed by the MMPC algorithm.

4.1.4.4. The Semi-Interleaved HITON Parents and Children Algorithm

The HITON Parents and Children (HITON-PC) algorithm was introduced by Constantin F. Aliferis, Ioannis Tsamardinos, and Alexander Statnikov in 2003 as an algorithm that uses Markov blankets[26] to discover the structure of a network in a manner that is sound[27], efficient, and scalable. The authors identify that previous attempts to provide a Markov blanket approach did not comply with all three elements, yielding network learning algorithms with dependencies that are excessive in number due to inefficiencies resulting from the absence of the collective elements above.

The HITON-PC algorithm produces a set of nodes deemed to be related to the target node, specifically those deemed to be either parent nodes or children nodes (Aliferis, et al., 2003). This is done by identifying the target node's associated Markov blanket from a general set of nodes (Aliferis, et al., 2003), where the nodes' relationships are unknown. The sets of parents and children are identified as those identified for the MMHC algorithm in Section 4.1.4.3.

The interleaved HITON-PC algorithm, introduced by Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D. Koutsoukos in 2010, attempts to populate the set of parents and children of a target node T, PC_T. The interleaved

[26] A Markov blanket is the set S of a variable X, whereby all other variables conditioned on S are independent of X (Aliferis, et al., 2003).

[27] A network learning algorithm is said to be sound if: (i) The joint PDF is faithful to the BN considered; (ii) There are sufficient data points to yield statistically credible results; and (iii) The classifiers used are sufficiently powerful (Aliferis, et al., 2003). For details surrounding a BN's faithfulness and the Markov condition, see Section 4.1.3.

HITON-PC algorithm begins with an empty candidate set and adds variables based on their univariate associations with the target node T (Aliferis, et al., 2010). As a new node is added to the candidate PC_T set, the impacts of the added node on those already included in the candidate set are re-evaluated, with reference to either the mentioned univariate causal relationships or the score of the network. For the former, any node whose association with the target node has been eliminated due to the inclusion of the new node is eliminated from the candidate set (Aliferis, et al., 2010). The algorithm concludes that a candidate set is the true PC_T when there are no more variables to include and, by definition, the candidate set is that which exhibits either the highest degree of association or the highest network score.

Finally, Aliferis, et al., 2010, introduce the semi-interleaved HITON-PC (SI-HITON-PC) algorithm. Its only difference relative to the interleaved HITON-PC algorithm is that the elimination step is not performed to the full extent of the interleaved HITON-PC algorithm (Aliferis, et al., 2010). Instead of eliminating nodes already included in the candidate PC_T set, the SI-HITON-PC algorithm attempts to achieve the same task as the interleaved HITON-PC algorithm by eliminating the node recently added to the set. This aspect reduces the algorithm's complexity relative to its interleaved HITON-PC counterpart (Aliferis, et al., 2010).

4.1.5. Inference

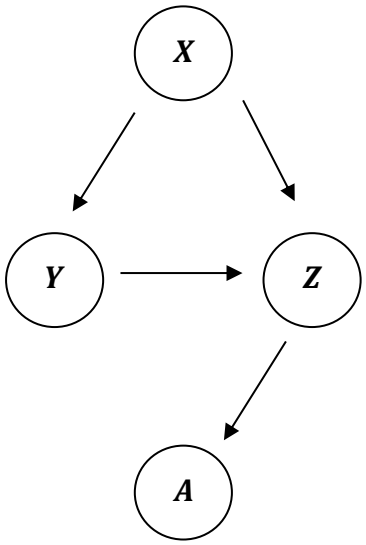
Inference, in the realm of BNs, is the concept of moving from the variables and the statistical data available to conclusions about the causal structure of the network. Inference in BNs has a flexible application due to the ease of updating information in the BN. The inference undertaken in this study was performed via the learning of the causal relationships using the algorithms described above.

Nilsson (1998) defines three main methods of inference: Causal inference, diagnostic inference, and explaining away inference. Causal inference, also known as top-down inference, infers the state of a child node (also known as the query node) from the state of a parent node (also known as the evidence node). Hence, the query node's state is said to be caused by the evidence node's state. Diagnostic inference, also known as bottom-up inference, infers the state of a parent node from that of its child. Hence, the effect of the child node on the parent node is first examined to infer the cause. Last, explaining away inference is the combined implementation of both top-down inference and bottom-up inference simultaneously.

4.1.6. Dynamic Bayesian Networks

Up to this point, references made to BNs implicitly referred to what are known as static BNs (SBNs), i.e., BNs that model causal relationships that exist at a specific point in time (Friedman, Murphy, & Russell, 2013). A generic example of a SBN is depicted in Figure 6. However, SBNs can be expanded to include a temporal element in them, thereby extending the capacity of the network to learn not only causal relationships between the networks' nodes at a specific point in time, but also the causal relationships within and between the nodes over time (Dagum, Galper, & Horvitz, 1992). These are called dynamic BNs (DBNs), a generic example of which is depicted in Figure 7.

Figure 6: A Static Bayesian Network



A Static Bayesian Network

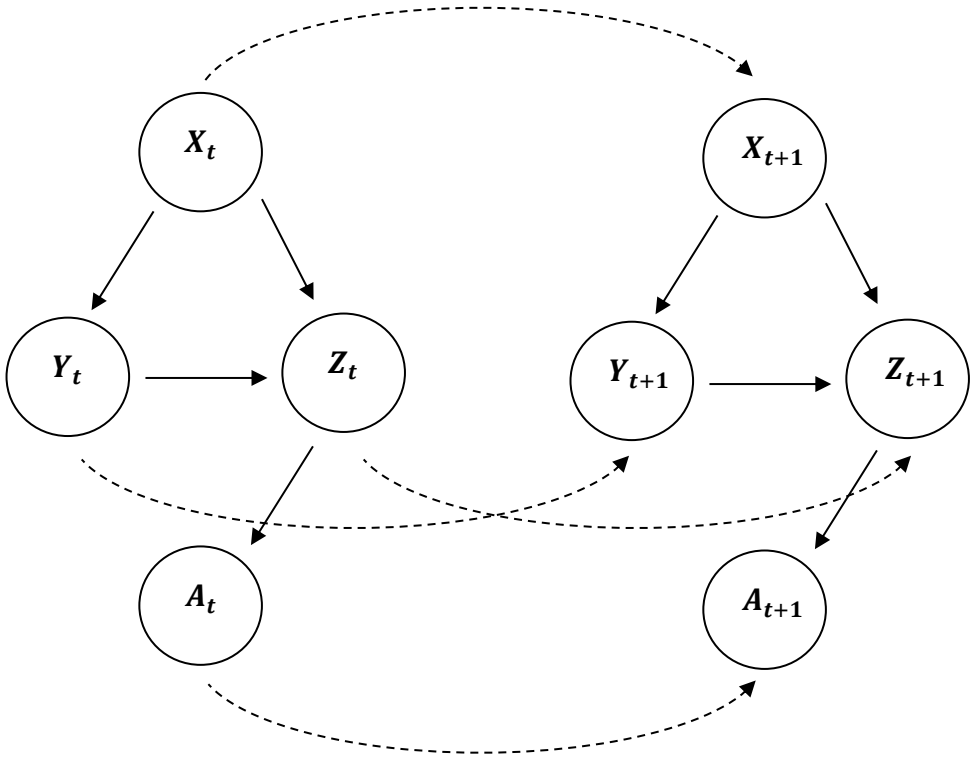
Note: This figure depicts a static Bayesian network (SBN). The SBN represents a directed acyclic graph (DAG) and causal relationships between the variables X, Y, Z, and A. The solid arrows represent the SBN's causal relationships within a period.

The temporal elements of a DBN are much better suited for the purposes of this study than the static elements of the SBN. Financial data are often time series data, and the time series data relating to a single variable are affected by that variable's filtration system up to that period, $\mathcal{F}_t$, as discussed in Section 3.1.3. Since this study requires the networks employed to forecast closing values of the S&P 500 index (discussed in more detail in Section 4.3.2), the modelling of the causal relationships between the networks' nodes, both over time and at each time step, is far more useful than any modelling that does not include a temporal component. This is highlighted in Figure 7.

The dashed arrows in Figure 7, representing the causal relationships of the DBN's variables between variables, represent the time series data of each node in the network. The added benefit

of the use of BNs is the discovery of causal relationships between variables within each period, as shown by the solid arrows in Figure 7. It is, therefore, evident that the use of both inter-period and intra-period causal relationships is better suited for the forecasting of financial data in this study, as illustrated by the importance of autoregressive models in Chapter 3.

Figure 7: A Dynamic Bayesian Network



A Dynamic Bayesian Network

Note: This figure depicts a dynamic Bayesian network (DBN). The DBN represents a directed acyclic graph (DAG) and causal relationships between the variables X, Y, Z, and A at each period t. The solid arrows represent the DBN's causal relationships within a period, while the dashed arrows represent the DBN's causal relationships between periods.

Moreover, the SBN learning algorithms discussed in Section 4.1.4 have been extended to be used in the applications of DBNs. For example, genetic algorithms (see Section 4.1.4.1) have been extended to a dynamic setting by Tucker and Liu (1999), Tucker, Liu, and Ogden-Swift (2001), and Wang, Yu, and Yao (2006), while the MMHC algorithm (see Section 4.1.4.3) was extended to a dynamic setting by Trabelsi, Leray, Ben Ayed, and Alimi (2013). Hence, their uses can be extended naturally to a temporal setting without amending the discussions provided.

Therefore, this study strictly uses DBNs in its methodology. The discussion of SBNs and DBNs is confined to this section of the study alone, and further references to BNs will mean DBNs, and references to any of the BNs learning algorithms discussed in Section 4.1.4 will mean their respective dynamic extensions.

4.2. Literature Review

This section discusses the relevant academic literature relating to the application of BNs in the process of asset pricing. A review of the literature indicates that no directly comparable studies have been undertaken to apply BNs to the process of calculating a bank's profit and loss account and measuring the bank's market risk using such a process. Hence, this literature review focuses on the application of BNs to asset pricing in general, as some commonality is found within this broader subject area. It then focuses on the few studies that do apply BNs to the calculation of financial risk, in general, and market risk, in particular, later in this section.

BNs have been regarded as a field of study since as early as 1985 and have been formally characterised in Judea Pearl's 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. As mentioned earlier, there is extensive literature available on the uses and application of BNs in fields such as medicine and weather prediction. When it comes to literature relating to BNs and banking and financial markets, BNs are often applied to fields such as equity investing (discussed below) and operational risk[28]. The literature concerning market risk and VaR and ES, on the other hand, is almost non-existent. Nonetheless, the rationale for applying BNs, even to the specific field of market risk management within banking, can be extrapolated from the application of BNs to other aspects of the financial field, as revealed in the literature below.

A crucial aspect that must first be established is the usefulness of the structure and dynamics of a BN within the field of finance. Due to the causal relationships depicted by BNs, they are well-suited as an aid for analysts examining both individual stocks as well as portfolios in the presence of uncertainty (Demirer, Mau, & Shenoy, 2006). This makes BNs an appropriate, useful, and relevant tool in the pricing of assets and portfolios and, by extension, banks' trading desk portfolios subject to regulation. Both Demirer, et al., (2006) and Shenoy and Shenoy (2000) provide basic outlines for constructing BNs for portfolio management.

An intuitive approach to structuring a BN for portfolio management application is the top-down approach outlined by Demirer, et al., (2006), which begins by examining macroeconomic variables that are thought to affect a pharmaceutical industry under examination. The authors then proceed to examine firm-specific attributes such as revenue and

[28] For some examples, see Neil, Fenton, and Tailor (2005); Aquaro, Bardoscia, Bellotti, Consiglio, De Carlo, and Ferri (2010); and Lockamy and McCormack (2012).

the cost of goods sold. Finally, Demirer, et al., model the causal relationships between the four firms under examination and the variables identified.

This top-down approach is not only intuitive, but it also accommodates an efficient investigation of factors. Arbitrage pricing theory (APT) variables identified in the literature can often be separated into macroeconomic and firm-specific attributes. This makes the investigation into the different factors feasible. These are discussed further in Section 4.2.

It is important to note at this stage that the structure of the network (i.e., the causal relationships modelled) is far more important than the variables used (Demirer, et al., 2006). Hence, when structuring the network, the relationships modelling employed should be considered carefully when selecting the factors to incorporate.

A significant benefit of BNs is their ability to combine both qualitative and quantitative aspects of the subject they are employed to model (Olbryś, 2009). In the case of asset pricing, the former includes aspects such as expert opinion and judgement[29], while the latter includes historical data.

Olbryś (2009) uses changes in inflation, the unemployment rate, industrial production, and the yield on one-year Treasury bills to calibrate a BN to forecast the returns of the Warsaw Stock Exchange's main index, as well as its sub-indices. Following the moments method adopted by Demirer, et al., (2006), the historical values of each node are then bucketed into a 'low' bucket, a 'medium' bucket, and a 'high' bucket, representing 25%, 50%, and 25% of the data points, respectively (Olbryś, 2009). The study employs a BN to discover the causal relationships between the nodes identified and the returns generated from the index used, showing the usefulness of BNs in determining variance-covariance matrices when calculating market performance.

Earlier studies incorporating BNs into asset pricing often focus on trading equity instruments. Jangmin, Lee, Park, and Zhang (2004) demonstrate the usefulness of a dynamic BN in technical analysis in the South Korean market. Zuo and Kita (2012a) use BNs to predict the trajectory of prices (either an upward or a downward trajectory) of stock prices and show BNs to be similarly useful to other existing trajectory prediction methods. Separately, Zuo and

[29] This aspect includes the causal relationships that may exist between the network's nodes.

Kita (2012b) also use BNs and price-to-equity ratios to find trends in the Nikkei 225 index in Japan.

Expanding on the previous work done on trading, several authors employ BNs to aid in longer-term investing. For example, in a study focusing on both trading (via futures and options trading) and longer-term investing (via buy/sell decisions), Chang and Tian (2015) conclude that the BN used allows for the incorporation of technical analysis (using the historical distribution of the underlying index, the S&P 500 index) as well as fundamental analysis (by incorporating the candidate causal relationships inherent in the network). Chang and Tian use the London Interbank Offer Rate (LIBOR), the US consumer price index (CPI), the unemployment rate, the money supply, and the volatility index (VIX), based on the S&P 500 index, among others, as the macroeconomic variables affecting the returns earned on the S&P 500 index. The authors find that the use of a BN to both trade short-term market fluctuations and invest with a longer-term perspective yields superior results over a ten-year period relative to a conventional buy-and-hold investment strategy (Chang & Tian, 2015).

A more recent study examining systemic risk in financial markets using BNs is that of Chan, Chu, and So (2023). In their study, the authors use a BN to model financial crises, using volatility as a proxy, by examining the interconnectedness of variables in financial markets. The study's primary relevance is the use of a rolling period to model the relationships believed to exist between the variables, using a BN. However, the authors' study does not make use of DBNs in the common sense of modelling the temporal relationships between the networks' nodes, but, rather, the authors create a time series of BNs and rank the various nodes based on their relative importance, based on a novel metric (Chan, et al., 2023).

Finally, and most relevant to this study, Apps (2020) applies a BN to calculate daily VaR forecasts relating to three United Kingdom (UK) banks' shares. Apps uses a simplified methodology to model stock returns by modelling whether the three-share portfolio returns are positive or negative using a Gaussian BN[30]. The network constructed includes three variables – a liquidity variable, a market variable, and the target variable, being whether the return achieved on the three-share portfolio is positive or negative (Apps, 2020).

The review above highlights the usefulness of BNs in incorporating both beliefs and past data when forming views of the future. BNs' established uses in the valuation of equities lend

[30] A Gaussian BN assumes multivariate normality (Grzegorczyk, 2010).

themselves to the application of determining the profit and loss account of a bank's equities trading desk. BNs further lend themselves to aid regulators in establishing causal relationships both within the economy, as shown by Chan, et al., (2023), and when a BN is applied to a specific portfolio's market risk, as shown by Apps (2020). Using the rolling period methodology presented by Chan, et al., although in its proper temporal form, and applying it to calculate 10-day 99% VaR forecasts, 10-day 97.5% ES forecasts, and their stressed versions, this study aims to expand on the simplified methodology applied by Apps to calculate market risk metrics.

This study aims to fill this gap by forecasting 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts, and their stressed counterparts, for the equities trading desk of a US bank using a BN-determined profit and loss account. The methodology undertaken in this study, which is discussed in the next section, uses several network learning algorithms to produce forecasts of several market risk metrics, as would be required by practitioners in charge of the market risk management of the equities desk of a US bank.

4.3. Data and Methodology

This section opens with a discussion of macroeconomic and financial variables which may serve as nodes in the BNs. The variables identified are those indicated by the literature as those having a potential causal relationship with the market proxy identified in Chapter 3, i.e., the S&P 500 index. Next, the data and rolling period methodology employed in constructing the BNs employed in this study, using the BN learning algorithms discussed in Section 4.1.4, are discussed. The backtests and forecasting error measures employed in this chapter follow, as introduced in Chapter 3. This chapter's results are then analysed with reference to the performances of the various BNs employed in this study to produce 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts, as well as their stressed counterparts. The results include the performances of the various BNs in producing the market risk metric forecasts with respect to the backtests and forecasting error measures employed in this study. Finally, a conclusion is provided to conclude this chapter.

