Instituto Superior de Economia e Gestão
UNIVERSIDADE TÉCNICA DE LISBOA
DESDE 1911

MESTRADO EM
ECONOMETRIA APLICADA E PREVISÃO

TRABALHO FINAL DE MESTRADO
DISSERTAÇÃO

EVALUATION OF VOLATILITY MODELS FOR FORECASTING VALUE AT RISK IN STOCK PRICES

RUI ALEXANDRE NARCISO MIGUENS LOURO

OUTUBRO - 2016

Instituto Superior de Economia e Gestão
UNIVERSIDADE TÉCNICA DE LISBOA
DESDE 1911

MESTRADO EM
ECONOMETRIA APLICADA E PREVISÃO

TRABALHO FINAL DE MESTRADO
DISSERTAÇÃO

EVALUATION OF VOLATILITY MODELS FOR FORECASTING VALUE AT RISK IN STOCK PRICES

RUI ALEXANDRE NARCISO MIGUENS LOURO

ORIENTAÇÃO:
PROFESSOR DOUTOR NUNO RICARDO MARTINS SOBREIRA

OUTUBRO – 2016

Abstract

The work depicted in this Thesis pertains to the calculation of Value at Risk and Expected Shortfall, presently the most relevant risk measurements for market risk. To that effect a number of GARCH models are used in conjunction with the Normal, Student-t and Generalized Error Distribution as well as their skewed counterparts. A factorial test plan is carried out through the use of a Rolling Window scheme where three different Rolling window sizes, three forecast horizons and two values of probability of loss are considered. The object of the study was the returns of the firms included the PSI20 stock index, as well as the index itself. The model data was then backtested to ensure that the data have an acceptable behavior, such as unconditional coverage, absence of clustering and proper calculation of the Expected Shortfall.

The results underline the difficulty in obtaining a model that can provide acceptable Value at Risk and Expected Shortfall values. They also show that the model parameters that provide the best results depend on the intended probability of loss and the forecast horizon. Also, skewed distributions generally do not perform any better than their non skewed counterparts.

Keywords: GARCH, Value at Risk, Expected Shortfall, backtesting

JEL Classification: C22, C52, C53, C58, G31

Resumo

O trabalho descrito nesta Tese é referente ao cálculo de *Value at Risk* e *Expected Shortfall* que presentemente são as medidas de risco de mercado com maior relevância. Para tal efeito são utilizados vários modelos GARCH em conjunção com a distribuição Normal, a t de Student e a distribuição Normal Generalizada assim como as suas equivalentes enviesadas. É levado a cabo um plano de ensaios fatorial através do uso de uma abordagem de janela deslizante, onde três tamanhos diferentes de janela deslizante, assim como três horizontes de previsão e dois valores de probabilidade de perda são considerados. O objeto do estudo são os retornos das companhias incluídas no índice PSI20 assim como o próprio índice. Os dados provenientes dos vários modelos foram sujeitos a testes estatísticos destinados a aferir se os dados têm um comportamento aceitável, como cobertura incondicional, ausência de agrupamentos e cálculo adequado do *Expected Shortfall*.

Os resultados obtidos reforçam a dificuldade associada à obtenção de um modelo que consiga fornecer valores aceitáveis de *Value at Risk* e *Expected Shortfall*. Também é evidenciado que os parâmetros dos modelos que fornecem os melhores resultados dependem do horizonte de previsão e da probabilidade de perda pretendidos. Adicionalmente, constata-se que as distribuições enviesadas não apresentam uma performance diferente das distribuições base.

Palavras Chave: GARCH, Value at Risk, Expected Shortfall, backtesting

Classificação JEL: C22, C52, C53, C58, G31

Acknowledgements

I would like to thank my supervisor Professor Nuno Sobreira for his guidance, support, availability and suggestions throughout this whole process.

Thanks are also due to Professor João Nicolau for his in depth analysis of this Thesis and the many useful comments that contributed to its improvement.

I would like to thank Mr. Luís Santos, my boss, for accepting the odd hours at work, as well as the occasional time off, that were necessary to carry out this Thesis.

Finally, I would like to thank my parents, brother, sister in law and nephews for all their support, patience and understanding throughout the time it took to develop this work.

Contents

List of Figures

List of Tables

1 Introduction

The work herein described pertains to the area of market risk. When investments are made in the stock market there is always an underlying risk that, at least, part of that investment is lost due to the unpredictability of market behavior. Given that risky investments have the potential to financially harm investors and firms, the need to measure and quantify market risk is of the utmost importance.

In the past few years, research about this topic has centered its attention in two main risk measures, Value at Risk (VaR) and Expected Shortfall (ES). The former is more established within the risk management community whereas the latter is gaining increased research interest.

Given the role played by stock markets in recent financial crises, regulators are now more demanding in terms of financial firms carrying out more risk than the one that can be covered by capital reserves. The exact amount is calculated with the use of risk measurement tools. This generates a conflict of interests in the calculation of the risk measurement, since an underestimated value will reduce the capital costs associated with the reserves but also trigger a regulatory reaction. An overestimated value will not bring about regulatory issues but will have needless capital costs. Therefore, an accurate calculation of the risk measurement is paramount.

The work depicted in this Thesis uses the class of GARCH models to compute VaR and ES for the firms included in the Portuguese Stock Index (PSI20), as well as for the index itself. These models are estimated with the Normal, Student-t and Generalized Error Distribution as well as their skewed counterparts. A factorial test plan is carried out through the use of a Rolling Window scheme where three different Rolling window sizes, three forecast horizons and two loss probabilities are used.