4.3.1. Arbitrage Pricing Theory

The APT was developed by the economist Stephen Ross and published in 1976 as an alternative to the traditional capital asset pricing model (CAPM). The APT may be regarded as a more abstract model relative to the CAPM. While the CAPM stipulates that the only variable to be considered is the market portfolio, the APT does not provide any explanation as to what

the variables relevant to returns are, but, rather, it simply provides a general, linear, relationship between the variables and the asset's return. This allows for the investigation of relationships between various financial, economic, company-specific, and market variables which are believed to affect returns. This investigation is useful in the identification of nodes in the belief network.

One of the earlier studies into the APT and its relevant equity variables to the US stock market was that of Chen, Roll, and Ross (1976). In their study, the authors suggest the presence of an effect of macroeconomic and financial variables on asset returns. They continue to identify these variables and their relations to expected returns. Their study concludes that, of the a priori variables tested, the industrial production, the premium attached to the risk of bonds, expected inflation and unexpected inflation, and the changes to the yield curve (particularly, twists) are all significant variables contributing to the returns of the assets considered in their sample over a 20-year period from 1953 to 1973.

Building on the work presented by Chen, et al., (1976), Fama (1990) examines the time variation of value-weighted returns earned on the New York Stock Exchange (NYSE) over the period 1953 to 1987, i.e., a period which encompasses that used by Chen, et al. Fama finds that around 33% of the variance of the annual returns earned on the NYSE is explained by the dividend yield, the term spread, and the default spread. This percentage is increased to 59% when including the growth rate of industrial production (used as a proxy predictor for corporate cash flows). These findings support those of Chen, et al., and highlight that industrial production (and its growth rate) is an important variable when analysing US equities.

Shanken and Weinstein (2006), on the other hand, contradict the findings made by Chen, et al., (1976) when testing for the significance of the mentioned variables over the period from 1968 to 1977. The authors point out a lack of robustness in the tests for the variables and conclude that studies performed over different periods may reach different conclusions. This conclusion further highlights the crucial need for expertise when applying abstract models such as the APT, as well as the time-varying dependencies that may exist between variables in financial markets.

In an early investigation into the determinants of the volatility experienced in US equity returns, specifically those included in the S&P 500 index, Lawrence (1989) suggests that differences between those companies included in the index and those excluded from the index include differences in the frequencies of trading of the shares, the respective price levels, the

sizes of the companies[31] (captured in terms of market capitalisation, as the S&P 500 index includes the largest 500 US companies by market capitalisation), and changes in beta values.

Further highlighting the abstract nature of the APT and its complicated application, Riesman (1992) states that expected returns could be approximated using the APT model as long as some proxy variables are used that are correlated to each other. Riesman shows that, practically, any set of correlated variables can be used, as long as their slope coefficients matrix (as used in the multivariate regression required on the variables) is invertible (Shanken, 1992).

Chen and Jordan (1993) identify inflation, the rate earned on US Treasury bills, the total return earned on long-term US government bonds, industrial production, the oil producer price index, and the return on industrial-related bond issues as significant variables. Moreover, the changes in these variables, both expected and unexpected, were identified as significant. These included expected and unexpected inflation, changes in expected inflation, changes in industrial production, changes in the price of oil, changes in the US term structure of interest rates, and changes in the risk premium[32] (Chen & Jordan, 1993).

In a recent application of machine learning to analyse the returns of US equities, Arrieta-Ibarra and Lobato (2015) compare the performances of random forests, support vector machines (SVMs), and ANNs against that of a generalised autoregressive conditional heteroscedasticity (1,1) (GARCH(1,1)) model using the S&P 500 index. The authors use inputs such as the three-month Treasury bill returns, commodity prices (specifically gold bullions and West Texas Intermediate, or WTI, oil), and the exchange rates of the Japanese Yen, the British Pound, and the Swiss Franc against the US Dollar to model the returns on US equities.

As discussed in Section 4.2, several authors who have applied BNs to asset pricing have used the S&P 500 index as their benchmark. Macroeconomic and financial variables such as the LIBOR, CPI, VIX, and the US unemployment rate have been found to exhibit causal relationships with the level of the S&P 500 index, while variables such as the level of money supply and housing levels have been found to not be correlated closely enough to the level of the index (Chang & Tian, 2015).

[31] As identified by Banz (1981).

[32] As determined by the difference between Aaa-rated industrial bonds and Baa-rated ones.

The discussion above highlights the differing opinions are expressed in the literature; some accredit return-related effects to some variables, while others criticise the existence of those same relationships. Nonetheless, the literature identifies industrial production, the premium attached to the risk of bonds, unexpected and expected inflation, and changes in the latter, and the changes to the yield curve (particularly, twists) as macroeconomic and financial variables that are worthy of attention. The prices (and changes thereof) of oil and gold, the unemployment rate, the level of the VIX (and changes thereof), and various exchange rates relative to the US Dollar are also worthy of examination. Some company-specific variables include the size of the company, as measured by market capitalisation, and beta values.

A further complication posed by the APT model is the fact that the identification of variables, macroeconomic, financial, or otherwise, may be made even more difficult due to the time dependencies of the variables on current market conditions. As the market conditions change, so will the influences of the variables on the returns. This means that, even if the required expertise to apply the model exists, findings may be contradicted in future research applied to a different period due to changes in market conditions and, therefore, sensitivities.

4.3.2. Data and Bayesian Network Rolling Period Methodology

The relevant macroeconomic and financial variable data, as well as the appropriate proxies, were gathered from the Bloomberg database for the period matching that used in Chapter 3, i.e., 15 March 1991 to 14 February 2020. Again, this period was chosen as it covers three business cycles (National Bureau of Economic Research, n.d.).

The closing values of the S&P 500 index, as the target output variable, were collected for the period. Moreover, the economic and financial variable data collected were for the variables identified in Section 4.3.1, as well as data related to other variables considered to be relevant and to exhibit causal relationships with the closing values of the S&P 500 index. Since the BNs use various network learning algorithms to learn the network structure using conditional probabilities and causal relationships (see Section 4.1.4), using expert judgement to add variables does not negatively impact the result, as these learning algorithms will determine no causal relationships between the variables included in the data and the S&P 500 index if a causal relationship is not identified.

The BN network learning algorithms discussed in Section 4.1.4 were applied using a rolling period methodology, as per the methodology employed in Chapter 3. This rolling period methodology produced a total of 7,286 one-day-ahead forecasts of the closing value of the S&P

500 index using the causal relationships learned between the economic and financial variables, i.e., the network's nodes, and the S&P 500 index during the out-of-sample period. At every iteration of the rolling period, the causal relationships were re-calibrated, which means that both the strengths and the existence of relationships may have changed with every iteration. These re-calibrations took place both at a point in time and between periods, utilising the temporal element of the learning algorithms used. This approach was deemed to be the most robust approach, as it allowed for new market, economic, and financial information to be incorporated into the prediction of the following trading day's closing value of the S&P 500 index, as would be the case in practice. Each of the forecasted closing values of the S&P 500 index was then incorporated into the return PDF to facilitate the calculation of 7,286 10-day 99% VaR forecasts and 7,286 10-day 97.5% ES forecasts, and their stressed counterparts.

Two rolling periods were used in this chapter. First, a rolling period was used to calibrate the BNs using the various network learning algorithms to produce one-day-ahead forecasts of the closing value of the S&P 500 index. Second, a rolling period was used to produce the relevant market risk metrics. Each of these rolling periods is 1,264 days in length, as used in Chapter 3, which equates to approximately five trading years. Hence, data preceding this study's start date are also collected, up to 10 trading years prior to 15 March 1991.

The rolling period methodology described above influenced the availability of economic and financial variables and data, since the data must exist from 18 March 1981 from the Bloomberg database to facilitate the various calibration and rolling periods. In total, data covering 41 economic and financial variables (see Appendix A) were collected for the study period and the relevant calibration and rolling periods. These variables included financial indices, economic indicators (e.g., US non-farm payroll data and US CPI), currency exchange rates, and commodity prices. As discussed previously, the network learning algorithms determine the existence of causal relationships between the variables and the S&P 500 index, if such exist, both at a point in time and between periods. Hence, while this study used 41 economic and financial variables, not all 41 variables necessarily exhibited directed causal relationships with the S&P 500 index, let alone exhibited these directed causal relationships at every single iteration of the rolling calibration period (since the networks were re-calibrated with every iteration).

Of the 41 variables used as inputs, it is worth noting that not all variables were updated daily. The updating frequencies of the variables used ranged from daily to quarterly. Since

daily closing values of the S&P 500 index were forecasted by the BNs employed in this study, and due to the inability of the network learning algorithms employed to deal with missing values, any non-daily variable data were transformed to daily data as follows. On any given trading day on which a new observation for the variable was unavailable, the variable's value was set to equal the last known value of the variable. For example, suppose that US CPI data are released on the 10th of every month. On the 10th, that new observation was recorded in the data, until the next month's observation became available, also on the 10th. Between these two dates, the first observation was recorded as the value for every day in the period. This method of dealing with the different frequencies of the data was adopted as it mimics the use of available data in practice (i.e., only the last known US CPI value is available to the risk manager at a US bank. Hence, the risk manager can only use this last known value in making any risk management decisions, until such a time that a new US CPI value is released to the market).

It is also worth noting that five of the 41 variables used as inputs did not have data for the entirety of the first of the two calibration periods, i.e., data for these variables were available starting between the beginning of the first calibration period and the end of this period. These variables were the daily prices of the WTI, the closing values of the UK's Financial Times Stock Exchange (FTSE) 100 index, the daily three-month US LIBOR rates, the daily spread on ten-year Aaa-rated US corporate bonds relative to Treasury bills, and the US policy uncertainty index. These variables were believed to potentially exhibit some causal relationships with the closing values of the S&P 500 index, if only at later stages of the study period. It was decided not to exclude these variables from the study due to their lack of data in the first calibration period.

Due to the inability of the network learning algorithms to deal with missing data, these variables' earliest known values were filled in from their earliest occurrence back to the beginning of the first calibration period. At most, this amounted to 981 trading days, or 10% of the total period. However, it is worth stressing that the methodology implemented in this study employed the re-calibration of the BN learning algorithms at every iteration of the rolling period throughout the study period. Therefore, while these filled-in values exist strictly in a portion of the first rolling period employed, i.e., at most from day 0 to day 981, these values were filtered out by the time the rolling period methodology began producing forecasts beyond the second rolling period, thereby learning the true causal relationships that existed between these variables, other variables, and the closing values of the S&P 500 index to produce one-day-ahead forecasts of the latter.

Finally, due to the re-calibration of the BN learning algorithms with every iteration of the rolling period, it is possible for these networks to not find any causal relationships on a particular trading day. Due to this, and potentially due to other reasons, the BN may have been unable to produce a one-day-ahead forecast for the closing value of the S&P 500 index on any given trading day. In this case, the last available forecast was used as the daily one-day-ahead forecast of the closing value of the S&P 500 index to facilitate the calculation of the relevant market risk metrics calculated in this study. This adjustment was only necessary for the genetic algorithm, as it failed to produce 664 one-day-ahead forecasts, or 9.11% of the total forecasts for this learning algorithm.

As per Chapter 3, 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts, and their stressed counterparts, were calculated at the equities desk level and then backtested, as banks are required to backtest their market risk metrics at a trading desk level (Basel Committee on Banking Supervision, 2019b). This offers three main advantages. First, the implementation of the BN models to calculate the market risk metrics at the trading desk level allows for a like-for-like comparison with the results of Chapter 3. Second, restricting the scope to the equities trading desk allows for the simplification of the BN models employed. This, in turn, makes the BN models outlined in this study significantly more viable for implementation by banks. Third, the separation allows for the measurement of the suitability and robustness of BN models to different trading desks in future research. This allows for conclusions to be made on the appropriateness of BN models for the equities trading desk, as well as others.

The approach outlined above is essentially the top-down approach outlined by Demirer, et al., (2006). The BN construction and the network learning were performed in R, using, amongst other resources, the dbnR package for the training of the DBNs employed in this study and forecasting of the one-day-ahead forecasts of the closing values of the S&P 500 index.

4.3.3. Backtesting

The backtesting techniques employed in this chapter are the same as those employed in Chapter 3. Specifically, the 10-day 99% VaR forecasts and the 10-day 99% SVaR forecasts were backtested using the BCBS's traffic light test, Kupiec's proportions of failure test, and Christoffersen's test for independence. Similarly, the 10-day 97.5% ES forecasts and the 10-day 97.5% SES forecasts were backtested using the BCBS's traffic light test, the conditional backtest, the unconditional backtest, and the minimally biased backtest.

Since the use of BNs in this study produced rolling forecasts that are incorporated into the PDF of the returns, the distribution of returns changed with every forecast. Hence, unlike in Chapter 3, the Du-Escanciano backtest (see Section 3.3.3.7) of ES and SES cannot be implemented in this chapter.

4.3.4. Forecasting Error Measures

The forecasting error measures techniques employed in this chapter are the same as those employed in Chapter 3. Specifically, the mean absolute error (MAE, see Equation (31)), the root mean square error (RMSE, see Equation (32)), the mean absolute percentage error (MAPE, see Equation (33)), and the median absolute percentage error (MdAPE, see Equation (34)). As per the methodology in Chapter 3, the symmetric MAPE (SMAPE, see Equation (35)) was used instead of the MAPE where the daily return is zero.

Moreover, the Diebold-Mariano test was used to statistically compare the relative forecasting abilities of the various BN models, as per Chapter 3. While all forecasts in this chapter were produced using BNs, different algorithms were used and, hence, these algorithms' forecasting abilities were statistically tested against each other.

4.4. Results

This section discusses the results using the data, methodology, and algorithms discussed in the preceding sections, and analyses of the results are provided. Specifically, this section examines the performance of each of the algorithms detailed in Section 4.1.4 to forecast 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts, and their stressed versions. This chapter is this study's primary contribution to the literature, and aims to determine which of the BN algorithms examined in this study produces a US bank's market risk metric forecasts that are most efficient, perhaps even more efficient than those of the traditional models used in Chapter 3.

Using each of the network learning algorithms considered in this study, each algorithm's forecasts were backtested using the backtests detailed in Section 4.3.3. Then, as carried out in Chapter 3, the forecasts were further evaluated on their efficiency in forecasting by assessing their out-of-sample performances relative to actual profit and loss amounts experienced by the equities trading desk of the bank. The performances of these forecasting error measures were further statistically tested using the Diebold-Mariano test, as detailed in Section 4.3.4.

In total, this chapter evaluates the performance of 10-day 99% VaR forecasts, 10-day 97.5% ES forecasts, 10-day 99% SVaR forecasts, and 10-day 97.5% SES forecasts using each of the four network learning algorithms, yielding a total of 16 BNs. Moreover, the performances of the 16 BNs were evaluated using the MAE, the RMSE, the MAPE, and the MdAPE, yielding values for 64 different forecasting error measures.

The rest of this section discusses each risk metric individually, exploring and analysing the results of the different network learning algorithms employed to calculate said metric in its context. The number of breaches experienced was calculated, and said breaches were backtested using the relevant backtesting techniques. Forecasting error measures were then employed to calculate the efficiency of the models used to forecast the risk metric, before using the Diebold-Mariano test to statistically evaluate the different models' forecasting ability and relative superiority.

4.4.1. Value at Risk

This section discusses and analyses the results of the network learning algorithms employed to learn the structure of the BNs in this study to forecast 10-day 99% VaR forecasts for a US bank using the S&P 500 index as the underlying, over the period 15 March 1991 to 14 February 2020. First, the numbers of breaches experienced for each of the various models are presented and discussed, followed by analyses of the backtesting results achieved using the BCBS's traffic light test, Kupiec's proportion of failure test, and Christoffersen's test for independence. Finally, a discussion of the forecasting errors using a variety of error measures, as detailed in Section 4.3.4, is presented, and these results are statistically tested for their relative predictive ability using the Diebold-Mariano test.

Table 39 contains the numbers of breaches observed for the various BNs employed in this study. Each of the four BNs experienced only three breaches during the 7,286-day out-of-sample period. All four BNs encountered breaches on the same days[33]: 27 October 1997, 31 August 1998, and 29 September 2008.

[33] These are the same dates on which the historical simulation model experienced breaches when producing 10-day 99% VaR forecasts, as detailed in Section 3.4.1.

Table 39: Breaches Observed for the 10-day 99% Value at Risk Metric using the Bayesian Network Learning Algorithms

	VaR Breaches
MMHC	3
Genetic	3
PC (Stable)	3
SI-HITON-PC	3

Note: This table reports the number of breaches experienced for applications of various Bayesian network (BN) learning algorithms to produce 10-day value at risk (VaR) forecasts at the 99% confidence level, using the daily logged returns of the Standard & Poor's (S&P) 500 index from 15 March 1991 to 14 February 2020. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the BN algorithms detailed in this table. The total number of VaR forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

The low numbers of breaches experienced by the BNs may suggest that the results of the BNs are in line with those experienced by banks using traditional models to produce 10-day 99% VaR forecasts, as discussed in Chapter 3. A comparison of the performances of the BNs' 10-day 99% VaR forecasts relative to those produced by the traditional models is reserved for Chapter 5. Nonetheless, the number of breaches observed for each of the BNs is in line with the literature surrounding banks' preferences for models that produce a low non-zero number of breaches (McAleer & da Veiga, 2008).

Table 40: Results of the Basel Committee on Banking Supervision's Traffic Light Test for the 10-day 99% Value at Risk Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Green zone
Genetic	Green zone
PC (Stable)	Green zone
SI-HITON-PC	Green zone

Note: This table reports the results of the Basel Committee on Banking Supervision (BCBS) Traffic Light test for banks' internal models based on the number of breaches of various 10-day value at risk (VaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the various Bayesian network (BN) learning algorithms. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the BN algorithms detailed in this table. The total number of VaR forecasts for the study period was 7,286 per algorithm. The number of breaches relative to the period length was then analysed using the BCBS's Traffic Light test to achieve one of three classifications. The classifications are 'Green zone' (if the binomial probability of the number of breaches relative to the number of trading days is greater than 99.99%), 'Yellow zone' (if the corresponding probability is lesser than 99.99% but greater than 99.99%), or 'Red zone' (if the corresponding probability is lesser than 95%). Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Due to the few breaches observed for each of the BNs, it is unsurprising that the BCBS's traffic light test results are in the 'Green zone' for all of the BNs employed, as shown in Table 40. The low numbers of breaches experienced by the BNs and the 'Green zone'

outcomes highlight that the use of BNs to produce 10-day 99% VaR forecasts, as undertaken in this study, would pass the BCBS's regulatory requirements when backtesting the forecasts produced.

The results of Kupiec's proportion of failure test at the 1% significance level, applied to the various BNs, are displayed in Table 41. As for the traditional models discussed in Chapter 3, the number of breaches expected for a 7,286-trading-day out-of-sample period at the 99% confidence level is between 50 and 94, inclusive. Since all BNs achieved three breaches in the out-of-sample period when producing 10-day 99% VaR forecasts, it is unsurprising that the backtest's null hypothesis is rejected for each of the four BNs. Hence, the BNs used to forecast 10-day 99% VaR forecasts are inaccurate at the 1% significance level.

Table 41: Results of the Kupiec Proportion of Failure Test for the 10-day 99% Value at Risk Metric using the Bayesian Network Learning Algorithms

	Breaches and Range	**Conclusion**
MMHC	$3 \notin [50,94]$	Reject
Genetic	$3 \notin [50,94]$	Reject
PC (Stable)	$3 \notin [50,94]$	Reject
SI-HITON-PC	$3 \notin [50,94]$	Reject

Note: This table reports the results of the Kupiec Proportion of Failure (PoF) backtest at the 1% significance level based on the number of breaches of various 10-day value at risk (VaR) Bayesian network (BN) learning algorithms at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the number of true breaches of the model is equal to the observed number of breaches. The results are based on the breaches of the logged returns earned on the Standard & Poor's (S&P) 500 index. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the BN algorithms detailed in this table. The total number of VaR forecasts for the study period was 7,286 per algorithm. The left column shows the number of breaches experienced under the BN learning algorithm together with the range predicted using the PoF test. Since the period has 7,286 trading days, the range of expected breaches is $[50,94]$ at the 99% significance level. The critical value of the χ_1^2 at the 99% confidence level is 6.64897. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

The final backtest employed in this section is Christoffersen's test for independence. Once again, the results contained in Table 42 are universal. The very few breaches experienced by each of the BNs used to produce 10-day 99% VaR forecasts are too few and sparse, leading to the failure to reject the null hypothesis. Hence, it is concluded that the breaches experienced under each of the BNs are independent at the 1% significance level.

Table 42: Results of the Christoffersen Test for Independence for the 10-day 99% Value at Risk Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Fail to reject
Genetic	Fail to reject
PC (Stable)	Fail to reject
SI-HITON-PC	Fail to reject

Note: This table reports the results of the Christoffersen test of independence backtest for banks' internal models based on the number of breaches of various 10-day value at risk (VaR) Bayesian network (BN) learning algorithms at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the probability of a breach following a non-breach is equal to the probability of a breach following a breach, i.e., breaches are independent. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the BN algorithms detailed in this table. The total number of VaR forecasts for the study period was 7,286 per algorithm. The critical value of the χ_1^2 at the 99% confidence level is 6.63490. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Table 43, below, depicts the MAE, the RMSE, the MAPE, and the MdAPE values for the various 10-day 99% VaR forecasts produced by the BNs employed in this study. As can be seen in Table 43, the various BNs used to produce 10-day 99% VaR forecasts achieved very similar results. This is expected given that the 10-day 99% VaR forecasts were calibrated using the return PDF, where only one of these returns is a forecast produced by the BN, while the (much larger) remainder is made up of the actual returns observed. Hence, on each day, the differences between the 10-day 99% VaR forecasts would be minor, and the results contained in Table 43 confirm this.

Since the differences in the forecasting error measures' values come down to the quality of the forecasts made by the BNs, it is concluded that the BN achieving the lowest of the forecasting error measures necessarily produced more accurate 10-day 99% VaR forecasts. Table 43 highlights that the BN achieving the highest value of each of the four forecasting error measures and, therefore, ranks as the least accurate BN when producing 10-day 99% VaR forecasts, is the BN using the MMHC algorithm. On the other hand, the BNs using the genetic algorithm, the PC (Stable) algorithm, and the SI-HITON-PC algorithm all score equally when using the MAE, the RMSE, and the MdAPE (as does the BN using the MMHC algorithm for the MdAPE). However, the MAPE shows that the BN scoring the lowest MAPE value (and, therefore, ranks as the most accurate) is the BN using the SI-HITON-PC algorithm.

Table 43: Forecasting Error Measures for the 10-day 99% Value at Risk Metric using the Bayesian Network Learning Algorithms

	MAE	**RMSE**	**MAPE**	**MdAPE**
MMHC	0.0945	0.1004	98.1677%	5.7790%
Genetic	0.0944	0.1003	98.1373%	5.7790%
PC (Stable)	0.0944	0.1003	98.1361%	5.7790%
SI-HITON-PC	0.0944	0.1003	98.1347%	5.7790%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day value at risk (VaR) Bayesian network (BN) learning algorithms at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index and the measures are based on the differences between the forecasted values of the learning algorithms and the actual returns achieved for each trading day. The total number of VaR forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Finally, the results of the Diebold-Mariano tests performed for the various 10-day 99% VaR forecasts produced by the BNs employed in this study at the 99% confidence level, are shown in Table 44, below. Recall that the null hypothesis of the Diebold-Mariano test states that the two models compared have equal forecasting abilities, while the alternative hypothesis states that the model stated second has superior forecasting abilities relative to the model stated first. In Table 44, positive Diebold-Mariano statistics lead to lower p-values, ultimately leading to the rejection of the null hypothesis, while negative Diebold-Mariano statistics lead to higher p-values, ultimately leading to the failure to reject the null hypothesis.

As can be seen from Table 44, when it comes to forecasting 10-day 99% VaR forecasts using BNs and the various network learning algorithms, the following can be concluded.

i. Every other BN learning algorithm has superior forecasting abilities relative to the MMHC learning algorithm, at the 1% significance level.
ii. The genetic algorithm has superior forecasting abilities to the MMHC algorithm and equal forecasting abilities to the PC (Stable) algorithm at the 1% significance level. The SI-HITON-PC algorithm is superior to the genetic algorithm at the 1% significance level.
iii. The PC (Stable) algorithm has superior forecasting abilities to the MMHC algorithm and equal forecasting abilities to the genetic algorithm, but is not superior to the SI-HITON-PC algorithm, all at the 1% significance level.
iv. The SI-HITON-PC algorithm is superior to all other network learning algorithms at the 1% significance level.

Table 44: Results of the Diebold-Mariano Test for the 10-day 99% Value at Risk Metric using the Bayesian Network Learning Algorithms

	Diebold-Mariano Statistic	**p-Value**
MMHC versus Genetic	6.54	3.355e-11
MMHC versus PC (Stable)	7.23	2.6270e-13
MMHC versus SI-HITON-PC	8.03	5.7070e-16
Genetic versus PC (Stable)	1.06	0.1439
Genetic versus SI-HITON-PC	2.86	0.0021
PC (Stable) versus SI-HITON-PC	2.60	0.0046

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day value at risk (VaR) Bayesian network (BN) learning algorithms at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. The total number of VaR forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

4.4.2. Stressed Value at Risk

This section discusses and analyses the results of the network learning algorithms employed to learn the structure of the BN in this study to forecast 10-day 99% SVaR forecasts for a US bank using the S&P 500 index as the underlying, over the period 15 March 1991 to 14 February 2020. First, the number of breaches experienced for each of the various models is presented and discussed, followed by an analysis of the backtesting results achieved using the BCBS's traffic light test, Kupiec's proportion of failure test, and Christoffersen's test for independence. Finally, the forecasting errors are measured using a variety of error measures, as detailed in Section 4.3.4, and these results are statistically tested for their relative predictive ability using the Diebold-Mariano test.

The breaches experienced by the various BNs used to produce the 10-day 99% SVaR forecasts are summarised in Table 45. All of the BNs employed produced no breaches at all over the 7,286-trading-day out-of-sample period. Given that all BNs produced only three breaches when used to produce 10-day 99% VaR forecasts, as shown in Table 39, and that the number of breaches is now calculated over a stressed period, it is unsurprising that no breaches are experienced when employing the BNs to produce 10-day 99% SVaR forecasts.

Table 45: Breaches Observed for the 10-day 99% Stressed Value at Risk Metric using the Bayesian Network Learning Algorithms

	SVaR Breaches
MMHC	0
Genetic	0
PC (Stable)	0
SI-HITON-PC	0

Note: This table reports the number of breaches experienced for applications of various Bayesian network (BN) learning algorithms to produce 10-day stressed value at risk (SVaR) forecasts at the 99% confidence level, using the daily logged returns of the Standard & Poor's (S&P) 500 index from 15 March 1991 to 14 February 2020. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the SVaR forecast obtained via one of the BN algorithms detailed in this table, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SVaR forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

The universal 'Green zone' result of the BCBS's traffic light test when used to evaluate the BNs used to produce 10-day 99% SVaR forecasts is unsurprising given that no breaches were experienced by each of the BNs employed. As for the conclusions reached in Chapter 3, it is still evident that the BCBS's traffic light test is only useful in concluding whether a forecasting model's breaches are statistically excessive in number relative to that expected.

Table 46: Results of the Basel Committee on Banking Supervision's Traffic Light Test for the 10-day 99% Stressed Value at Risk Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Green zone
Genetic	Green zone
PC (Stable)	Green zone
SI-HITON-PC	Green zone

Note: This table reports the results of the Basel Committee on Banking Supervision (BCBS) Traffic Light test for banks' internal models based on the number of breaches of various 10-day stressed value at risk (SVaR) models at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the various Bayesian network (BN) learning algorithms. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the SVaR forecast obtained via one of the BN algorithms detailed in this table, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SVaR forecasts for the study period was 7,286 per algorithm. The number of breaches relative to the period length was then analysed using the BCBS's Traffic Light test to achieve one of three classifications. The classifications are 'Green zone' (if the binomial probability of the number of breaches relative to the number of trading days is greater than 99.99%), 'Yellow zone' (if the corresponding probability is lesser than 99.99% but greater than 99.99%), or 'Red zone' (if the corresponding probability is lesser than 95%). Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Table 47 contains the results of the Kupiec proportion of failure test for the 10-day 99% SVaR forecasts obtained using the BNs employed in this study. As per the discussion in the previous section, the number of breaches expected for the 7,286-trading-day out-of-sample

period is the inclusive range from 50 to 94 breaches at the 99% confidence level. Given that all BNs produced no breaches over the period, the universal rejection of the null hypothesis of the Kupiec proportion of failure test is unsurprising.

Table 47: Results of the Kupiec Proportion of Failure Test for the 10-day 99% Stressed Value at Risk Metric using the Bayesian Network Learning Algorithms

	Breaches and Range	**Conclusion**
MMHC	$0 \notin [50,94]$	Reject
Genetic	$0 \notin [50,94]$	Reject
PC (Stable)	$0 \notin [50,94]$	Reject
SI-HITON-PC	$0 \notin [50,94]$	Reject

Note: This table reports the results of the Kupiec Proportion of Failure (PoF) backtest at the 1% significance level based on the number of breaches of various 10-day stressed value at risk (SVaR) Bayesian network (BN) learning algorithms at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the number of true breaches of the model is equal to the observed number of breaches. The results are based on the breaches of the logged returns earned on the Standard & Poor's (S&P) 500 index. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the SVaR forecast obtained via one of the BN algorithms detailed in this table, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SVaR forecasts for the study period was 7,286 per algorithm. The left column shows the number of breaches experienced under the BN learning algorithm together with the range predicted using the PoF test. Since the period has 7,286 trading days, the range of expected breaches is $[50,94]$ at the 99% significance level. The critical value of the χ_1^2 at the 99% confidence level is 6.64897. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Table 48: Results of the Christoffersen Test for Independence for the 10-day 99% Stressed Value at Risk Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Fail to reject
Genetic	Fail to reject
PC (Stable)	Fail to reject
SI-HITON-PC	Fail to reject

Note: This table reports the results of the Christoffersen test of independence backtest for banks' internal models based on the number of breaches of various 10-day stressed value at risk (SVaR) Bayesian network (BN) learning algorithms at the 99% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the probability of a breach following a non-breach is equal to the probability of a breach following a breach, i.e., breaches are independent. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the SVaR forecast obtained via one of the BN algorithms detailed in this table, where the SVaR forecast is calculated over the most severe one-year period in the period preceding the return's date, i.e., over a stressed period. The total number of SVaR forecasts for the study period was 7,286 per algorithm. The critical value of the χ_1^2 at the 99% confidence level is 6.63490. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

The final backtest employed to test the 10-day 99% SVaR forecasts produced by the various BNs employed in this study is Christoffersen's test for independence, the results of which are depicted in Table 48, above. Since no breaches were experienced by any of the BNs employed

in this study, it is not possible to conclude that there exists any dependence between the non-existent breaches. Hence, the backtest's null hypothesis cannot be rejected at the 1% significance level, or at any significance level.

It is evident that the statistical backtests available to US banks are of little use when it comes to statistically assessing models that produce no breaches. The only useful analysis is, therefore, that related to the relative efficiency of the 10-day 99% SVaR forecasts produced by the various BNs employed in this study. Table 49, below, depicts the results of the various forecasting error measures employed in this study, namely the MAE, the RMSE, the MAPE, and the MdAPE, as calculated for the various BNs used to produce the 10-day 99% SVaR forecasts.

Table 49: Forecasting Error Measures for the 10-day 99% Stressed Value at Risk Metric using the Bayesian Network Learning Algorithms

	MAE	**RMSE**	**MAPE**	**MdAPE**
MMHC	0.1465	0.1523	154.6608%	8.2242%
Genetic	0.1427	0.1483	151.5520%	8.0387%
PC (Stable)	0.1482	0.1544	156.9743%	8.2343%
SI-HITON-PC	0.1482	0.1544	156.9730%	8.2343%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day stressed value at risk (SVaR) Bayesian network (BN) learning algorithms at the 99% confidence levels over the period 15 March 1991 to 14 February 2020, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index and the measures are based on the differences between the forecasted values of the learning algorithms and the actual returns achieved for each trading day. The total number of SVaR forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Unlike the discussion in the preceding section, the results of the various forecasting error measures with respect to the 10-day 99% SVaR forecasts differ more than their unstressed counterparts. This may be due to how a stressed period is calibrated in this study. Recall that the stressed period used to calculate the 10-day 99% SVaR forecasts includes the worst returns experienced over the period, i.e., a stressed period. This period would contain a BN-produced forecast if said forecast would result in the next day's expected return to be included in the stressed period. On the next day, the previous day's return is now known, and the expected return is replaced with the actual return. This return may or may not be included in the stressed period. Hence, the differences in the forecasting error measures detailed in Table 49 may be due to the discrepancies in the one-day-ahead forecasts produced by the BN learning algorithms employed in this study. This may be arguably more meaningful than the differences in 10-day

99% VaR, as discussed in the previous section, given the risk management focus of the bank in this context.

As seen in Table 49, the least accurate 10-day 99% SVaR forecasts, as measured by all four forecasting error measures, were produced using the PC (Stable) algorithm. Performing marginally better on only one of the four forecasting error measures is the SI-HITON-PC algorithm (which ranked as the most accurate algorithm when producing 10-day 99% VaR forecasts). On the other hand, the most accurate 10-day 99% SVaR forecasts were produced using the genetic algorithm. Given the discussion surrounding the NA values experienced when producing the closing values of the S&P 500 index using the genetic algorithm in Section 4.3.2, it is perhaps important to note the second-most accurate BN algorithm, i.e., that using the MMHC algorithm.

Finally, consider the results of the Diebold-Mariano tests performed for the various 10-day 99% SVaR forecasts produced by the BNs employed in this study at the 99% confidence level, as depicted in Table 50, below. Once again, recall that the null hypothesis of the Diebold-Mariano test states that the forecasting abilities of the two models tested are equal, while the alternative hypothesis states that the second model stated has superior forecasting abilities relative to the first model stated. Again, positive Diebold-Mariano statistics in Table 50 lead to lower p-values, ultimately leading to the rejection of the null hypothesis, while negative statistics lead to higher p-values, ultimately leading to the failure to reject the null hypothesis.