The models are then subject to an array of statistical tests to ensure their ability to provide acceptable VaR and ES values. This process is known as backtesting and is concerned with a range of phenomena that affect financial time series. The purpose of the tests is to ensure that the models provide values that have a proper unconditional coverage, do not exhibit clustering phenomena and properly calculate the ES.

This Thesis is structured so that firstly the subject of the risk measurement, its implications and calculation are explained. Then the GARCH models used in this Thesis and corresponding statistical distributions are presented. The backtests, their objectives and specificities follow. A small review of literature to explain how other studies approached this subject is then presented. This will hopefully clarify some options taken in the methodology section which comes next. Then the main results are described and finally the conclusions are presented and suggestions for future work are made.

2 Measures of Market Risk

Given the serious effects to the overall economy of extreme movements in stock prices, it is important to develop ways of assessing risk associated with stock market movements, *i.e.*, market risk. This has led to the development of two main measurements of market risk: Value at Risk (VaR) and Expected Shortfall (ES). Studies about the VaR are of substantial relevance as financial Regulators demand regularly that firms report the VaR values pertaining to their market positions. Hence, a substantial amount of research has been conducted about this subject.

ES has been developed in recent years as an alternative to VaR and a way to overcome some of the shortcomings of VaR. However, the amount of research about this matter is very recent and naturally scarcer. Nevertheless, some Regulators also require ES values to be reported even though no requirements concerning the backtesting of this risk measurement are made. The VaR estimation is required to undergo a backtesting procedure, which will be explained further ahead in this Thesis.

The main difference between both risk measurements, and one of the main reasons for the development of ES, is that VaR is not a *coherent measure of risk*, as defined by Artzner et al (1999), in contrast with ES. A *coherent measure of risk* must satisfy the following four axioms:

- Translation Invariance: $\varphi(X + m) = \varphi(X) - m$
- Sub-additivity: $\varphi(X_1 + X_2) \leq \varphi(X_1) + \varphi(X_2)$

- Positive Homogeneity: $\varphi(aX) = a\varphi(X)$
- Monotonicity: $\varphi(X_1) \geq \varphi(X_2)\ if\ X_1 \leq X_2$

$Where\ \varphi\ is\ a\ measurement\ of\ risk, X\ is\ a\ market\ position, m\ and\ a \in \mathbb{R}\ and\ a > 0$

The VaR does not comply with the sub-additivity axiom. The value obtained by adding the VaR pertaining to each of the components of a portfolio can often be larger than the VaR calculated for that portfolio. This would entail that a more diversified portfolio tends to be riskier, which is known to be untrue.

The use of the VaR has been criticized precisely by the fact that is not a *coherent measure of risk*. Some stakeholders feel that only coherent measures of risk should be routinely used. However, given the practice and widespread use of VaR and the lack of well-established alternatives, the VaR will most probably continue to be the most popular market risk measure in the near future.

2.1 Basel Committee Requirements

Among the many rules established by the *Basel Committee on Banking Supervision* there are a few that deal with market risk. Insofar as financial firms are concerned, it is required that capital reserves held by the firm, intended to offset a market loss, are calculated according to the following equation:

$$CR_t = mf_t \times RM_t \tag{1}$$

Where, CR_t is the capital reserve at period t, mf_t a time varying multiplying factor and RM_t the selected risk measurement at period t (VaR or ES for instance)

$$mf_t = 3 + F \tag{2}$$

The rules require that an evaluation of the past performance of the risk measurement is undertaken. This is done through a so called *Traffic Light Approach* where the number of violations, *i.e.*, the number of times that the returns are lower than the risk measurement in the last 250 periods, is taken into consideration. This number is then used to obtain a value for *F*, see equation (2), which can increase the capital reserves. Factor *F* is obtained from Table I.

The idea behind the Stoplight analogy is to make it evident when the manner by which the risk measurements are obtained are acceptable (green), cause concern (yellow) or are unacceptable (red). The scenario where an acceptable model systematically provides values that will cause more than 10 violations per 250 time periods is extremely unlikely. In this situation the Regulator may require the financial institution to change the way by which the risk measurement is calculated. In the meantime, the financial firm is required to increase its capital reserves, which entails added costs. For more information regarding this subject see Basel Committee on Banking Supervision (2013).

The capital reserves imply a capital cost since interests are due on this capital. Therefore, it is in the financial firm's best interest to reduce this capital by selecting a relatively low value for the risk measurement. However, if the risk measurements are so low that there are too many violations the firm will incur in increased capital requirements via this system. As such, an ideal manner of obtaining the risk measurements should not be too risky or too conservative but mimic the actual return distribution.

2.2 Value at Risk (VaR)

The VaR is defined as the minimum amount of money lost with an investment with a given probability, given the past behavior of the asset. Equation (3) below provides the mathematical definition of the VaR:

$$P(\Delta V_{t+h} < VaR|\mathcal{F}_t) = \propto \tag{3}$$

Where, $\mathcal{F}_t$ is the available information set until time t , $\Delta V_{t+h} = V_{t+h} - V_t$, V_t is the value of the asset at time t, h is the forecast horizon and $\propto$ is the probability value.

Calculating VaR is not a straightforward task. There are parametric and non-parametric approaches, the former of which will be the only one considered throughout this Thesis. This approach is based on *log returns*, which are calculated by equation (4).

$$r_{t+h} = ln(V_{t+h}) - ln(V_t) \tag{4}$$

It can be proved, see Nicolau (2012), that under this context the calculation for the VaR can be carried out using equation (5).

$$VaR_{t,h,\alpha} = \left(E[r_{t+h}|\mathcal{F}_t] + q_\alpha^Z\sqrt{VAR[r_{t+h}|\mathcal{F}_t]}\right)V_t \tag{5}$$

Where, q_α^Z is the α quantile of the standardized distribution of returns, E[] is the Expected Value and VAR[] the variance

Additional information regarding VaR, more specifically, demonstrations and other properties can be found in Nicolau (2012) and Tsay (2005).