As can be seen from Table 50, when it comes to forecasting 10-day 99% SVaR forecasts using BNs and the various network learning algorithms, the following can be concluded.

i. The PC (Stable) and SI-HITON-PC algorithms have equal forecasting abilities to the MMHC algorithm, while the genetic algorithm has superior forecasting abilities relative to the MMHC algorithm, at the 1% significance level.
ii. Every other BN learning algorithm has equal or inferior forecasting abilities relative to the genetic algorithm at the 1% significance level.
iii. The PC (Stable) algorithm has equal forecasting abilities to the MMHC algorithm and the genetic algorithm, and worse forecasting abilities relative to the SI-HITON-PC algorithm, at the 1% significance level.

iv. The SI-HITON-PC algorithm has superior forecasting abilities relative to the PC (Stable) algorithm, and equal forecasting abilities to the MMHC algorithm and the genetic algorithm, at the 1% significance level.

Table 50: Results of the Diebold-Mariano Test for the 10-day 99% Stressed Value at Risk Metric using the Bayesian Network Learning Algorithms

	Diebold-Mariano Statistic	**p-Value**
MMHC versus Genetic	62.85	<2.2e-16
MMHC versus PC (Stable)	-53.38	1
MMHC versus SI-HITON-PC	-53.32	1
Genetic versus PC (Stable)	-91.96	1
Genetic versus SI-HITON-PC	-91.90	1
PC (Stable) versus SI-HITON-PC	2.48	0.0066

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day stressed value at risk (SVaR) Bayesian network (BN) learning algorithms at the 99% confidence levels over the period 15 March 1991 to 14 February 2020, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. The total number of SVaR forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

4.4.3. Expected Shortfall

This section discusses and analyses the results of the network learning algorithms employed to learn the structure of the BN in this study to forecast 10-day 97.5% ES forecasts for a US bank using the S&P 500 index as the underlying, over the period 15 March 1991 to 14 February 2020. First, the number of breaches experienced for each of the various models is presented and discussed, followed by an analysis of the backtesting results achieved using the BCBS's traffic light test, the conditional, the unconditional, and the minimally biased backtests. Finally, the forecasting errors are measured using a variety of error measures, as detailed in Section 4.3.4, and these results are statistically tested for their relative predictive ability using the Diebold-Mariano test.

Table 51, below, shows the breaches experienced for each of the BNs used to obtain 10-day 97.5% ES forecasts during the 7,286-trading-day out-of-sample test period. Each of the four BNs experienced only three breaches during the out-of-sample test period. The dates on which the breaches were observed are the same three dates across the four BNs, and correspond to

the dates on which the BNs experienced breaches when forecasting 10-day 99% VaR forecasts (27 October 1997, 31 August 1998, and 29 September 2008; see Section 4.4.1).

Table 51: Breaches Observed for the 10-day 97.5% Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	ES Breaches
MMHC	3
Genetic	3
PC (Stable)	3
SI-HITON-PC	3

Note: This table reports the number of breaches experienced for applications of various Bayesian network (BN) learning algorithms to produce 10-day expected shortfall (ES) forecasts at the 97.5% confidence level, using the daily logged returns of the Standard & Poor's (S&P) 500 index from 15 March 1991 to 14 February 2020. An ES breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the ES forecast obtained via one of the BN algorithms detailed in this table. The total number of ES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Table 52: Results of the Basel Committee on Banking Supervision's Traffic Light Test for the 10-day 97.5% Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Green zone
Genetic	Green zone
PC (Stable)	Green zone
SI-HITON-PC	Green zone

Note: This table reports the results of the Basel Committee on Banking Supervision (BCBS) Traffic Light test for banks' internal models based on the number of breaches of various 10-day expected shortfall (ES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the various Bayesian network (BN) learning algorithms. An ES breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the ES forecast obtained via one of the BN algorithms detailed in this table. The total number of ES forecasts for the study period was 7,286 per algorithm. The number of breaches relative to the period length was then analysed using the BCBS's Traffic Light test to achieve one of three classifications. The classifications are 'Green zone' (if the binomial probability of the number of breaches relative to the number of trading days is greater than 99.99%), 'Yellow zone' (if the corresponding probability is lesser than 99.99% but greater than 99.99%), or 'Red zone' (if the corresponding probability is lesser than 95%). Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

The results of the BCBS's traffic light test for the 10-day 97.5% ES forecasts produced by the BNs employed in this study are shown in Table 52, above. Once again, the extremely low numbers of breaches experienced by the various BNs unsurprisingly lead to a universal 'Green zone' outcome when using the BCBS's traffic light test. Hence, the use of BNs to produce 10-day 97.5% ES forecasts would yield models that are accepted from a regulatory standpoint.

Turning to statistical backtests beyond the BCBS's traffic light test, the first ES-specific backtest employed is the conditional backtest. Using this backtest, the 10-day 97.5% ES forecasts were evaluated together with their respective 10-day VaR forecasts. Given that all of the learning algorithms used in this study experienced at least one corresponding 10-day VaR breach, this conditional backtest can be carried out across all learning algorithms. The null hypothesis states the observed 10-day 97.5% ES forecasts are the true 10-day 97.5% ES forecasts. This null hypothesis is rejected at the 97.5% confidence level for each of the four learning algorithms used to produce 10-day 97.5% ES forecasts in this study, as shown in Table 53, below.

As for the conditional backtest, the next backtest, the unconditional backtest, is also dependent on the existence of at least one corresponding 10-day VaR breach. Once again, since all learning algorithms experienced at least one breach, the unconditional backtest can be carried out for all four learning algorithms. The null hypothesis and alternative hypothesis of the unconditional backtest are the same as those of the conditional backtest. Each of the BNs' set of 10-day 97.5% ES forecasts results in the rejection of the null hypothesis at the 97.5% confidence level, as shown in Table 54, below.

Table 53: Results of the Conditional Backtest for the 10-day 97.5% Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Reject
Genetic	Reject
PC (Stable)	Reject
SI-HITON-PC	Reject

Note: This table reports the results of the conditional backtest for banks' internal models based on the number of breaches of various 10-day expected shortfall (ES) Bayesian network (BN) learning algorithms at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the ES forecasts observed are the true ES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. The test can only be carried out if there is at least one value at risk (VaR) breach. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the algorithms detailed in this table. The total number of ES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Table 54: Results of the Unconditional Backtest for the 10-day 97.5% Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Reject
Genetic	Reject
PC (Stable)	Reject
SI-HITON-PC	Reject

Note: This table reports the results of the unconditional backtest for banks' internal models based on the number of breaches of various 10-day expected shortfall (ES) Bayesian network (BN) learning algorithms at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the ES forecasts observed are the true ES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. The test can only be carried out if there is at least one value at risk (VaR) breach. A VaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the algorithms detailed in this table. The total number of ES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Table 55: Results of the Minimally Biased Backtest for the 10-day 97.5% Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Reject
Genetic	Reject
PC (Stable)	Reject
SI-HITON-PC	Reject

Note: This table reports the results of the minimally biased backtest for banks' internal models based on the number of breaches of various 10-day expected shortfall (ES) Bayesian network (BN) learning algorithms at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the ES forecasts observed are the true ES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. The total number of ES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

The final ES-specific backtest employed in this section to statistically evaluate the 10-day 97.5% ES forecasts produced by the BNs is the minimally biased backtest. Its null hypothesis and alternative hypothesis are the same as those of the previous two backtests. Unlike the previous two backtests, it does not depend on the existence of at least one corresponding 10-day VaR breach, even though this condition would have been fulfilled had it existed. As shown in Table 55, above, once again, the null hypothesis is rejected across the board at the 97.5% confidence level, leading to the conclusion that all BNs used in this study do not produce accurate 10-day 97.5% ES forecasts.

Table 56, below, depicts the MAE, the RMSE, the MAPE, and the MdAPE values for the various 10-day 97.5% ES forecasts produced by the BNs employed in this study. As can be seen in Table 56, as for the forecasting error measures' results for the 10-day 99% VaR

forecasts shown in Table 43, the various BNs achieve very similar results. This is, once again, expected, as the PDFs of the BNs only use a single one-day-ahead forecast when calculating the 10-day 97.5% ES forecasts, while the (much larger) remainder of the calibration period is made up of historical returns.

Table 56: Forecasting Error Measures for the 10-day 97.5% Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	MAE	**RMSE**	**MAPE**	**MdAPE**
MMHC	0.0918	0.0988	96.9988%	5.2595%
Genetic	0.0917	0.0987	96.9736%	5.2595%
PC (Stable)	0.0917	0.0987	96.9737%	5.2595%
SI-HITON-PC	0.0917	0.0987	96.9730%	5.2595%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day expected shortfall (ES) Bayesian network (BN) learning algorithms at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index and the measures are based on the differences between the forecasted values of the learning algorithms and the actual returns achieved for each trading day. The total number of ES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Once again, the differences in the forecasting error measures' values come down to the quality of the forecasts made by the BNs employed. Hence, the BN algorithm that produced the most accurate forecasts is the SI-HITON-PC algorithm, as shown in Table 56. While the genetic algorithm and the PC (Stable) algorithm both scored the same for the MAE, the RMSE, and the MdAPE when compared to the SI-HITON-PC algorithm, the genetic algorithm ranks as the second-most accurate algorithm after scoring marginally better than the PC (Stable) algorithm when using the MAPE. Hence, the genetic algorithm ranks as the second-most accurate algorithm when evaluating the learning algorithms employed by BNs to produce 10-day 97.5% ES forecasts, while the PC (Stable) algorithm ranks third-most accurate. The MMHC algorithm ranks as the least accurate learning algorithm when employed by a BN to produce 10-day 97.5% ES forecasts.

Last, attention is turned to the performances of the various models when applying the Diebold-Mariano test at the 99% confidence level to statistically determine which models produce more accurate 10-day 97.5% ES forecasts. Table 57 shows the results of the Diebold-Mariano test for the various BN learning algorithms used in this study. As stated in the preceding sections, positive statistics in Table 57 are accompanied by higher p-values, and

suggest that the null hypothesis should be rejected, while the opposite is true for negative statistics.

Table 57: Results of the Diebold-Mariano Test for the 10-day 97.5% Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	Diebold-Mariano Statistic	**p-Value**
MMHC versus Genetic	6.79	5.8940e-12
MMHC versus PC (Stable)	6.57	2.6590e-11
MMHC versus SI-HITON-PC	7.23	2.6050e-13
Genetic versus PC (Stable)	-0.33	0.6277
Genetic versus SI-HITON-PC	3.61	0.0002
PC (Stable) versus SI-HITON-PC	2.19	0.0143

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day expected shortfall (ES) Bayesian network (BN) learning algorithms at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. The total number of ES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

The results depicted in Table 57 are summarised as follows, as they pertain to 10-day 97.5% ES forecasts produced using the various BN learning algorithms employed in this study.

i. Every other BN learning algorithm has superior forecasting abilities relative to the MMHC learning algorithm, at the 1% significance level.
ii. The genetic algorithm has superior forecasting abilities relative to the MMHC algorithm, equal forecasting abilities relative to the PC (Stable) algorithm, and inferior forecasting abilities relative to the SI-HITON-PC algorithm, all at the 1% significance level.
iii. The PC (Stable) algorithm has superior forecasting abilities relative to the MMHC algorithm, and equal forecasting abilities relative to the genetic algorithm and the SI-HITON-PC algorithm, all at the 1% significance level[34].

[34] The p-value of the Diebold-Mariano test comparing the PC (Stable) algorithm to the SI-HITON-PC algorithm is 0.0143.

iv. The SI-HITON-PC algorithm has superior forecasting abilities relative to the MMHC algorithm and the genetic algorithm, but equal forecasting abilities relative to the PC (Stable) algorithm, all at the 1% significance level.

4.4.4. Stressed Expected Shortfall

This section discusses and analyses the results of the network learning algorithms employed to learn the structure of the BN in this study to forecast 10-day 97.5% SES forecasts for a US bank using the S&P 500 index as the underlying, over the period 15 March 1991 to 14 February 2020. First, the number of breaches experienced for each of the various models is presented and discussed, followed by an analysis of the backtesting results achieved using the BCBS's traffic light test, the conditional, the unconditional, and the minimally biased backtests. Finally, the forecasting errors are measured using a variety of error measures, as detailed in Section 4.3.4, and these results are statistically tested for their relative predictive ability using the Diebold-Mariano test.

Table 58: Breaches Observed for the 10-day 99% Stressed Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	SES Breaches
MMHC	0
Genetic	0
PC (Stable)	0
SI-HITON-PC	0

Note: This table reports the number of breaches experienced for applications of various Bayesian network (BN) learning algorithms to produce 10-day stressed expected shortfall (SES) forecasts at the 97.5% confidence level, using the daily logged returns of the Standard & Poor's (S&P) 500 index from 15 March 1991 to 14 February 2020. A SES breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the SES forecast obtained via one of the BN algorithms detailed in this table, where the SES forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

The breaches observed for the various BN algorithms used to produce the 10-day 97.5% SES forecasts are summarised in Table 58, above. As seen for the 10-day 99% SVaR forecasts in Section 4.4.2, all of the BNs employed produced no breaches at all over the 7,286-trading-day out-of-sample period. Since all of the BNs produced very few breaches when producing 10-day 97.5% ES forecasts (see Table 51), it is unsurprising that the number of breaches over the stressed period is zero for each of the BNs employed in this study.

Table 59: Results of the Basel Committee on Banking Supervision's Traffic Light Test for the 10-day 97.5% Stressed Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Green zone
Genetic	Green zone
PC (Stable)	Green zone
SI-HITON-PC	Green zone

Note: This table reports the results of the Basel Committee on Banking Supervision (BCBS) Traffic Light test for banks' internal models based on the number of breaches of various 10-day stressed expected shortfall (SES) models at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index using the various Bayesian network (BN) learning algorithms. A SES breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the SES forecast obtained via one of the BN algorithms detailed in this table, where the SES forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SES forecasts for the study period was 7,286 per algorithm. The number of breaches relative to the period length was then analysed using the BCBS's Traffic Light test to achieve one of three classifications. The classifications are 'Green zone' (if the binomial probability of the number of breaches relative to the number of trading days is greater than 99.99%), 'Yellow zone' (if the corresponding probability is lesser than 99.99% but greater than 99.99%), or 'Red zone' (if the corresponding probability is lesser than 95%). Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Next, the BCBS's traffic light test was employed to assess the 10-day 97.5% SES forecasts produced by the various BNs employed in this study, resulting in a universal 'Green zone' result, as shown in Table 59, above. Given the lack of breaches for each of the BNs employed, the BCBS's traffic light test cannot detect that any of the BNs tested produced excessive breaches.

Turning to backtests which are more interesting from a statistical standpoint, the conditional backtest was applied to the 10-day 97.5% SES forecasts produced by the BNs employed in this study. Its null hypothesis states that the observed 10-day 97.5% SES forecasts observed were the true 10-day 97.5% SES forecasts. Recall that the conditional backtest can be employed only if there exists at least one corresponding 10-day SVaR breach in the out-of-sample period. As discussed in Section 4.4.2 and shown in Table 45, no 10-day SVaR breaches were observed for any of the BNs, rendering this backtest unemployable. This universal result is depicted in Table 60, below.

Table 60: Results of the Conditional Backtest for the 10-day 97.5% Stressed Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Cannot perform backtest
Genetic	Cannot perform backtest
PC (Stable)	Cannot perform backtest
SI-HITON-PC	Cannot perform backtest

Note: This table reports the results of the conditional backtest for banks' internal models based on the number of breaches of various 10-day stressed expected shortfall (SES) Bayesian network (BN) learning algorithms at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the SES forecasts observed are the true SES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. The test can only be carried out if there is at least one stressed value at risk (SVaR) breach. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the algorithms detailed in this table, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

As for the conditional backtest, the unconditional backtest's null hypothesis also states that the observed 10-day 97.5% SES forecasts observed were the true 10-day 97.5% SES forecasts. It, too, can only be employed for the 10-day 97.5% SES forecasts if there exists at least one corresponding 10-day SVaR breach in the out-of-ample period. Hence, it, too, cannot be employed, given the zero 10-day SVaR breaches experienced in the out-of-sample period. This universal result is depicted in Table 61, below.

Table 61: Results of the Unconditional Backtest for the 10-day 97.5% Stressed Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Cannot perform backtest
Genetic	Cannot perform backtest
PC (Stable)	Cannot perform backtest
SI-HITON-PC	Cannot perform backtest

Note: This table reports the results of the unconditional backtest for banks' internal models based on the number of breaches of various 10-day stressed expected shortfall (SES) Bayesian network (BN) learning algorithms at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the SES forecasts observed are the true SES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. The test can only be carried out if there is at least one stressed value at risk (SVaR) breach. A SVaR breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the VaR forecast obtained via one of the algorithms detailed in this table, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Table 62: Results of the Minimally Biased Backtest for the 10-day 97.5% Stressed Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	Backtest Result
MMHC	Reject
Genetic	Reject
PC (Stable)	Reject
SI-HITON-PC	Reject

Note: This table reports the results of the minimally biased backtest for banks' internal models based on the number of breaches of various 10-day stressed expected shortfall (SES) Bayesian network (BN) learning algorithms at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020. The test's null hypothesis states that the SES forecasts observed are the true SES figures. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. A SES breach was recorded where the loss incurred on the S&P 500 index (as measured by its daily logged return) exceeded the ES forecast obtained via one of the algorithms detailed in this table, where the SES forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The total number of SES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Finally, the minimally biased backtest was employed. As for the conditional backtest and the unconditional backtest, its null hypothesis also states that the observed 10-day 97.5% SES forecasts observed were the true 10-day 97.5% SES forecasts. However, unlike the two preceding backtests, it does not rely on the existence of at least one breach for the corresponding 10-day SVaR forecasts. Table 62, above, summarises the results of the backtest, showing the universal rejection of the null hypothesis for all of the BNs employed in this study. Hence, it is concluded that none of the BN employed produced statistically accurate 10-day 97.5% SES forecasts.

The backtests employed in this section produce little comfort surrounding the statistical accuracy of the 10-day 97.5% SES forecasts produced by the various BNs, where such backtests could be employed. The relative efficiency of each of the BNs employed when producing 10-day 97.5% SES forecasts was evaluated next. The MAE, the RMSE, the MAPE, and the MdAPE were calculated for each of the BNs used to produce the 10-day 97.5% SES forecasts.

As seen in Table 63, below, the various BNs employed in this study produced relatively similar forecasting error measures across the MAE and the RMSE. The PC (Stable) learning algorithm was the most accurate when producing 10-day 97.5% SES forecasts based on the MAE and the RMSE, with all other algorithms scoring marginally higher (i.e., worse) for each of the forecasting error measures. The PC (Stable) algorithm also scored better than the other algorithms when using the MAPE and the MdAPE, with the other three algorithms scored equally more when using the MdAPE. Using the MAPE, the second-most accurate algorithm

was the SI-HITON-PC algorithm, followed by the genetic algorithm. The MMHC algorithm was the least accurate algorithm based on the MAPE, scoring the highest for the measure between the various algorithms.

Table 63: Forecasting Error Measures for the 10-day 97.5% Stressed Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	MAE	**RMSE**	**MAPE**	**MdAPE**
MMHC	0.1700	0.1728	178.4088%	9.9945%
Genetic	0.1700	0.1728	178.3908%	9.9945%
PC (Stable)	0.1698	0.1727	178.3133%	9.9685%
SI-HITON-PC	0.1700	0.1728	178.3905%	9.9945%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day stressed expected shortfall (SES) Bayesian network (BN) learning algorithms at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020, where the SES forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index and the measures are based on the differences between the forecasted values of the learning algorithms and the actual returns achieved for each trading day. The total number of SES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

Finally, the Diebold-Mariano test was applied to the various BNs used to produce 10-day 97.5% SES forecasts, with the results of these tests, tested at the 99% confidence level, summarised in Table 64, below. As previously stated, positive Diebold-Mariano statistics are accompanied by lower p-values, and suggest that the null hypothesis should be rejected, while the opposite is true for negative statistics.

The results depicted in Table 64 are summarised as follows, as they pertain to the 10-day 97.5% SES forecasts produced by the BNs employed in this study.

i. Every other BN learning algorithm has superior forecasting abilities relative to the MMHC algorithm at the 1% significance level.
ii. The PC (Stable) algorithm has superior forecasting abilities relative to the genetic algorithm, while the SI-HITON-PC algorithm has equal forecasting abilities[35], and the MMHC algorithm has inferior forecasting abilities, all at the 1% significance level.
iii. The PC (Stable) algorithm has superior forecasting abilities relative to all other learning algorithms at the 1% significance level.

[35] The p-value of this test is 0.0157.

iv. The SI-HITON-PC algorithm has superior forecasting abilities relative to the MMHC algorithm, and equal or inferior forecasting abilities relative to all other algorithms, all at the 1% significance level.

Table 64: Results of the Diebold-Mariano Test for the 10-day 97.5% Stressed Expected Shortfall Metric using the Bayesian Network Learning Algorithms

	Diebold-Mariano Statistic	**p-Value**
MMHC versus Genetic	6.49	4.5330e-11
MMHC versus PC (Stable)	24.61	<2.2e-16
MMHC versus SI-HITON-PC	6.68	1.2540e-11
Genetic versus PC (Stable)	30.01	<2.2e-16
Genetic versus SI-HITON-PC	2.15	0.0157
PC(Stable) versus SI-HITON-PC	-30.09	1

Note: This table reports the results of the Diebold-Mariano test, testing the forecast accuracies of banks' internal models based on the forecasting errors of various 10-day stressed expected shortfall (SES) Bayesian network (BN) learning algorithms at the 97.5% confidence levels over the period 15 March 1991 to 14 February 2020, where the SES forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The test's null hypothesis states that the two models compared have the same level of forecasting accuracy, while the alternative hypothesis is that the second model named has greater forecasting accuracy relative to the first model named. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index. The total number of SES forecasts for the study period was 7,286 per algorithm. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

4.5. Conclusion

This chapter introduced BNs to the literature on market risk management by banks by using four learning algorithms to produce 10-day 99% VaR forecasts, 10-day 97.5% ES forecasts, and their stressed counterparts. Similarly to Chapter 3, these forecasts were produced using historical return data for the S&P 500 index from 15 March 1991 to 14 February 2020. However, unlike the methodology used in Chapter 3, the use of BNs introduced a forward-looking element to market risk management, producing one-day-ahead forecasts for the closing values of the S&P 500 index and incorporating the expected returns on those into the return distribution used to produce the market risk forecasts.

Across all measures, all BNs yielded very few breaches (if any), and all sets of forecasts were classified as 'Green zone' models using the BCBS's traffic light test. Hence, while it is prudent to conclude that the BCBS's traffic light test is of little use in the practical testing of the accuracy of either VaR forecasts or ES forecasts, or their stressed versions, as per the conclusions made in Chapter 3, it is concluded that the introduction to and use of BNs in

producing these market risk metrics does not render the resulting forecasts and models impermissible by the regulatory test dictated by the BCBS. Moreover, the use of BNs still resulted in few breaches, regardless of which of the learning algorithms was employed. Hence, the results of the use of BNs are in line with the literature, showing that banks prefer few breaches to no breaches (see Chapter 3 and, for example, McAleer and da Veiga, 2008).

Once again, the supporting statistical backtests do not add much value, given the level of conservatism displayed by the BNs' forecasts across all four market risk metrics. This is a function of the equal-weight assigned to each trading day in the risk metric calibration period, where one such trading day is the BN's one-day-ahead forecast, while the remainder consists of historical returns. Hence, as per the conclusions for Chapter 3, BNs, too, produce conservative forecasts across all four market risk metrics and, therefore, do not stray from the existing literature covering traditional models (see Section 3.5).

Across the learning algorithms used to produce 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts, the SI-HITON-PC algorithm produced the most accurate forecasts. When considering the stressed metrics, there was no consensus between those. The genetic algorithm and the MMHC algorithm produced the most accurate 10-day 99% SVaR forecasts[36]. The PC (Stable) algorithm produced the most accurate 10-day 97.5% SES forecasts.

The primary finding of this chapter is that the use of BNs to forecast market risk metrics produced results that are comparable to those achieved using the traditional market risk models discussed in Chapter 3. That being said, the direct application of BNs to produce return distributions and forecasts did not produce market risk forecasts that are any more accurate than the commonly used historical simulation model.

For example, the 10-day 99% VaR forecasts produced using a BN and the SI-HITON-PC learning algorithm produced the same forecasting error measure values to four decimal places as its historical simulation model counterpart. This, and the forecasting error measures of the other three market risk metrics when it comes to both the BN application and the historical simulation model, are discussed further in Chapter 5. As seen in Chapter 3, when it comes to 10-day 99% VaR forecasts, the EGARCH model proved to produce forecasts that were more

[36] The genetic algorithm produced the most accurate 10-day 99% SVaR forecasts, while the MMHC algorithm was a close second. Due to the treatment of NA values produced by the genetic algorithm (see Section 4.3.2), the MMHC algorithm is also mentioned, as the genetic algorithm's more accurate results may be a function of the intervention discussed.

accurate than those produced using the historical simulation model, which scored equally to the forecasts produced using the BN and the SI-HITON-PC algorithm.

This makes sense when considering that the return PDF consists of the forecast of the closing value of the S&P 500 index and, by extension, the next trading day's forecasted return, produced by the BN, which only constitutes one part of the 1,264-part calibration period. The remaining 1,263 parts are the historical returns observed in the past. Hence, even though a BN incorporates a forward-looking methodology, which should, in theory, yield more accurate market risk metrics, the equal weighting of this BN-derived one-day-ahead return forecast relative to the rest of the return distribution essentially has minimal impact.

Giving the forecast observation more weight would require either a shortening of the calibration period length or a different weighting methodology. Looking again at the 10-day 99% VaR metric, the MAE values for the historical simulation model, the MMHC algorithm, the genetic algorithm, the PC (Stable) algorithm, and the SI-HITON-PC algorithm, for calibration periods of length 1,264 days (the original length used in this study), 250 days, 25 days, and five days, are summarised in Table 65. Table 65 illustrates what effect a change in the weighting of the forward-looking nature of the BNs employed in this study would have on the differences in error measures.

Table 65: The Impact of Changes in Calibration Period Length on the Mean Absolute Error of Various Models for the 10-day 99% Value at Risk Metric

Period Length	**1,264**	**250**	**25**	**5**
Historical simulation	0.0944	0.0865	0.0634	0.0372
MMHC	0.0945	0.0867	0.0681	0.0428
Genetic	0.0944	0.0866	0.0638	0.0382
PC (Stable)	0.0944	0.0866	0.0637	0.0377
SI-HITON-PC	0.0944	0.0865	0.0634	0.0372

Note: This table reports the mean absolute errors for the various 10-day 99% value at risk (VaR) models for different calibration period lengths. Bayesian network (BN) learning algorithms and the historical simulation model were used to produce 10-day 99% VaR forecasts over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard & Poor's (S&P) 500 index and the measures are based on the differences between the forecasted values of the learning algorithms and the actual returns achieved for each trading day. The total number of VaR forecasts for the study period was 7,286 per algorithm and per model. Note that the BN learning algorithms considered were the max-min hill-climbing (MMHC) algorithm, the genetic algorithm, the Peter and Clark Stable (PC (Stable)) algorithm, and the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm.

These results highlight two conclusions. First, the MAE values of the historical simulation model and the BN using the SI-HITON-PC algorithm are always the lowest of the five models, indicating that the conclusion regarding the higher accuracy of the SI-HITON-PC algorithm

when producing 10-day 99% VaR forecasts is insensitive to changes in the calibration period length, at least when the MAE is used. Second, the differences in MAE values of the different models increase significantly with the reduction in calibration period length or, equivalently, with the increased weight of the forecast relative to the historical returns making up the remainder of the calibration period.

This second finding suggests that the weighting of the forecasted return produced by the BN in the return PDF relative to the other (historical) returns making up the return distribution may yield more accurate market risk forecasts than a non-weighted application as carried out in this chapter. Specifically, the findings in Chapter 3 show that the weightings incorporated by different autoregressive models produced more accurate market risk metric forecasts relative to the historical simulation model. Hence, an autoregressive application incorporating the forecasts produced by the BN learning algorithms may be even more beneficial than any of these methods in isolation. This novel combined approach, termed integrated forecast dynamic Bayesian networks (IFDBNs), is developed and applied in the next chapter.

5. Market Risk Management using Integrated Forecast Dynamic Bayesian Networks

This chapter develops a novel methodological approach to the calibration and specification of Bayesian networks (BNs). For this approach, the term integrated forecast dynamic Bayesian networks (IFDBNs) is coined. The conclusion to the previous chapter highlighted the potential advantages of changes to the weight assigned to the forecasted return relative to the remaining historical returns used to produce a return probability density function (PDF). Therefore, to address this, the IFDBN methodology developed in this chapter combines the forward-looking forecasting of the BNs introduced in Chapter 4 and the weight-assigning functionality of the autoregressive models introduced in Chapter 3.

Specifically, the autoregressive model identified as the most accurate traditional model in Chapter 3 (as these autoregressive models assign weightings to the returns making up the return PDF), and the BN learning algorithm identified as the most accurate learning algorithm in Chapter 4, are combined to produce forecasts for each of the four market risk metrics considered in this study. This integration essentially takes the forecasted return from the BN and then applies the autoregressive model's weighting to the set of returns. In this way, the IFDBN methodology allows for greater weight to be placed on the next trading day's forecasted return in the return PDF and, by extension, yields a market risk metric that places more emphasis on the forward-looking BN return forecast.

The IFDBN approach presented in this chapter was applied to the context of producing 10-day 99% value at risk (VaR) forecasts and 10-day 97.5% expected shortfall (ES) forecasts, and their stressed counterparts, from 15 March 1991 to 14 February 2020 for the equities trading desk of a United States (US) bank in the context of the regulations put forward by the Basel Committee on Banking Supervision (BCBS). However, the novelty of the IFDBN methodology of combining the BN forecasts with the weight-assigning functionality of autoregressive models is applicable beyond the confines of this setting.

The remainder of this chapter is structured as follows. First, this chapter's data and methodology are discussed, where the autoregressive models and BN learning algorithms identified in the previous chapters are revisited to specify each of the IFDBNs used for each of the four market risk metrics. Next, the results of the IFDBNs employed, together with the results of the best-in-class models from the previous chapters and the historical simulation

model are discussed, the latter included due to its wide popularity in practice. Finally, a conclusion surrounding the performances of the IFDBNs is provided.

5.1. Data and Methodology

This section introduces the novel methodology to construct IFDBNs. Each of the four market risk metrics explored in this study has its respective combination of autoregressive traditional model and BN learning algorithm which, individually, yielded its most accurate forecasts in Chapters 3 and 4, respectively. The data used in this chapter were the historical returns of the S&P 500 index and the one-day-ahead forecasts of the closing values of the S&P 500 index, which were used to produce forecast returns in Chapter 4.

The results of Chapter 3, discussed in Section 3.4, show that, for each of the four market risk metrics in this study, autoregressive models produced more accurate forecasts. Specifically, the exponential generalised autoregressive conditional heteroscedasticity (EGARCH) model produced the most accurate 10-day 99% VaR forecasts (see Section 3.4.1) and 10-day 97.5% ES forecasts (see Section 3.4.3), while the generalised autoregressive conditional heteroscedasticity (GARCH) model produced the most accurate 10-day 99% stressed VaR (SVaR) forecasts (see Section 3.4.2) and 10-day 97.5% stressed ES (SES) forecasts (see Section 3.4.4).

On the other hand, the results of Chapter 4, discussed in Section 4.4, show that, for each of the four market risk metrics in this study, different BN learning algorithms perform better for the different market risk metrics. Specifically, the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm produced the most accurate 10-day 99% VaR forecasts (see Section 4.4.1) and 10-day 97.5% ES forecasts (see Section 4.4.3). For the 10-day 99% SVaR forecasts, the genetic algorithm proved the most accurate, although this may be accidental due to the treatment of any NAs produced by the algorithm (see Section 4.3.2 for the relevant discussion). The algorithm that produced the next most accurate 10-day 99% SVaR forecasts was the max-min hill-climbing (MMHC) algorithm (see Section 4.4.2). Hence, both algorithms are considered in this chapter. Finally, the Peter and Clark (PC) (Stable) algorithm produced the most accurate 10-day 97.5% SES forecasts (see Section 4.4.4).

Hence, the following combinations of traditional market risk management models and BN learning algorithms are combined for the various market risk metrics. The IFDBN used to produce 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts is created by integrating the exponential weighting properties of the EGARCH model (see Section 3.1.3.3) and the most

accurate forecasts produced by the SI-HITON-PC algorithm (see Section 4.1.4.4). Next, 10-day 99% SVaR forecasts were produced using two IFDBNs that use the GARCH model (see Section 3.1.3.2) as the underlying forecasting model: One that uses the genetic algorithm (see Section 4.1.4.1), and another that uses the MMHC algorithm (see Section 4.1.4.3). Finally, 10-day 97.5% SES forecasts were produced using an IFDBN that combines the GARCH model and the PC (Stable) algorithm (see Section 4.1.4.2).

The use of the autoregressive processes as the basis for increasing the relative weight of the dynamic BNs' (DBNs') forecasts relative to the remainder of the calibration period may potentially yield more accurate market risk metric forecasts. Moreover, it also mimics the process a risk manager will implement in practice, i.e., awarding more weight to the next day's forecast relative to past market movements, without disregarding the latter completely.

5.2. Results

This section discusses the results using the data and methodology discussed in the previous section. Analyses of the results are also provided. The results of the forecasts of each of the four market risk metrics are discussed in turn, as per previous chapters.

5.2.1. Value at Risk

The IFDBN used to produce 10-day 99% VaR forecasts used forecasts of the closing values of the S&P 500 index generated by the BN employing the SI-HITON-PC learning algorithm and the EGARCH model. The 10-day 99% VaR forecasts achieved no breaches during the period, which matches the experience of the EGARCH model, as shown in Section 3.4.1. This means that, as for the other VaR models evaluated in this study, the IFDBN achieved a 'Green zone' outcome using the BCBS's traffic light test. Moreover, Kupiec's proportion of failures test rejected its null hypothesis for the IFDBN's 10-day 99% VaR forecasts, while Christoffersen's test for independence failed to reject its null hypothesis for the same forecasts.

The performances of the IFDBN with respect to the four forecasting error measures used in this study are summarised in Table 66, below. For ease of comparison, the results of the historical simulation model (as the benchmark model), the EGARCH model, and the BN using the SI-HITON-PC learning algorithm are also provided in Table 66.