2.3 Expected Shortfall (ES)

The definition for ES hinges on the definition of VaR. While VaR can be seen as a limit, which if surpassed means that the asset return has an extreme behavior, ES is the value that one expects to lose if such behavior occurs. Given this scenario, the returns are on the tail end of the return distribution. There are few extreme points in a data set, so the tail is often difficult to accurately model. Thus a model that provides adequate VaR values does not necessarily provide adequate ES values. ES can be better understood through equation (6).

$$ES_{h,\alpha} = E[r_{t+h}|r_{t+h} < VaR_{t,h,\alpha}] = \frac{1}{\alpha}\int_0^\alpha VaR_{t,h,x}dx \tag{6}$$

A more detailed explanation of ES can be found in Roccioletti (2015). Taking into account equations (5) and (6) it is possible to obtain equation (7), which provides a straightforward manner to calculate the value of ES.

$$ES_{h,\alpha} = \left(E[r_{t+h}|\mathcal{F}_t] + \frac{\sqrt{VAR[r_{t+h}|\mathcal{F}_t]}}{\alpha}\int_0^\alpha q_x^Z dx \right) V_t \tag{7}$$

As a way to simplify the analysis described in this Thesis, from this point onwards the initial investment V_t will be comprised of one Monetary Unit.

3 Generalized Autoregressive Conditional Heteroscedasticity Modelling

As previously seen, the calculation of both VaR and ES requires estimates for the expected value and the standard deviation of the returns. Given the behavior of financial time series, this is not a simple task. The main difficulty pertains to the fact that the log returns are very hard to predict given past behavior and exhibit conditional heteroscedasticity. Whatever model is used it must account these properties as well as the remaining features of financial time series.

Although there are a number of available models for this purpose, the work carried out in the scope of this Thesis pertains to the class of GARCH models. The General Autoregressive Conditional Heteroscedasticity (GARCH) models are one of the most common ways to accurately obtain the required information. These models are an extension of ARCH models and are defined by equation (8).

$$
\begin{aligned}
r_t &= \mu_t + u_t \\
u_t &= \sigma_t \varepsilon_t \\
\sigma_t^2 &= F\left(u_{t-1}^2, \dots, u_{t-q}^2, \sigma_{t-1}^2, \dots, \sigma_{t-p}^2\right)
\end{aligned}
\tag{8}
$$

Where, μ_t is the conditional mean of the log returns, σ_t is the conditional volatility and ε_t is an i.i.d. random variable with mean 0 and variance 1.

For standard time series, an ARMA model can be used to model the conditional mean. However, given that the log returns often display very low autocorrelation values, dynamic ARMA terms are often not relevant. In regards to the conditional volatility, there are a number of GARCH models that can be used. Four models were used throughout this Thesis: GARCH, IGARCH, EGARCH, GJR-GARCH and RiskMetrics. These models are described further ahead.

Naturally, for any of these models to be useful for the purpose of this Thesis, model parameters must be estimated. These estimates are obtained by using numeric approaches based on the principle of Maximum Likelihood. Six distributions for ε_t were assumed throughout this Thesis: Normal, Student-t, Generalized Error Distribution, Skewed Normal, Skewed Student-t and Skewed Generalized Error Distribution. The models and distributions that were selected possess properties that are compatible with the most well-known characteristics of financial time series.

3.1 Characteristics of Financial Time Series

Throughout the years several statistical studies carried out on financial time series have concluded that many possess the following characteristics:

- Negative Skewness – The distribution of returns tends to be skewed in the negative direction

- High Kurtosis – The Kurtosis of the distribution is always above 3 making it a leptokurtic or “flat” distribution
- Calendar Effects – The volatility can be related to certain calendar dates, such as the ones earmarked for disclosing accounting information and forecasts.
- Low autocorrelation between returns
- Volatility clustering – The extreme values of the returns tend to be grouped at certain points in time.
- Asymmetric Effect – The time intervals that have the greatest volatility also entail the greatest financial losses
- Profitability and volatility co-movements – When two or more financial time series are analyzed it is apparent that when the profitability of one series increases the profitability of the other series also increases, and vice-versa. The same is also true for volatility.

A more detailed explanation of these characteristics, as well as their causes and implications, can be found in Nicolau (2012).

3.2 GARCH Models

3.2.1 GARCH

The GARCH model, which was first described by Bollerslev (1986), is the forefather of the subsequent types of GARCH models. It provides a simple and straightforward way to deal with the problems affecting the ARCH family of models. In this model, the conditional volatility is calculated through equation (9):

$$\sigma_t^2 = \omega + \delta_1 u_{t-1}^2 + \cdots + \delta_q u_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_p \sigma_{t-p}^2 \tag{9}$$

The orders (p and q) of the GARCH model must be selected before fitting the data.

3.2.2 IGARCH

The IGARCH, or integrated GARCH model, first described by Engle and Bollerslev (1986), is similar to the GARCH model and the volatility is calculated through equation (9). It differs from the

GARCH model by requiring that the condition from equation (10) be verified by the fitting algorithm.

$$\sum_{i=i}^{q} \delta_i + \sum_{j=i}^{p} \beta_j = 1 \tag{10}$$

This condition implies that a unit root is present in the process governing u_t^2. However, it can be proved that this series is strictly stationary under the IGARCH specification (Nicolau,2012). This model has a number of useful properties such as persistence in variance however, changes in model structure can often be mistaken for a model of this type.

3.2.3 EGARCH

This model, developed by Nelson (1991), tries to overcome some of the weaknesses of standard GARCH models in dealing with the characteristics of financial time series, namely the fact that they do not take into account the presence of asymmetric effects. The conditional volatility is calculated through equation (11).

$$ln(\sigma_t^2) = \omega + \sum_{i=1}^{q} \big(\delta_i \varepsilon_{t-i} + \gamma_i(|\varepsilon_{t-i}| - E[\varepsilon_{t-i}])\big) + \sum_{j=1}^{p} \beta_j ln\big(\sigma_{t-j}^2\big) \tag{11}$$

3.2.4 GJR-GARCH

This model was developed by Glosten et al (1993). It came about as a method for dealing with the presence of asymmetric effects in financial time series. Since the GARCH and IGARCH model are only concerned with the persistence of σ_t^2 and positive and negative u_t^2 are treated equally, these models are unable to properly deal with the asymmetric effect.

The GJR-GARCH(1,1) model calculates the conditional volatility using equation (12) below.

$$\sigma_t^2 = \omega + \delta_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \gamma W_{t-1} u_{t-1}^2 \tag{12}$$

Where, W_t is equal to 1 if u_t is less than zero and 1 otherwise

The term W_t enables this model to deal separately with positive and negative values of u_t and thus take the asymmetric effect into account.

3.2.5 RiskMetrics

The RiskMetrics method was developed by J.P. Morgan and it is possible to show that it is equivalent IGARCH(1,1), without a constant term. This method is only used in combination with the Normal distribution. The RiskMetrics methodology to obtain σ_t^2 is described by equation (13).

$$\sigma_t^2 = (1-\beta)u_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{13}$$

3.3 Statistical Distributions

In order for a GARCH model, of any type, to be fully identified it must undergo an estimation procedure, where the values of the relevant parameters are obtained. The most common methods for GARCH models rely on the Maximum Likelihood (ML) Principle, since these methods ensure that if the distribution is properly selected, the Fréchet Cramer Rao inferior limit will be attained and thus the ML estimator is the most efficient according to this criterion. For more details on this subject see Nicolau (2012) and Tsay (2005). For this principle to be applied a statistical distribution for ε_t must be selected.

The statistical distributions that were used in the scope of the work detailed in this Thesis are well known and are fully described in several references. A total of six distributions were used: three regular distributions, as well as their skewed counterparts.

The Normal or Gaussian distribution is the most commonly used and well known. This is in no small part due to the implications of the Central Limit Theorem. Therefore, its properties and methods for dealing with this distribution are well known. The probability density function is given by equation (14).

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{14}$$

Although fairly easy to work with, this distribution is neither asymmetric or leptokurtic, which may be seen as a disadvantage given the characteristics of financial time series described in Section 3.1.

The Student-t distribution is more complex to work than the normal distribution, although is also well known for quite some time. This distribution is leptokurtic to varying degrees, depending on the number of degrees of freedom associated with the distribution. This enables the extreme data of a financial time series to be more accurately modelled than with the normal distribution. The probability density function is given by equation (15)

$$f(x|\mu,\nu,\beta)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\beta\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{(x-\mu)^2}{\beta\nu}\right) \tag{15}$$

Where, *Γ()* is the Gamma Function

The Generalized Error Distribution, also known as the Generalized Gaussian Distribution, is an extension of the Normal distribution altered in such a way as to being possible to increase the kurtosis of the distribution, thus generating a leptokurtic distribution. It differs from the Student-t distribution in the fact that the shape of the area around the mean is also changed. The probability density function can be found in equation (16).

$$f(x|\mu,\nu,\beta)=\frac{\beta e^{-\frac{1}{2}\left|\frac{x-\mu}{\nu}\right|^{\beta}}}{2^{1+\beta^{-1}}\nu\Gamma(\beta^{-1})} \tag{16}$$

All of the previously mentioned distributions are symmetric, which, as previously stated, conflicts with the known characteristics of financial time series, more specifically their negative skewness. In order to overcome these limitations skewed version of these distributions were also used. The skewed versions of these distributions are obtained from the base versions of the distribution according to the method described by Ghalanos (2015).

4 Backtesting

Once a model has been developed it is possible to obtain the value of the VaR and ES. However, at this stage, it is not possible to make any judgement about the reliability of these values and, consequently, the adequacy of the model that generated them.

Backtests are a category of statistical tests applied to risk measurements provided by a given model and the actual returns in order to evaluate the suitability of that model to provide acceptable risk measurements. A model must be backtested prior to being used by a financial institution.

There are separate backtests for VaR and ES, even though some backtests for ES do rely on backtests for VaR. The backtests for VaR are focused on the fact that one can see this risk measurement as a limit. Therefore, the number of violations, or the number of times the actual return is lower than the VaR, see equation (17), is the subject of the analysis. In the case of ES, the relationship between the value of ES and the actual return, when a VaR violation occurs, is the focus of the analysis.

$$I_t = \begin{cases} 1 \; if \; r_t < VaR_t \\ 0, otherwise \end{cases} \tag{17}$$

The VaR backtests are focused on two main aspects, unconditional coverage and clustering. Since VaR is a limit, it is expected that the relationship described in the following equation be verified.

$$\alpha \times N = NV, \qquad NV = \sum_{t=1}^{N} I_t \tag{18}$$

Where, N is the number of time periods in the interval where VaR is being backtested and NV the number of violations that exist in the evaluation data set

This is called unconditional coverage. In the event that this relationship is not verified the VaR does not comply with either the firms' objective of minimum capital costs, if there are too few violations, or the regulatory requirements, if there are too many violations.