Table 66: Forecasting Error Measures for the 10-day 99% Value at Risk Metric using an Integrated Forecast Dynamic Bayesian Network

	MAE	RMSE	MAPE	MdAPE
Historical Simulation	0.0944	0.1003	98.1347%	5.7790%
EGARCH(1,1)	0.0694	0.0776	59.7189%	4.8893%
SI-HITON-PC	0.0944	0.1003	98.1347%	5.7790%
IFDBN	0.0694	0.0776	59.8151%	4.8964%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day value at risk (VaR) forecasts produced in this study at the 99% confidence level over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard and Poor's (S&P) 500 index and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. The exponential generalised autoregressive conditional heteroscedasticity (EGARCH) model used the normal distribution as the underlying statistical distribution. The semi-interleaved HITON parents and children (SI-HITON-PC) algorithm was used as the learning algorithm by a Bayesian network (BN) to produce one-day-ahead forecasts of the closing values of the S&P 500 index. The integrated forecast dynamic Bayesian network (IFDBN) used the SI-HITON-PC learning algorithm and the EGARCH model to produce the 10-day 99% VaR forecasts.

The results obtained for the 10-day 99% VaR IFDBN score equally well to the EGARCH model when using both the MAE and the RMSE measures, ranking these two models as the most accurate 10-day 99% VaR forecasting models. Using the MAPE and the MdAPE measures, the IFDBN ranks as the second-most accurate model, after the EGARCH model, when producing 10-day 99% VaR forecasts. Hence, it is concluded that the incorporation of a forward-looking element to the traditional EGARCH model employed does not yield any statistically significant advantages relative to the strictly backwards-looking EGARCH model when used to produce 10-day 99% VaR forecasts. However, the IFDBN does produce more accurate 10-day 99% VaR forecasts relative to the BN using the SI-HITON-PC algorithm used to produce 10-day 99% VaR forecasts.

5.2.2. Stressed Value at Risk

The 10-day 99% SVaR forecasts were produced using two different IFDBNs due to the similar performances of the genetic algorithm and MMHC learning algorithm, and the former's treatment of NAs (see Section 4.3.2). Each learning algorithm's set of forecasts for the daily closing values of the S&P 500 index was used to produce 10-day 99% SVaR forecasts using the GARCH model as the base autoregressive model. The genetic IFDBN's 10-day 99% SVaR forecasts achieved two breaches over the 7,286-trading-day period, while the MMHC IFDBN's 10-day 99% SVaR forecasts achieved only a single breach over the same period. This compares to two breaches using the GARCH model (see Section 3.4.2). Unsurprisingly, the two IFDBNs' sets of forecasts scored the models a 'Green zone' result using the BCBS's traffic light test, a

rejection of the null hypothesis using Kupiec's proportions of failure test, and a failure to reject the null hypothesis using Christoffersen's test for independence.

Once again, the performances of the IFDBNs with respect to the four forecasting error measures used in this study are summarised in Table 67, below. For ease of comparison, the results of the historical simulation model (as the benchmark model), the GARCH model, and the BNs using the genetic algorithm and the MMHC learning algorithm are also provided in Table 67.

Table 67: Forecasting Error Measures for the 10-day 99% Stressed Value at Risk Metric using an Integrated Forecast Dynamic Bayesian Network

	MAE	**RMSE**	**MAPE**	**MdAPE**
Historical Simulation	0.1482	0.1544	156.9730%	8.2343%
GARCH(1,1)	0.0931	0.0965	97.2833%	5.6640%
MMHC	0.1465	0.1523	154.6608%	8.2242%
Genetic	0.1427	0.1483	151.5520%	8.0387%
IFDBN (MMHC)	0.1197	0.1272	126.9094%	6.3477%
IFDBN (genetic)	0.1197	0.1272	126.9094%	6.3477%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day stressed value at risk (SVaR) forecasts produced in this study at the 99% confidence level over the period 15 March 1991 to 14 February 2020, where the SVaR forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The results are based on the logged returns earned on the Standard and Poor's (S&P) 500 index and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. The generalised autoregressive conditional heteroscedasticity (GARCH) model used the normal distribution as the underlying statistical distribution. The genetic algorithm and the max-min hill-climbing (MMHC) algorithm were used as the learning algorithms by each of the Bayesian networks (BN) to produce one-day-ahead forecasts of the closing values of the S&P 500 index. The integrated forecast dynamic Bayesian networks (IFDBNs) used either the genetic algorithm or the MMHC learning algorithm and the GARCH model to produce the 10-day 99% SVaR forecasts.

Interestingly, the IFDBNs performed equally well across the four forecasting error measures, regardless of which of the two BN learning algorithms was used. Moreover, unlike the results for the IFDBN used to produce 10-day 99% VaR forecasts, the IFDBNs used to produce 10-day 99% SVaR forecasts did not perform equally well as the GARCH model, the latter producing significantly more accurate 10-day 99% SVaR forecasts, regardless of which of the four forecasting error measures was used. Hence, while the GARCH model ranked as the most accurate model when producing 10-day 99% SVaR forecasts, it is worth noting that the IFDBNs ranked as the second most-accurate models, and displayed a significantly improved accuracy relative to the respective BNs using either the genetic algorithm or the MMHC algorithm.

5.2.3. Expected Shortfall

Next, attention is turned to ES. As for the 10-day 99% VaR forecasts IFDBN, the IFDBN used to produce 10-day 97.5% ES forecasts was also constructed using forecasts of the closing values of the S&P 500 index using the SI-HITON-PC learning algorithm and the EGARCH model. Unlike the 10-day 99% VaR forecasts, the 10-day 97.5% ES forecasts achieved eight breaches over the period, a number of breaches that is equal to that achieved using the EGARCH model (see Section 3.4.3). As for the previous three IFDBNs, the ES IFDBN also scored a 'Green zone' on the BCBS's traffic light test. Turning to ES-specific backtests, both the conditional backtest and unconditional backtest could not be performed, due to zero breaches being observed for the corresponding VaR forecasts[37]. The minimally biased backtest, on the other hand, rejected its null hypothesis, leading to the conclusion that the 10-day 97.5% ES forecasts produced by the IFDBN are not the true 10-day 97.5% ES values.

This study's forecasting error measures' values for the 10-day 97.5% ES forecasts produced using the IFDBN constructed using the SI-HITON-PC learning algorithm and the EGARCH model are summarised in Table 68. Once again, for comparison, the corresponding results for the historical simulation model (as the benchmark model), the EGARCH model, and the BN using the SI-HITON-PC learning algorithm are also provided in Table 68.

Table 68: Forecasting Error Measures for the 10-day 97.5% Expected Shortfall Metric using an Integrated Forecast Dynamic Bayesian Network

	MAE	**RMSE**	**MAPE**	**MdAPE**
Historical Simulation	0.0917	0.0987	96.9729%	5.2595%
EGARCH(1,1)	0.0546	0.0669	45.6498%	3.3802%
SI-HITON-PC	0.0917	0.0987	96.9730%	5.2595%
IFDBN	0.0546	0.0669	45.6344%	3.3871%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day expected shortfall (ES) forecasts produced in this study at the 99% confidence level over the period 15 March 1991 to 14 February 2020. The results are based on the logged returns earned on the Standard and Poor's (S&P) 500 index and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. The exponential generalised autoregressive conditional heteroscedasticity (EGARCH) model used the normal distribution as the underlying statistical distribution. The semi-interleaved HITON parents and children (SI-HITON-PC) algorithm was used as the learning algorithm by a Bayesian network (BN) to produce one-day-ahead forecasts of the closing values of the S&P 500 index. The integrated

[37] The corresponding VaR forecasts were produced using the SI-HITON-PC-generated daily forecasts of the closing values of the S&P 500 index and the EGARCH model to enable a like-for-like evaluation.

forecast dynamic Bayesian network (IFDBN) used the SI-HITON-PC learning algorithm and the EGARCH model to produce the 10-day 99% ES forecasts.

The results summarised in Table 68 are intriguing. First, it is clear that the IFDBN produced more accurate 10-day 97.5% ES forecasts relative to the BN that used the SI-HITON-PC algorithm to produce the equivalent forecasts. When compared to the EGARCH model, the IFDBN and the EGARCH model score equally well for the MAE and the RMSE. However, the IFDBN scored a lower score for the MAPE relative to the EGARCH model, suggesting that it produced more accurate 10-day 97.5% ES forecasts, but scored a higher score for the MdAPE relative to the EGARCH model, suggesting that it produced less accurate 10-day 97.5% ES forecasts. Either way, the differences between the scores of the EGARCH model and the IFDBN are insignificant.

5.2.4. Stressed Expected Shortfall

Finally, this section discusses the results of the IFDBN that was used to produce 10-day 97.5% SES forecasts. This IFDBN was constructed using daily forecasts of the closing value of the S&P 500 index using the PC (Stable) learning algorithm and the GARCH model as the underlying autoregressive model. The 10-day 97.5% SES forecasts achieved four breaches over the 7,286-trading-day out-of-sample period. This compares to three breaches observed for the GARCH model (see Section 3.4.4). This IFDBN, too, achieved a 'Green zone' conclusion using the BCBS's traffic light test. Unlike the 10-day 97.5% ES forecasts, the 10-day 97.5% SES forecasts' corresponding VaR forecasts[38] did result in breaches, meaning that all three ES-specific backtests could be carried out. All three backtests rejected their null hypotheses and concluded that the 10-day 97.5% SES forecasts produced by the IFDBN are not the true 10-day 97.5% SES values.

Finally, the performances of the IFDBN with respect to the four forecasting error measures used in this study are summarised in Table 69, below. For ease of comparison, the results of the historical simulation model (as the benchmark model), the GARCH model, and the BN using the PC (Stable) learning algorithm are also provided in Table 69.

[38] The corresponding VaR forecasts were produced using the PC (Stable)-generated daily forecasts of the closing values of the S&P 500 index and the GARCH model to enable a like-for-like evaluation.

Table 69: Forecasting Error Measures for the 10-day 97.5% Stressed Expected Shortfall Metric using an Integrated Forecast Dynamic Bayesian Network

	MAE	**RMSE**	**MAPE**	**MdAPE**
Historical Simulation	0.1700	0.1728	178.3905%	9.9945%
GARCH(1,1)	0.0834	0.0892	88.8212%	4.6616%
PC (Stable)	0.1698	0.1727	178.3133%	9.9685%
IFDBN	0.0806	0.0857	85.8459%	4.5943%

Note: This table reports the results of the four forecasting error measures, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE) for the various 10-day stressed expected shortfall (SES) forecasts produced in this study at the 99% confidence level over the period 15 March 1991 to 14 February 2020, where the SES forecast is calculated over the most severe period preceding the return's date, i.e., over a stressed period. The results are based on the logged returns earned on the Standard and Poor's (S&P) 500 index and the measures are based on the differences between the forecasted values of the models and the actual returns achieved for each trading day. The total number of forecasts for the study period was 7,286 per model. The generalised autoregressive conditional heteroscedasticity (GARCH) model used the normal distribution as the underlying statistical distribution. The Peter and Clark Stable (PC (Stable)) algorithm was used as the learning algorithm by a Bayesian network (BN) to produce one-day-ahead forecasts of the closing values of the S&P 500 index. The integrated forecast dynamic Bayesian network (IFDBN) used the PC (Stable) learning algorithm and the GARCH model to produce the 10-day 99% SES forecasts.

Once again, it is clear that the IFDBN produced 10-day 97.5% SES forecasts that are more accurate than those produced by a BN that uses the PC (Stable) algorithm. Moreover, and unlike the three other market risk metrics considered in this study, the IFDBN produced 10-day 97.5% SES forecasts that are more accurate than any other model considered, including the GARCH model.

5.3. Conclusion

This chapter introduced the novel methodology of IFDBNs, which combines the forward-looking forecasting nature of BNs and the weight-assigning functionality of autoregressive models. This methodology was applied using the best-in-class autoregressive models identified in Chapter 3 and the best-in-class BN learning algorithms identified in Chapter 4 to produce IFDBNs to calculate 10-day 99% VaR forecasts, 10-day 97.5% ES forecasts, and their stressed counterparts.

All IFDBNs obtained breaches in line with their traditional autoregressive counterparts, and backtested similarly to those, too. Hence, as for the conclusions in Chapter 4, the use of IFDBNs does not stray from those of traditional models, and the use of IFDBNs will be as appropriate as the use of traditional models from a regulatory perspective, as all five IFDBNs tested in this chapter obtained a 'Green zone' result using the BCBS's traffic light test.

For each of the four market risk metrics, it was shown that the IFDBN(s) considered produced more accurate forecasts than the corresponding BN that does not employ the

integrated forecast methodology. When compared to the traditional autoregressive models, the results of the IFDBNs were mixed.

The VaR IFDBN scored relatively similarly or marginally worse than the corresponding EGARCH model, depending on which forecasting error measure was used. For ES, the IFDBN scored similarly for some forecasting error measures, marginally better for the MAPE measure, and marginally worse for the MdAPE measure, when compared to the EGARCH model, suggesting no improved accuracy overall. When looking at the stressed market risk metrics, the SVaR IFDBNs scored significantly worse than the GARCH model used to produce 10-day 99% SVaR forecasts across all four forecasting error measures. However, the SES IFDBN was the only IFDBN to outperform its traditional GARCH model counterpart, producing more accurate forecasts regardless of which forecasting error measure is considered.

Studies such as those put forward by Shenoy and Shenoy (2000) and Demirer, Mau, and Shenoy (2006) were early to suggest the potential advantages of applying BNs to quantifying market risk. However, they fail to provide any practical methodology for how this could be done, beyond a theoretical approach to identifying nodes which may be related to stock market returns. In a recent application of BNs to forecasting VaR, Apps (2020) builds on the theoretical work presented by these earlier studies by estimating the directional move of a three-stock portfolio's return to produce VaR forecasts. Apps's study also uses only three variables, namely a liquidity variable, a market variable, and the target variable (the return of the three-stock portfolio). Given the simplistic implementation, this methodology falls short of offering either academics or practitioners a well-defined and practical methodology for the implementation of BNs for measuring risk in practice, especially in the stricter setting of banking regulations. In contrast, this study builds on current industry norms in risk measurement investigated in a previous chapter by developing and implementing the IFDBN methodology detailed in this chapter.

As this chapter and the preceding one show, the perceived advantages offered by BNs to quantify market risk, as suggested by earlier studies, do not necessarily materialise, based on the results of this study. Both traditional BNs, developed in the previous chapter, and the IFDBNs, developed in this chapter, scored similarly to their backwards-looking counterparts, with the SES IFDBN being the exception. This discrepancy between the expectations of early studies and the results of this study may be due to the application of BNs and IFDBNs in the

context of the banking regulatory framework put forward by the BCBS, and an (unscaled) application outside of these confines may be addressed in future research.

6. Conclusion

Banks in the United States (US) and, specifically, those governed by the rules of the Basel Committee on Banking Supervision (BCBS), have applied numerous market risk management techniques to quantify their tail risks. The internal models used by US banks are only subject to a basic backtesting technique that lacks statistical rigour. Nonetheless, these internal models are almost exclusively backwards-looking and historical in nature, with the historical simulation model being a popular choice (Pérignon & Smith, 2010).

This study was constructed in three main parts, each offering novelty. First, an array of traditional backwards-looking models were used to produce 10-day 99% value at risk (VaR) forecasts, 10-day 97.5% expected shortfall (ES) forecasts, and forecasts of their stressed counterparts, for the period 15 March 1991 to 14 February 2020 using the closing values of the Standard & Poor's (S&P) 500 index. The autoregressive models were found to offer forecasts that are the most accurate. The generalised autoregressive conditional heteroscedasticity (GARCH) model was found to be the most accurate model when producing 10-day 99% stressed VaR (SVaR) forecasts and 10-day 97.5% stressed ES (SES) forecasts. For the 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts, the exponential GARCH (EGARCH) model was found to produce the most accurate forecast for each of the measures. These findings are in line with those in the literature, for example, O'Brien and Szerszeń (2017), potentially due to the models' abilities to capture volatility clustering (Angelidis, Benos, & Degiannakis, 2004) relative to the other models used.

These models were tested using both the normal distribution and the skewed Student's t distribution as the underlying statistical distributions. The outperformances of the GARCH models and EGARCH models above were recorded when using the normal distribution as the underlying statistical distribution. The skewed Student's t distribution proved to offer poor forecasts across all four market risk metrics when compared to the performance of the normal distribution, a result that contradicts studies suggesting that distributions with higher skew relative to the normal distribution may produce more accurate results (see, for example, McNeil & Frey, 2000; Wong, Fan, & Zeng, 2012). This difference is explained by the poorer tail fit of the skewed Student's t distribution relative to the normal distribution, as the market risk metrics considered in this study are tail risks, and since the skewed Student's t distribution exhibits fatter tails than those of the normal distribution.

This study contributes to the existing literature as it offers a wide array of traditional backwards-looking models assessed using two distributions and four market risk metrics. The stressed metrics, specifically, have very little coverage in the literature, especially with respect to the performances of different statistical distributions.

Moreover, the calibration of these stressed metrics also required some due care. The BCBS does not dictate how a stressed period is to be determined, but, rather, dictates that the stressed period must be one year in length and that it must include 2007 (Basel Committee on Banking Supervision, 2023). Hence, it is not clear whether this stressed period is to be a year of consecutive trading days, or one made of non-consecutive trading days. This study used a stress period equal in length to the regular calibration period used (to facilitate a like-for-like comparison with non-stressed metrics), where this stressed period was made of the days on which the worst returns were achieved in its historical period, whether consecutive or otherwise, thereby yielding the absolute worst calibration set. This methodology was chosen as it offers the most conservative of stressed periods, and removes the ability to adjust any effect of a stressed period made of consecutive trading days by adjusting its start and end dates. This methodology, too, is a novel contribution to the literature.