Clustering occurs in situations where, even though the actual number of existing violations is acceptable they are concentrated at certain periods in time. Given that one of the characteristics of financial time series is the existence of volatility clustering, the occurrence of VaR clustering is a serious possibility. To exemplify the issue of clustering let us imagine the case of an evaluation set where 50 violations are expected and 50 violations occur sequentially. This means that for 50 business days, corresponding to roughly to two and a half months, the financial firm will lose more money than what its capital reserves are designed to withstand. Such a long period can lead to the firm's collapse. In short, the selection of a model with clustering issues can make a firm ill prepared to deal with a crisis scenario and also experience a regulatory backlash.

Backtesting ES is a much more complex task than VaR since ES, unlike VaR, does not possess the property of elicitability, which is an important property for evaluating forecasting performance. This somewhat complex matter is thoroughly addressed by Wimmerstedt (2015) and Roccioletti (2015). Some tests that do not make use of the elicitability property have recently been developed (for example, Acerbi & Szekely,2014).

4.1 Value at Risk

A total of six backtests were used to evaluate the suitability of each model for calculating VaR. Of these tests two are concerned with the unconditional coverage, three are concerned with the clustering phenomenon and one with the combination of both aspects.

4.1.1 Proportion of Failures (POF)

As previously stated, if the model is providing adequate values of VaR then the number of expected violations (αN) should be close to the number of actual violations (NV). Based on this relationship Kupiec (1995) suggested a test based on the statistic of equation (19).

$$POF = -2ln\left(\frac{(1-\alpha)^{N-NV}\alpha^{NV}}{\left(1-\frac{NV}{N}\right)^{N-NV}\left(\frac{NV}{N}\right)^{NV}}\right), \qquad POF \sim \chi^2_{(1)} \tag{19}$$

The null hypothesis is that NV/N=α. If the null hypothesis is not rejected, it is assumed that the unconditional coverage provided by the model is adequate. This test rejects the null hypothesis when there are too many or too few violations, thus being suited to the needs of both financial firms and regulators. However, it does not take into account VaR clustering. Furthermore, it does not possess good finite sample properties (until roughly one year of data, see Niepola, 2009).

4.1.2 Independence of Violations Test (IND)

The independence test suggested by Christofferssen (1998) deals solely with the issue of VaR clustering. The test is predicated on the assumption that if there is no clustering then the probability that in the I_t series there is a transition from zero to one is the same as the probability of a transition from one to one occurring. This entails that the probability of a violation occurring is independent of what occurred in the previous time period. The test uses the following statistic.

$$IND = -2ln\left(\frac{\left(1-\frac{n_{01}+n_{11}}{N-1}\right)^{(n_{00}+n_{10})}\left(\frac{n_{01}+n_{11}}{N-1}\right)^{(n_{01}+n_{11})}}{\left(1-\frac{n_{01}}{N-NV}\right)^{n_{00}}\left(\frac{n_{01}}{N-NV}\right)^{n_{01}}\left(1-\frac{n_{11}}{NV}\right)^{n_{10}}\left(\frac{n_{11}}{NV}\right)^{n_{11}}}\right), \qquad IND \sim \chi^2_{(1)} \tag{20}$$

Where, n_{ij} is the number of transitions from i to j that occur in the evaluation data set.

The null hypothesis states that the probability of a transition from one to zero is identical to the probability of a transition from one to one, hence there is no VaR clustering in the data.

This test, while able to detect the worse forms of the clustering phenomenon, is not able to detect some types of clustering situations. The test statistic is based solely on data from transitions that are one sampling period apart. This means that for clustering situations where violations are closely grouped but are separated by more than two sampling periods the test statistic will not differ from a situation where no clustering is present and therefore the null hypothesis will not be rejected.

4.1.3 Conditional Coverage Test (CC)

Christofferssen (1998) notes that neither of the previous tests is complete on its own. In order for a model to provide acceptable VaR values it must simultaneously verify the unconditional coverage and the independence tests. This test is based on a combination of the POF and IND test statistics as combined in equation (21).

$$CC = POF + IND, \qquad CC \sim \chi^2_{(2)} \tag{21}$$

This test's null hypothesis is that the VaR values provided by the model ensure an adequate unconditional coverage and are not affected by the clustering phenomenon. Given that this test is based on the IND test statistic it also has the same problems in detecting some forms of clustering.

4.1.4 Time Until First Failure Test (TUFF)

This test, described in Kupiec (1995), tests for the unconditional coverage of the VaR. The idea behind the test is to analyze the time interval between the beginning of the evaluation set and the occurrence of the first violation. The test statistic is described in equation (22).

$$TUFF = -2ln\left(\frac{\alpha(1-\alpha)^{\tau-1}}{\frac{1}{\tau}\left(1-\frac{1}{\tau}\right)^{\tau}}\right), \qquad TUFF \sim \chi^2_{(1)} \tag{22}$$

Where, τ is the number of sampling periods between the beginning of the evaluation set and the first violation.

This test has low power for identifying unsuitable VaR models and for this reason it is not often used. However, it can provide better results than the POF test if the evaluation set is small enough to cause problems with that test.

4.1.5 Dynamic Quantiles Test (DQ)

The purpose of this test is to evaluate the occurrence of clustering in the I_t vector. It is described in Engle & Manganelli (2004). The test requires a number of steps, the first of which is the creation of a H_t vector, as per equation (23).

$$H_t = I_t - \alpha \tag{23}$$

This vector is then combined with other information concerning the VaR into an X matrix. This matrix contains lagged values of H_t and values of VaR_t. Engle & Manganelli (2004) are not entirely clear on which information the X matrix should contain as well as the number of lags. However, they present an example where the contemporary values of VaR are used as well as 4 lags of H_t. The columns of the X matrix are described in equation (24). The next step is to carry out a regression of H_t in X, as described in equation (25).

$$X = [H_{t-1}| \cdots |H_{t-k}|VaR_t] \tag{24}$$

$$H_t = \lambda X \tag{25}$$

The implementation of this test was slightly altered in regards to the original description in the sense that a first regression with k=4 is carried out and then the residual vector is analyzed for the presence of autocorrelations by means of a Ljung-Box test. If significant autocorrelations are encountered the number of lags k is increased and the process repeated until k=10, in order to avoid infinite test runs and keep the number of lags at a reasonable number. During the course of the test implementation no situation where the presence of autocorrelation would alter the result of the DQ test was detected.

Once the correlation has been satisfactorily carried out it is possible to calculate the test statistic, which is described in equation (26). The null hypothesis is that no data clustering exists.

$$DQ = \frac{\hat{\lambda}' X' X \hat{\lambda}}{\alpha(1-\alpha)}, \qquad DQ \sim \chi^2_{(k)} \tag{26}$$

Generally, this test leads to the stricter results of all of its counterparts that deal with clustering. However, the lack of precise information in published sources regarding its implementation makes it somewhat complex to work with.

4.1.6 Duration Based Independence Test (DBI)

This test is concerned with the presence of clustering in the data. The idea behind the DBI test is to analyze the duration, elapsed time between one violation and the next, of the various violations in the evaluation set and then ascertain if the durations are relatively random or not. The data is then used to obtain the b parameter of a Weibull distribution, see equation (27).

$$f(D; a, b) = a^b b D^{b-1} e^{-(aD)^b} \tag{27}$$

Where D is the duration between violations.

The Null hypothesis is that b=1, indicating that the short and long durations of the violations are sufficiently evenly matched. If b<1 there will be an excessive number of short durations and if b>1 an excessive number of very long durations, either of which indicates clustering.

Since the test is conducted on the duration between violations and required numerical fitting operations, among others, a sizeable number of violations are required to apply the test. Therefore, a substantial number of data points in the evaluation set are required to provide acceptable results, Ghalanios (2015) mentions at least 1000.

Further details regarding this test can be found in Christofferssen & Pelletier (2004).

4.2 Expected Shortfall

The backtests that deal with ES are concerned with its value versus the value of the actual returns when a VaR violation occurs. They are not concerned with the clustering issue. A total of four tests were used for the work herein described.

4.2.1 McNeil, Frye and Embrechts Test (MFE)

McNeil et al (2000) proposes a test for ES that is akin to a t-test. The first step of the test is to obtain the test statistic through equation (28).

$$MFE = \frac{E[res_t]}{\sqrt{VAR[res_t]}} \tag{28}$$

Where $res_t = ES_{\alpha,t} - r_t$

The null hypothesis is that $E[res_t] = 0$ implying there are no significant deviations between the returns, conditional on being smaller than VaR, and the calculated values of ES. Given the lack of knowledge concerning the statistical distribution that should be used to obtain the critical values in this case, the *bootstrap* method is applied to obtain the p-values. This method tends to work better when a substantial number of violations occur, which is not the case for most situations.

4.2.2 Acerbi and Szekely Tests (AS)

Acerbi & Szekely (2014) propose a set of three tests that are designed specifically to deal with the issue of ES not possessing the property of elicitability. The test statistics do not rely on this property and, as such, are not hindered by this problem. The test statistic pertaining to the first test can be obtained using equation (29) while the statistic pertaining to the second test is presented in equation (30). The third test was not used in the scope of this work owing to the complex issues surrounding its implementation.

$$AS1 = \frac{1}{NV}\sum_{t=1}^{N}\frac{r_t I_t}{-ES_{\alpha,t}} + 1 \tag{29}$$

$$AS2 = \sum_{t=1}^{N}\frac{r_t I_t}{-ES_{\alpha,t}\alpha N} + 1 \tag{30}$$

In both cases the null hypothesis is that ES value is appropriate. These test statistics have unknown distributions and, as such, the only way to obtain critical values and make a decision concerning the rejection or non-rejection of the null hypothesis is to use simulation based methods. In this case a *Monte Carlo* method, which consists of the following steps, is used:

1. Simulate n numbers of returns (sr_t^i, i=1,…,n) at each time period of the evaluation data set. The simulation is carried out using the statistical data obtained in the scope of the of the GARCH modelling to randomly generate return data. By carrying out this procedure one will obtain n vectors of simulated returns. (For the work detailed in this Thesis n=5000 was used).
2. Calculate the AS1 and AS2 statistics pertaining to the n simulated vectors of returns. This requires that the VaR and violation vector be calculated for each of these simulated returns vectors. This will lead to n values of statistic AS1 and n values of statistic AS2.
3. Compare the value of AS1 and AS2 pertaining to the actual returns and the n values AS1 and AS2 pertaining to the simulated data. The p values are obtained from equation (31) for the first test and from equation (32) for the second test.

$$p_{AS1} = \frac{\sum_{i=1}^{n} C\left(AS1\left(sr_t^i\right) < AS1(r_t)\right)}{n} \tag{31}$$

$$p_{AS2} = \frac{\sum_{i=1}^{n} C\left(AS2\left(sr_t^i\right) < AS2(r_t)\right)}{n} \tag{32}$$

Where, $C(A < B) = \begin{cases} 1 \; if \; A < B \\ 0 \; otherwise \end{cases}$

Under certain circumstances the *Monte Carlo* process is not required in the case of the second test. Acerbi & Szekely (2014) report that when the returns are calculated using the Normal and the Student-t distribution the critical values of -0.7 and -1.8, corresponding respectively to α=0,05 and α=0,01, are suitable for the majority of situations. In the work described in this Thesis the *Monte Carlo* method was used in all occasions, since no such consideration was made in the case of the Generalized Error Distribution and the skewed variants of the Normal and Student-t distributions by Acerbi & Szekely (2014).