The outperformance of the autoregressive models above indicated that the incorporation of some forward-looking element to the forecasts produced may perform better than those models that are strictly backwards-looking, as is the case for the historical simulation model, for example. The autoregressive models offer some forward-looking elements, as these models use past return data to produce return volatility forecasts using some weighting mechanisms. This, together with recent attempts to apply Bayesian networks (BNs) to produce VaR forecasts (see Apps, 2020), form the second part of this study. The application of BNs to produce market risk metrics may yield forecasts that are even more accurate than those obtained using the autoregressive models alone. This is due to BNs' ability to produce one-day-ahead forecasts of the closing values of the S&P 500 index, in this study's context, thereby incorporating a forward-looking element beyond the return volatility aspect. These one-day-ahead forecasts were incorporated into the calibration period to produce a return probability density function (PDF), of which one part was the forecasted return, and the remaining parts being the historical returns, to facilitate the calculation of the various tail market risk metrics.

Four algorithms were implemented to learn the structure of the BN for each of the four market risk metrics considered in this study. When evaluating these algorithms' performances

in producing 10-day 99% VaR forecasts and 10-day 97.5% ES forecasts, the semi-interleaved HITON parents and children (SI-HITON-PC) algorithm was found to produce the most accurate forecasts. When evaluating the algorithms' performances in producing forecasts for the stressed metrics, the genetic algorithm and the max-min hill-climbing (MMHC) algorithm produced the most accurate 10-day 99% SVaR forecasts[39], while the Peter and Clark (PC) (Stable) algorithm produced the most accurate 10-day 97.5% SES forecasts.

This second part's contributions to the literature are three-fold. First, this study offers a comprehensive development of a methodology incorporating BNs in asset pricing, with a specific focus on financial risk management. Early studies, such as those by Shenoy and Shenoy (2000) and Demirer, et al., (2006) suggest the use of BNs in asset pricing. However, many studies stop short of identifying potential nodes in the networks and evaluating the causal relationships. Even recent applications, such as that by Apps (2020), use a simplified methodology to produce VaR forecasts, and only in the context of directionality, as opposed to anything more specific. Second, this study expanded the methodology detailed in earlier studies by applying various network learning algorithms to produce 7,286 exact one-day-ahead forecasts for the closing value of the S&P 500 index for the entirety of the out-of-sample period using a rolling-period methodology. This methodology can now be used in other areas of asset pricing, in general, and financial risk management, in particular. Last, this study contributes to the literature by producing not only 10-day 99% VaR forecasts using the BN methodology, but also 10-day 97.5% ES forecasts, 10-day 99% SVaR forecasts, and 10-day 97.5% SES forecasts, with the stressed metrics using the same stress period methodology described earlier, providing a comprehensive review of the performances of these market risk metrics.

The primary finding of this study's second part is that the one-part contribution of the BN's expected return using a forecasted one-day-ahead closing value of the S&P 500 index, relative to the remainder of the return PDF, made of historical returns, yields little added benefit relative to the historical simulation model. Changes in the weighting of this expected return relative to the remainder of the return PDF or, equivalently, a reduction in the calibration period length

[39] The genetic algorithm produced the most accurate 10-day 99% SVaR forecasts, while the MMHC algorithm was a close second. Due to the treatment of NA values produced by the genetic algorithm (see Section 4.3.2), the MMHC algorithm is also mentioned, as the genetic algorithm's more accurate results may be a function of the intervention discussed.

contributing to the return PDF construction, yielded more significant differences between the performances of the various BN learning algorithms.

This finding suggests that the weighting of the forecasts produced by the BN relative to the remainder of the return PDF may yield more accurate market risk forecasts than the non-weighted application performed. Hence, this study's third and final part involves the development of a novel methodology that incorporates the weighting elements of the autoregressive models identified in the first part and the forward-looking expected return forecasts produced by the BN learning algorithms identified in the second part, to produce superior market risk forecasts. For this novel methodology, the term integrated forecast dynamic Bayesian networks (IFDBNs) is coined.

The IFDBNs' results with respect to the various market risk metrics were mixed. While all IFDBNs developed produced more accurate market risk forecasts than their BN counterparts, this result was not universally true when compared to the autoregressive models. The VaR IFDBN and the ES IFDBN scored similarly to the EGARCH models used to produce their respective forecasts across the forecasting error measures used in this study. The SVaR IFDBN produced significantly less accurate forecasts relative to its GARCH counterpart, while the SES IFDBN produced more accurate forecasts relative to its GARCH counterpart. Hence, the risk practitioner must evaluate the usefulness and improved accuracy of either BNs or IFDBNs to quantify and manage market risk on a case-by-case basis.

7. Limitations and Suggestions for Future Research

The limitations encountered when performing this study can be grouped into two main topics, namely data availability and computing power availability. The study period considered was relatively long, covering 15 March 1991 to 14 February 2020. For all three empirical chapters in this study, at most two calibration periods, each 1,264 days in length (i.e., approximately five working years), were necessary to calibrate either the models or the network learning algorithms used.

Addressing the availability of data first, the use of two calibration periods when training the Bayesian networks (BNs) using the various algorithms in this study meant that economic and financial variable data needed to be available approximately ten working years before the study's start date, i.e., 15 March 1991. This meant that certain variables that were thought to exhibit causal relationships with the closing values of the Standard & Poor's (S&P) 500 index (the target node) lacked data for the combined period (made up of the study period and the two calibration periods) and, further, meant that these variables had to be excluded. Their inclusion, especially if a shorter study period was to be chosen, may have increased the accuracy of the forecasts produced by the network and, therefore, yielded more accurate market risk metric forecasts.

Next, the computing power available to learn the BN structures also limited the applications of various BN learning algorithms. In total, ten network learning algorithms were available. However, running each of the remaining six algorithms not employed in this study would have taken an additional year due to the computing power available. This rendered their application to be impractical from a bank's risk manager's perspective, who is tasked with producing market risk forecasts daily, i.e., before trading activities resume the next day. Hence, a run time exceeding a few hours would not be useful to the bank's risk practitioner. Once again, should a shorter study period be chosen in future research, these algorithms may then be considered once more.

Future research avenues are plentiful and will be discussed here in order of this study's three empirical chapters. First, the traditional models evaluated in this study used the normal distribution and one specification of the skewed Student's t distribution of many available. Those other specifications, together with other distributions (such as extreme value theory distributions, for example), may be considered to evaluate whether their tail fits are better for the data used in this study. Moreover, only a limited selection of autoregressive models were

chosen, with those chosen representing the autoregressive models commonly used in the literature covering financial time series. The inclusion of more autoregressive models may yield improved forecasting of market risk metrics if the models' specifications, perhaps with respect to the weights assigned to the time series data, prove to be more accurate relative to the models used in this study. The last aspect of interest relating to this study's first empirical chapter is the definition applied to the stressed period used. The stressed period used in this study was the collection of the worst returns seen to date, without restricting these to be from consecutive dates, as some studies do. This constraint may prove more practical from the practitioner's point of view, as the use of a consecutive year may be easier to calibrate, store, and retrieve, depending on the data and systems available. This may, by extension, yield different results and conclusions for the stressed market risk metrics used in this study.

Second, and as discussed above, there exist many other BN learning algorithms which could be applied, either when computing power allows for them to be applied on the same dataset or if the study period was to be shortened. These BN learning algorithms may yield even better forecasts of the target variable, resulting in even better results overall. Moreover, all of the BN learning algorithms used in this study produced one-day-ahead forecasts. The use of longer projection periods may also be considered, with or without a weighting methodology applied to each of the forecasts made in this longer projection period. This study showed that changes in the length of the calibration period may yield more significant differences in performance among the various BN learning algorithms used. Hence, it could be of interest to see how these and other BN learning algorithms perform with shorter calibration periods. Moreover, this study used the hill-climb search algorithm as the underlying scoring algorithm for all learning algorithms except the genetic algorithm, and all learning algorithms used the Akaike information criterion as the scoring criterion when ranking the different potential solutions. A future avenue for research would include the use of different greedy scoring algorithms and different scoring criteria. The numbers of parents and children nodes were not limited in this application, and could be limited, especially when using algorithms which take longer to run. Last, this study used BNs to produce 10-day versions of market risk metrics. Future research can use the models and BN learning algorithms used in this study to compare the one-day versions of these.

Finally, this study introduced the integrated factor dynamic Bayesian network (IFDBN) methodology, a novel methodology that combines the weight-assigning properties of autoregressive models and the forward-looking forecasts produced by the BNs. In this study,

the underlying autoregressive models used were only the generalised autoregressive conditional heteroscedasticity (GARCH) model and the exponential GARCH (EGARCH) model, based on the results of earlier chapters. The use of a wider variety of autoregressive models and BN learning algorithms in future research may highlight new combinations which may prove to produce more accurate market risk forecasts than those achieved by the IFDBNs employed in this study.

References

Acerbi, C. & Székely, B., 2014. Backtesting Expected Shortfall. *Risk,* 27(11), pp. 76-81.

Acerbi, C. & Székely, B., 2017. *General Properties of Backtestable Statistics,* s.l.: SSRN.

Acerbi, C. & Tasche, D., 2002. Expected Shortfall: A Natural Coherent Alternative to Value at Risk. *Economic Notes by Banca Monte dei Paschi di Siena SpA,* 31(2), pp. 379-388.

Agarwal, V. & Naik, N. Y., 2004. Risks and Portfolio Decisions involving Hedge Funds. *The Review of Financial Studies,* 17(1), pp. 63-98.

Alexander, C., 2008. *Market Risk Analysis IV: Value-at-Risk Models.* Chichester: John Wiley & Sons, Ltd.

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani S. & Koutsoukos, X. D., 2010. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research,* Volume 11, pp. 171-234.

Aliferis, C. F., Tsamardinos, I. & Statnikov, A., 2003. *A Novel Markov Blanket Algorithm for Optimal Variable Selection.* Bethesda, s.n., pp. 21-25.

Angelidis, T., Benos, A. & Degiannakis, S., 2004. The Use of GARCH Models in VaR Estimation. *Statistical Methodology I,* pp. 105-128.

Apps, E., 2020. *Applying a Bayesian Network to VaR Calculations (Working Paper),* Liverpool: University of Liverpool Management School.

Aquaro, V., Bardoscia, M., Bellotti, R., Consiglio, A., De Carlo, F. & Ferri, G., 2010. A Bayesian Networks Approach to Operational Risk. *Physica A,* Volume 389, p. 1721-1728.

Arrieta-Ibarra, I. & Lobato, I. N., 2015. Testing for Predictability in Financial Returns using Statistical Learning Procedures. *Journal of Time Series Analysis,* 36(5), pp. 672-686.

Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D., 1997. Thinking Coherently. *Risk,* Volume 10, pp. 67-71.

Banz, R. W., 1981. The Relationship between Return and Market Value of Common Stocks. *Journal of Financial Economics,* 9(1), pp. 3-18.

Barber, D., 2012. Machine Learning Concepts. In: *Bayesian Reasoning and Machine Learning.* New York: Cambridge University Press, pp. 305-321.

Basel Committee on Banking Supervision, 1988. *International Convergence of Capital Management and Capital Standards,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 1996. *Amendment to the Capital Accord to Incorporate Market Risks,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2004. *International Covergence of Capital Measurement and Capital Standards: A Revised Framework,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2006. *Basel II: International Convergence of Capital Measaurement and Capital Standards: A Revised Framework,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2009. *Revisions to the Basel II Market Risk Frameowk,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2013. *Fundamental Review of the Trading Book: A Revised Market Risk Framework,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2017. *Basel III: Finalising Post-Crisis Reforms,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2019a. *Internal Models Approach: Backtesting and P&L Attribution Test Requirements,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2019b. *Minimum Capital Requirements for Market Risk,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2022. *Instructions for Basel III Monitoring,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2023. *Calculation of RWA for Market Risk,* Basel: Bank for International Settlements.

Begley, T. A., Purnanandam, A. & Zheng, K., 2017. The Strategic Under-Reporting of Bank Risk. *Review of Financial Studies,* 30(10), pp. 3376-3415.

Berkowitz, J. & O'Brien, J., 2002. How Accurate are Value-at-Risk Models at Commercial Banks?. *The Journal of Finance,* 57(3), pp. 1093-1111.

Berkowitz, J., Christoffersen, P. & Pelletier, D., 2009. Evaluating Value-at-Risk Models with Desk-Level Data. *Management Science,* pp. 1-15.

Bollerslev, T., 2007. *Glossary to ARCH (GARCH),* s.l.: Duke University.

Cano, A., Gómez-Olmedo, M. & Moral, S., 2008. *A Score Based Ranking of the Edges for the PC Algorithm.* Hirtshals, Proceedings of the Fourth European Workshop on Probabilistic Graphical Models (PGM 2008), pp. 41-48.

Cano, R., Sordo, C. & Gutiérrez, J. M., 2004. Applications of Bayesian Networks in Meteorology. *Advances in Bayesian Networks,* pp. 309-327.

Chan, L. S. H., Chu, A. M. Y. & So, M. K. P., 2023. A Moving-Window Bayesian Network Model for Assessing Systemic Risk in Financial Markets. *PLoS ONE,* 18(1), pp. 1-24.

Chang, C.-L., Jiménez-Martín, J.-Á., Maasoumi, E. & McAleer, 2019. Choosing Expected Shortfall over VaR in Basel III using Stochastic Dominance. *International Review of Economics and Finance,* Volume 60, pp. 95-113.

Chang, K. & Tian, Z., 2015. *Market Analysis and Trading Strategies with Bayesian Networks.* Washington, DC, Elsevier, pp. 1922-1929.

Chen, J. M., 2018. On Exactitude in Financial Regulation: Value-at-Risk, Expected Shortfall, and Expectiles. *Risks,* 6(2).

Chen, N.-F., Roll, R. & Ross, S. A., 1986. Economic Forces and the Stock Market. *The Journal of Business,* 59(3), pp. 383-403.

Chen, S.-J. & Jordan, B. D., 1993. Some Empirical tests in the Arbitrage Pricing Theory: Macro Variables vs. Derived Factors. *Journal of Banking and Finance,* 17(1), pp. 65-89.

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

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

Citigroup Inc., 2022. *Investor Relations | SEC Filings.* [Online]
Available at: https://www.citigroup.com/citi/investor/sec.htm
[Accessed 31 August 2022].

Colombo, D. & Maathuis, M. H., 2014. Order-Independent Constraint-Based Causal Structure Learning. *Journal of Machine Learning Research,* Volume 15, pp. 3921-3962.

Correa, E. & Goodacre, R., 2011. A Genetic Algorithm-Bayesian Network Approach for the Analysis of Metabolomics and Spectroscopic Data: Application to the Rapid Identification of Bacillus Spores and Classification of Bacillus Species. *BMC Bioinformatics,* 12(33).

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. & Spiegelhalter, D. J., 1999. *Probabilistic Networks and Expert Systems.* New York: Springer-Verlag.

da Veiga, B., Chan, F. & McAleer, M., 2012. It Pays to Violate: How Effective are the Basel Accord Penalties in Encouraging Risk Management? *Accounting & Finance,* Volume 52, pp. 95-116.

Dagum, P., Galper, A. & Horvitz, E., 1992. *Dynamic Network Models for Forecasting.* Stanford, s.n.

Daníelsson, J., Embrechts, P., Goodhart, C., Keating, C., Muennich, F., Renault, O., & Shin, H. S., 2001. *An Academic Response to Basel II,* London: London School of Economics Financial Markets Group.

Daníelsson, J., 2002. The Emperor has no Clothes: Limits to Risk Modelling. *Journal of Banking & Finance,* Volume 26, pp. 1273-1296.

Daníelsson, J., 2013. *The New Market-Risk Regulations.* [Online]
Available at: https://cepr.org/voxeu/columns/new-market-risk-regulations

Daníelsson, J., Hartmann, P. & de Vries, C. G., 1998. The Cost of Conservatism: Extreme Returns, Value-at-Risk, and the Basle 'Multiplication Factor'. *Risk,* 11(1), pp. 101-103.

Dash, D. & Druzdzel, M. J., 1999. *A Hybrid Anytime Algorithm for the Construction of Causal Models from Sparse Data.* San Francisco, Morgan Kaufmann, pp. 142-149.

Demirer, R., Mau, R. R. & Shenoy, C., 2006. Bayesian Networks: A Decision Tool to Improve Portfolio Risk Analysis. *Journal of Applied Finance,* Volume 16, pp. 106-119.

Diebold, F. X. & Mariano, R. S., 1995. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics,* 13(3), pp. 253-263.

Diebold, F. X., 2015. Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests. *Journal of Business & Economic Statistics,* 33(1).

Du, Z. & Escanciano, J. C., 2017. Backtesting Expected Shortfall: Accounting for Tail Risk. *Management Science,* 63(4), pp. 940-958.

Efron, B. & Tibshirani, R. J., 1993. *An Introduction to the Bootstrap.* New York: Chapman & Hall, Inc..

Engle, R., 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrics,* Volume 50, pp. 897-1007.

Engle, R., 2001. GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. *Journal of Economic Perspectives,* 15(4), pp. 157-168.

Fama, E. F., 1965. The Behavior of Stock-Market Prices. *Journal of Business,* 38(1), pp. 34-105.

Fama, E. F., 1990. Stock Returns, Expected Returns, and Real Activity. *Journal of Finance,* 45(4), pp. 1089-1108.