In the case of statistic AS1 only the deviation of the actual return from the value of ES at time periods corresponding to violations is tested. This means that the number of violations is irrelevant for the test result. As such, the test can return acceptable values in models which do not have an acceptable unconditional coverage.

Test statistic AS2 does take into account the number of expected violations. In a situation where the number of existing violations are greater than the expected violations the test statistic is adversely affected and the likelihood of the null hypothesis being rejected increases. However, in the opposite situation, where the number of existing violations is smaller than the number of expected violations, the likelihood of not rejecting the null increases. Therefore, one should take into account the results pertaining to both tests when making a decision concerning the ability of a model to provide acceptable values of ES.

While these tests are acceptable from a regulatory standpoint they will not reject the null hypothesis when the VaR has been overestimated and the unconditional coverage is not acceptable and therefore are not, on their own, suitable for optimal risk management. Furthermore, these tests require significant computational requirements due to the need for *Monte Carlo* techniques.

4.2.3 Emmer, Krautz and Tasche Test (EKT)

Emmer et al (2015) state that if the VaR pertaining to four values of α are successfully backtested then ES is also successfully backtested. This corresponds to testing the distribution tail coverage at four locations thus ensuring that the tail of the fitted distribution is close to the actual return values. The set of α values that must be tested are described in equation (33).

$$\begin{cases} \alpha_1 = \alpha \\ \alpha_2 = 1 - \big((1-\alpha) * 0.75 + 0.25\big) \\ \alpha_3 = 1 - \big((1-\alpha) * 0.5 + 0.5\big) \\ \alpha_4 = 1 - \big((1-\alpha) * 0.25 + 0.75\big) \end{cases} \tag{33}$$

It is also stated that the upper tail of the observations should be inspected, although only a visual inspection is mentioned. No method for conducting this inspection in an automated manner is presented or suggested. Furthermore, it is stated that for many cases five or more values of α may be required to properly test ES using this method. Again, no method for ascertaining the appropriate number of values of α is either presented or mentioned. In the scope of this Thesis only the four values of α mentioned are used.

This test is highly conservative since it requires that nearly the whole tail of the distribution is properly modelled instead of its average, which the quantity of interest in the ES analysis.

5 Review of Literature

An important step when conducting any type of research is to ascertain what publications concerning the matter at hand exist and what conclusions were obtained, since this can have a strong influence on the analysis. A considerable number of publications were analyzed and the ones that are considered more relevant are described below.

Dendramis et al (2014) evaluated the VaR forecasting capabilities of a number of volatility processes. GARCH models such as the standard GARCH and EGARCH were used as well as Extreme Value Theory and Markov Regime Switching based models. The Normal distribution, the skewed Student-t and the skewed Generalized Error distribution (GED) were considered in the analysis. The data used were stock and bond indexes as well as exchange rates. A Rolling Window of 1800 observations was used and the VaR was obtained for α=0,01 and α=0,05. The VaR values were backtested using the POF and DQ backtests. This study concluded that the skewed Student-t and the skewed GED distributions enable the GARCH and EGARCH models to increase their VaR forecasting accuracy relative to the Normal distribution. On its own the EGARCH model type was not able to describe the skewness and kurtosis properties of the analyzed time series. The authors concluded that the EGARCH and GARCH models can, if combined with skewed fat tail distributions, have comparable results to the Extreme Value Theory approach.

Braione & Scholtes (2016) also evaluate the VaR forecasting capabilities but focus their study on the distributional assumptions. The Normal, Student-t, Multivariate Exponential Power (MEP) distributions are used as well as their skewed counterparts. Multivariate GARCH models are also used. The volatility process associated with the multivariate model type is the Rotated BEKK. In the univariate case the GARCH, NCT-GARCH and NCT-APARCH models are used. The results are evaluated using the POF, TUFF, IND, CC, DQ and DBI tests. The values of α being considered are 0,01 and 0,05. A GARCH(1,1) is used for the conditional volatility equation. A Rolling Window size of 1500 is used. However, the model is only updated at every 20 observations. Only one step forecasts are considered. The data consisted of log returns pertaining to the stocks of various NYSE listed companies. The authors concluded that the skewed Student-t distribution produced the most accurate VaR forecasts. The use of this distribution allows a GARCH model to have a performance similar to a far more complex NCT-APARCH model.

Bams et al (2005) evaluates the ability of various models to provide useful VaR values for exchange rates. A GARCH model type coupled with a student-t distribution is used as well as a PGARCH model coupled with a Stable Paretian Distribution. The GARCH (1,1) was used to model the conditional volatility and an AR(1) model for the conditional mean of the log returns. The values of α being considered are 0,01, 0,05 and 0,10. The forecasting horizons are 5, 10 and 20 days. A Rolling Window of 1700 observations is used. The evaluation is conducted by comparing the number of existing violations with the number of expected violations. The clustering phenomenon is not taken into account. The authors conclude that a GARCH model with a Student-t distribution is able to provide acceptable values of VaR whereas more complex tail models tend to overestimate the VaR values.