Friedman, N., Murphy, K. & Russell, S., 2013. *Learning the Structure of Dynamic Probabilistic Networks,* Berkley: arXiv preprint arXiv:1301.7374.

Friedman, N., Nachman, I. & Pe'er, D., 1999. *Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm.* Stockholm, Morgan Kaufmann Publishers Inc, pp. 601-620.

Gastineau, G. L., 1994. Beating the Equity Benchmarks. *Financial Analysts Journal,* 50(4), pp. 6-11.

Giot, P. & Laurent, S., 2003. Value-at-Risk for Long and Short Trading Positions. *Journal of Applied Econometrics,* Volume 18, pp. 641-664.

Giot, P. & Laurent, S., 2004. Modelling Daily Value-at-Risk using Realized Volatility and ARCH Type Models. *Journal of Empirical Finance,* Volume 11, pp. 379-398.

Gneiting, T., 2011. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association,* 106(494), pp. 746-762.

González-Rivera, G., Lee, T.-H. & Yoldas, E., 2007. Optimality of the RiskMetrics VaR Model. *Finance Research Letters,* Volume 4, pp. 137-145.

Grinold, R. C., 1992. Are Benchmark Portfolios Efficient?. *Journal of Portfolio Management,* Volume 19, pp. 14-21.

Grzegorczyk, M., 2010. An Introduction to Gaussian Bayesian Networks. In: Q. Yan, ed. *Systems Biology in Drug Discovery and Development: Methods and Protocols.* Totowa: Humana Press, pp. 121-147.

He, X. D., Kou, S. & Peng, X., 2022. Risk Measures: Robustness, Elicitability, and Backtesting. *Annual Review of Statistics and its Application,* Volume 9, pp. 141-166.

Herring, R. J., 2007. The Rocky Road to Implementation of Basel II in the United States. *Atlantic Economic Journal,* Volume 35, pp. 411-429.

Holland, J. H., 1975. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence.* Michigan: The University of Michigan Press.

Hutchinson, J. M., Lo, A. W. & Poggio, T., 1994. A Nonparametric Approach to Pricing and Hedging Derivative Securities via Learning Networks. *The Journal of Finance,* 49(3), pp. 851-889.

Jackson, P., Maude, D. J. & Perraudin, W., 1997. Bank Capital and Value at Risk. *Journal of Derivatives,* pp. 73-111.

Jain, A. K. & Mao, J., 1996. Artificial Neural Networks: A Tutorial. *IEEE Computer,* Volume 29, pp. 31-44.

Jangmin, O., Lee, J. W., Park, S.-B. & Zhang, B.-T., 2004. *Stock Trading by Modelling Price Trend with Dynamic Bayesian Networks.* Berlin, Heidelberg, Springer, pp. 794-799.

Jensen, F. V., 1996. Bayesian Networks Basics. *AISB Quarterly,* 94(1), pp. 9-22.

Jiménez-Martín, J.-Á., McAleer, M. & Pérez-Amaral, T., 2009. The Ten Commandments for Managing Value at Risk under the Basel II Accord. *Journal of Economic Surveys,* 23(5), pp. 850-855.

Jorion, P., 2001. *Value at Risk: The New Benchmark for Managing Financial Risk.* 2nd ed. New York City: McGraw-Hill.

JP Morgan Chase & Co., 2022. *Investor Relations | SEC Filings & Other Disclosures.* [Online]
Available at: https://jpmorganchaseco.gcs-web.com/
[Accessed 31 August 2022].

Kerkhof, J. & Melenberg, B., 2004. Backtesting for Risk-based Regulatory Capital. *Journal of Banking and Finance,* Volume 28, pp. 1845-1865.

Kondor, I., Pafka, S. & Nagy, G., 2007. Noise Sensitivity of Portfolio Selection under Various Risk measures. *Journal of Banking & Finance,* Volume 31, pp. 1545-1573.

Korb, K. B. & Nicholson, A. E., 2004. *Bayesian Artificial Intelligence.* London: Chapman & Hall/CRC Press UK.

Koski, T. & Noble, J. M., 2009. Learning the Graph Structure. In: D. J. Balding, et al. eds. *Bayesian Networks: An Introduction.* s.l.:John Wiley & Sons, Ltd, pp. 167-195.

Kuester, K., Mittnik, S. & Paolella, M. S., 2006. Value-at-Risk Prediction: A Comparison of Alternative Strategies. *Journal of Financial Econometrics,* 4(1), pp. 53-89.

Kupiec, P., 1995. Techniques for Verifying the Accuracy of Risk Measurement Models. *The Journal of Derivatives,* 3(2), pp. 73-84.

Lambert, P. & Laurent, S., 2016. Modelling Financial Time Series Using GARCH-Type Models and a Skewed Student Density. *Journal of Finance and Economics,* 4(2), pp. 54-62.

Laurens, F., 2012. Basel III and Prudent Risk Management in Banking: Continuing the Cycle of Fixing Past Crises. *Risk Governance & Control: Financial Markets & Institutions,* 2(3), pp. 17-22.

Lauritzen, S. L., 1996. *Graphical Models.* New York: Oxford University Press.

Lawrence, H., 1989. S&P 500 Cash Stock Price Volatilities. *Journal of Finance,* 44(5), pp. 1155-1175.

Lehmann, E. L. & Romano, J. P., 2005. *Testing Statistical Hypotheses.* 3rd ed. New York City: Springer.

Lemmer, J. F., 1996. The Causal Markov Condition, Fact or Artifact? *SIGART Bulletin,* 7(3), pp. 3-16.

Linsmeier, T. J. & Pearson, N. D., 2000. Value at Risk. *Financial Analyst Journal,* 56(2), pp. 47-67.

Liu, F. & Stentoft, L., 2021. Regulatory Capital and Incentives for Risk Model Choice under Basel 3. *Journal of Financial Econometrics,* 19(1), pp. 53-96.

Lockamy, A. I. & McCormack, K., 2012. Modeling Supplier Risks using Bayesian Networks. *Industrial Management & Data Systems,* 112(2), pp. 313-333.

Lucas, A. & Siegmann, A., 2008. The Effect of Shortfall as a Risk Measure for Portfolios with Hedge Funds. *Journal of Business Finance & Accounting,* Volume 35, pp. 200-226.

Lucas, P. J. F., van der Gaag, L. C. & Ameen, A.-H., 2004. Bayesian Networks in Biomedicine and Health-Care. *Artificial Intelligence in Medicine,* 30(3), pp. 201-214.

McAleer, M. & da Veiga, B., 2008. Forecasting Value-at-Risk with a Parsimonious Portfolio Spillover GARCH (PS-GARCH) Model. *Journal of Forecasting,* Volume 27, pp. 1-19.

McMillan, D. G. & Kambourourdis, D., 2009. Are RiskMetrics Forecasts Good Enough? Evidence from 31 Stock Markets. *International Review of Financial Analaysis,* Volume 18, pp. 117-124.

McNeil, A. J. & Frey, R., 2000. Estimation of Tail-related Risk Measures for Heteroscedastic Financial Time Series: An Extreme Value Approach. *Journal of Empirical Finance,* Volume 7, pp. 271-300.

Mitchell, M., 1996. *An Introduction to Genetic Algorithms.* 2nd ed. Massachusetts: Massachusetts Institute of Technology.

Mitchell, T. M., 2006. *The Discipline of Machine Learning.* [Online]
Available at: http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf
[Accessed 26 August 2019].

National Bureau of Economic Research, n.d. *US Business Cycle Expansions and Contractions.* [Online]
Available at: https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions
[Accessed 12 August 2022].

Neil, M., Fenton, N. & Tailor, M., 2005. Using Bayesian Networks to Model Expected and Unexpected Operational Losses. *Risk Analysis,* 25(4), pp. 963-972.

Nelson, D. B., 1991. Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrics,* 59(2), pp. 347-370.

Nieto, M. R. & Ruiz, E., 2016. Frontier in VaR Forecasting and Backtesting. *International Journal of Forecasting,* pp. 475-501.

Nilsson, N. J., 1998. *Artificial Intelligence: A New Synthesis.* San Francisco: Morgan Kaufmann Publishers Inc..

Nunes, M., Gerding, E., McGroarty, F. & Niranjan, M., 2018. Artificial Neural Networks in Fixed Income Markets for Yield Curve Forecasting. *SSRN.*

O'Brien, J. & Szerszeń, P. J., 2017. An Evaluation of Bank Measures for Market Risk Before, During and After the Financial Crisis. *Journal of Banking and Finance,* Volume 80, pp. 215-234.

Olbryś, J., 2009. Forecasting Portfolio Return based on Bayesian Network Model. In: W. Milo, S. G. & W. P., eds. *Financial Markets: Principles of Modelling, Forecasting and Decision-Making.* Łódź: Lodz University Press, pp. 157-171.

Pafka, S. & Knodor, I., 2001. Evaluating the RiskMetrics Methodology in Measuring Volatility and Value-at-Risk in Financial Markets. *Physica A: Statistical Mechanics and its Applications,* 299(1-2), pp. 305-310.

Pearl, J. & Russell, S., 2001. Bayesian Networks. In: M. A. Arbib, ed. *The Handbook of Brain Theory and Neural Networks.* Cambridge: MIT Press, pp. 157-160.

Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Francisco: Morgan Kaufmann Publishers Inc..

Peiró, A., 1999. Skewness in Financial Returns. *Journal of Banking & Finance,* Volume 23, pp. 847-862.

Pérignon, C. & Smith, D. R., 2010. The Level and Quality of Value-at-Risk Disclosed by Commercial Banks. *Journal of Banking & Finance,* Volume 34, pp. 362-377.

Pérignon, C., Deng, Z. Y. & Wang, Z. J., 2008. Do Banks Overestimate their Value-at-Risk? *Journal of Banking & Finance,* Volume 32, pp. 783-794.

Pritsker, M., 2006. The Hidden Dangers of Historical Simulation. *Journal of Banking & Finance,* Volume 30, pp. 561-582.

Ramsey, J., Zhang, J. & Spirtes, P. L., 2006. *Adjacency-Faithfulness and Conservative Causal Inference.* Cambridge, s.n.

Reisman, H., 1992. Variables, Factor Structure, and the Approximate Multibeta Representation. *The Journal of Finance,* 47(4), pp. 1303-1314.

RiskMetrics Group, Inc., 2001. *Return to RiskMetrics: The Evolution of a Standard,* New York: RiskMetrics Group, Inc..

Robinson, R. W., 1977. Counting Unlabeled Acyclic Digraphs. In: C. H. C. Little, ed. *Combinatorial Mathematics V. Lecture Notes in Mathematics, Volume 622.* s.l.:Springer, Berlin, Heidelberg, pp. 28-43.

Shanken, J. & Weinstein, M. I., 2006. Economic Forces and the Sock Market Revisited. *Journal of Empirical Finance,* Volume 13, pp. 129-144.

Shanken, J., 1992. The Current State of the Arbitrage Pricing Theory. *The Journal of Finance,* 47(4), pp. 1569-1574.

Sharma, M., 2012. Evaluation of Basel III Revision of Quantitative Standards for Implementation of Internal Models for Market Risk. *IIMB Management Review,* Volume 24, p. 234-244.

Shenoy, C. & Shenoy, P. P., 2000. Bayesian Network Models of Portfolio Risk and Return. In: Y. S. Abu-Mostafa, B. LeBaron, A. W. Lo & A. S. Weigned, eds. *Computational Finance.* Cambridge: MIT Press, pp. 87-106.

Silverstein, C., Brin, S., Motwani, R. & Ullman, J., 2000. Scalable Techniques for Mining Causal Structures. *Data Mining and Knowledge Discovery,* Volume 4, pp. 163-192.

Spirtes, P. & Glymour, C., 1991. An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review,* 9(1), pp. 62-72.

Spirtes, P., Glymour, C. & Scheines, R., 2000. *Causation, Prediction, and Search.* 2nd ed. New York City: Springer.

Stahl, G., 1997. Three Cheers: Why the Basle Committee's Market Risk Multiplication Factor is Fully Justified. *Risk,* 10(5), pp. 67-69.

Steel, D., 2006. Homogeneity, Selection, and the Faithfulness Condition. *Minds and Machines,* Volume 16, pp. 303-317.

Stephenson, T. A., 2000. *An Introduction to Bayesian Network Theory and Usage,* Martigny: Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP).

Suppes, P. c., 1970. *A Probabilistic Theory of Causality.* Amsterdam: North-Holland.

Suthaharan, S., 2014. Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning. *Performance Evaluation Review,* 41(4), pp. 70-73.

Taylor, J. W., 2020. Forecast Combinations for Value at Risk and Expected Shortfall. *International Journal of Forecasting,* Volume 36, pp. 428-441.

Trabelsi, G., Leray, P., Ben Ayed, M. & Alimi, A. M., 2013. *Dynamic MMHC: A Local Search Algorithm for Dynamic Bayesian Network Structure Learning.* London, s.n., pp. 392-403.

Tsamardinos, I. & Aliferis, C. F., 2003. *Towards Principles Feature Selection: Relevancy, Filters and Wrappers.* Key West, Proceedings of Machine Learning Research, pp. 300-307.

Tsamardinos, I., Brown, L. E. & Aliferis, C. F., 2006. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning,* Volume 65, pp. 31-78.

Tucker, A. & Xiaohui, L., 1999. *Extending Evolutionary Programming Methods to the Learning of Dynamic Bayesian Networks.* Orlando, s.n., pp. 923-929.

Tucker, A., Xiaohui, L. & Ogden-Swift, A., 2001. Evolutionary Learning of Dynamic Probabilistic Models with Large Time Lags. *International Journal of Intelligent Systems,* Volume 16, pp. 621-645.

United States' Securities and Exchange Commission, 1997. *Release No. 7836,* s.l.: s.n.

Verma, T. & Pearl, J., 1990. Causal Networks: Semantics and Expressiveness. In: R. D. Scharter, T. S. Levitt, L. N. Kanal & J. F. Lemmer, eds. *Uncertainty in Artificial Intelligence 4.* New York: Elsevier, pp. 69-76.

Wang, H., Yu, K. & Yao, H., 2006. *Learning Dynamic Bayesian Networks using Evolutionary MCMC.* Guangzhou, s.n., pp. 45-50.

Wang, Q., Wang, R. & Ziegel, J., 2023. *E-backtesting.* [Online]
Available at: https://arxiv.org/pdf/2209.00991v3
[Accessed 2 April 2024].

Wong, W. K., 2008. Backtesting Trading Risk of Commercial Banks Using Expected Shortfall. *Journal of Banking & Finance,* Volume 32, pp. 1404-1415.

Wong, W. K., Fan, G. & Zeng, Y., 2012. Capturing Tail Risks beyond VaR. *Review of Pacific Basin Financial Markets and Policies,* 15(3), pp. 1-25.

Yamai, Y. & Yoshiba, T., 2005. Value at Risk versus Expected Shortfall: A Practical Perspective. *Journal of Banking and Finance,* Volume 29, pp. 997-1015.

Zuo, Y. & Kita, E., 2012a. Up/Down Analysis of Stock Index by using Bayesian Network. *Engineering Management Research,* 1(2), pp. 46-52.

Zou, Y. & Kita, E., 2012b. Stock Price Forecast using Bayesian Network. *Expert Systems with Applications,* Volume 39, pp. 6729-6737.

Appendices

Appendix A: Variables used to Train the Bayesian Networks

Table 70: Variables used to Train the Various Bayesian Networks

Variable Name	Classification
S&P 500 index (closing value)	Target variable
Australian Dollar/US Dollar	Currency Exchange Rate
Bloomberg Commodities Index	Financial
Bloomberg US Treasury Total Return Index	Financial
Canadian Dollar/US Dollar	Currency Exchange Rate
Canola Price	Commodity
Corn Price	Commodity
Euro/US Dollar	Currency Exchange Rate
Federal Reserve Lending Rate	Economic
Financial Times Stock Exchange 100 Index	Financial
Gold Price	Commodity
Great British Pound/US Dollar	Currency Exchange Rate
Japanese Yen/US Dollar	Currency Exchange Rate
Nikkei Index	Financial
Oats Price	Commodity
Russell 1000 Index	Financial
Russell 2000 Index	Financial
Russell 3000 Index	Financial
Silver Price	Commodity
Soybean Meal Price	Commodity
Soybean Oil Price	Commodity
Soybean Price	Commodity

S&P Australian Stock Exchange Index	Financial
S&P Toronto Stock Exchange Index	Financial
Three-month US London Interbank Offer Rate	Financial
Topix Index	Financial
US Consumer Price Index	Economic
US Corporate Aaa-rated Ten-year Spread	Financial
US Corporate Bonds Index	Financial
US Disposable Income Growth Index	Economic
US Full Employment	Economic
US Jobless Claims	Economic
US M1 Money Supply	Economic
US M2 Money Supply	Economic
US Manufacturing Index	Economic
US Manufacturing Tendency Index	Economic
US Non-Farm Payroll	Economic
US Part Time Employment	Economic
US Policy Uncertainty Index	Political
West Texas Intermediate Price	Commodity
Wheat Price	Commodity

Note: This table lists the variables used to train the Bayesian networks (BNs) using the various learning algorithms in this study. The data for the variables were obtained from the Bloomberg database over the period 15 March 1991 to 14 February 2020, as well as for two calibration periods preceding this start date. The data relate to variables from the United States (US) as identified (from the literature or otherwise) to causally relate to the target variable, the Standard & Poor's (S&P) 500 index.