Angelidis et al (2004) focus their research on several types of GARCH models for obtaining VaR values. The focus of the work are daily log returns of several stock market indexes. The values for α are 0,01 and 0,05, a one-step forecast and Rolling Window sizes of 500, 1000, 1500 and 2000 are used. The research concentrated on the following model types: GARCH, TARCH and EGARCH. The Normal, Student-t and GED distributions, as well as their skewed counterparts, were used. These GARCH models were used in conjunction with several AR(k) model configurations for the conditional mean part of the model and several (p,q) configurations for the volatility part of the model, where k=0,..,4; p=0,..,2 and q=0,..,2. The authors conclude that the mean specification process does not play a significant role in the VaR calculation. Also, they concluded that leptokurtic distributions, such as the Student-t lead to better results. Furthermore, the analysis concluded that the appropriate size of the Rolling Window depends on the algorithm being used, since some require more data than others.

6 Methodology

6.1 Data description

The work carried out within the scope of this Thesis was applied to the stock prices of the companies that are included in the PSI20 on 22 February 2016, and to the index itself. At that exact date, the PSI20 included the following companies: Altri, BCP, BPI, CTT, EDP, EDP Renováveis, GALP, Impresa, Jeronimo Martins, Mota & Engil, Navigator, NOS, Pharol, REN, Semapa, SONAE and Teixeira Duarte.

Therefore, a substantial amount of time series data from individual Portuguese firms was used for this work which it is believed to be an important contribution to this literature. Given that CTT and Teixeira Duarte had fairly recent Initial Public Offerings the available data pertaining to these companies are not sufficient to pursue our analysis. For this reason, both companies were not included in the final dataset.

The data collected for each company comprised data between 3 January 2000 or the earliest date available. As it is standard in the literature, daily stock prices were converted to daily returns, through the use of equation (4), before any analysis was conducted.

6.2 GARCH model specification

The GARCH model building procedure requires identification of the model for the conditional mean. Generally, the conditional mean is defined as an ARMA model. However, it is well-known that autocorrelations between log-returns typically assume very low values (Nicolau, 2012), which precludes the ARMA model as a good description of the dynamic properties of the log-returns. In fact, most of the publications from the literature review defined the conditional mean to be either zero or a constant term. Angelidis et al (2014) state that the conditional mean specification has nearly no practical relevance for the VaR value calculation. Braione & Scholtes (2016) considered a zero conditional mean for the GARCH model and argued that this assumption has negligible impact on the VaR forecasts.

To confirm the robustness of the results to the zero conditional mean assumption, we also obtained the set of VaR values assuming an unrestricted constant mean for the log-returns. These values were compared with the zero mean VaR values. The difference between the number of realized violations and expected violations was analyzed. Finally, it was concluded that, on average, the zero conditional mean model produced lower differences. This result confirms the argument from Braione & Scholtes (2016) for this dataset. This experiment was conducted using various GARCH models, distributions and all variants that are mentioned in the next section. Therefore, the reported results assumed a zero conditional mean for the log returns.

To completely specify the GARCH model or any extensions considered one also needs to specify the lag orders. According to Nicolau (2012) the GARCH(1,1) specification has been relatively successful in many applications. Furthermore, the vast majority of sources from the literature review also use this model specification. Some examples are Dendramis et al (2014) and Braione & Stokes (2016). Finally, a small number of parameters is desirable because a more parsimonious model is less likely to have problems related with overfitting and the fitting process is much less time consuming and computationally intensive. This is important for large scale simulation experiments such as the one we are conducting. As such, the GARCH(1,1) model for the conditional volatility model was used for the work detailed in this Thesis.

6.3 Simulation parameters

For an effective management of the capital reserves intended to offset any market loss one needs to be aware of future capital needs as soon as possible. To that effect, several forecast horizons of the risk measure being considered are needed. Therefore, the VaR and ES estimates produced under the scope of this work pertain to three forecast horizons: 1, 5 and 10. This corresponds to one day, one week and two weeks' forecasts. To one's knowledge this is the first documented work that conducts a large scale experiment which studies simultaneously VaR and ES backtesting with three different forecast horizons.

A given set of data is required to estimate a given GARCH model and produce the forecasts needed for the computation of the VaR and ES of a certain period. We use a fixed size rolling window scheme to obtain the relevant set of VaR and ES values for backtesting. An important question in such an exercise is the sample size of the rolling window. It is evident that recent time series information is crucial for the model. However, the same cannot be said for information that pertains to the time series behavior in the distant past. For instance, it would be extremely unlikely that the behavior of the returns in the 1920's would have any influence on their present behavior. The size of the rolling window is not an easy choice and no academic research has been found concerning this matter. In fact, the literature review reveals little uniformity concerning this matter. Previous studies quite often adopt rolling window sizes of 1000 to 2000 periods, but values outside this interval may also be found. Hence, in this research for the size of the rolling windows the following values were considered: 1000 (corresponding to roughly 4 years of data), 1500 (6 years) and 2000 (8 years). A notable exception is EDP Renováveis who had a fairly recent IPO. For that reason, only rolling windows of sizes 1000 and 1500 were adopted as not enough data is available for the remaining case.

Another parameter that is relevant to this analysis is the value of α. The values reported by researchers are strongly influenced by the Basel Committee Rules, which have changed throughout the years. Nevertheless, values of 0,05 and 0,01 are mentioned frequently and cover the majority of situations. As such, these are the two values that will be considered in the scope of this work.

We also considered several possible GARCH models for this experiment. The ones that were considered in the analysis were: GARCH, IGARCH, EGARCH, GJR-GARCH and RiskMetrics methodology. All of these models have been previously described in section 3.2 and were the most commonly used in the previous literature.

Finally, the statistical distribution for ε_t must be defined for the estimation step of the GARCH mode. The distributions that were considered in scope of this work were the following: Normal, Student-t, Generalized Error Distribution, Skewed Normal, Skewed Student-t and Skewed Generalized Error Distribution. These distributions have been previously in Section 3.3.

6.4 Simulation

The simulations were carried out using a factorial test plan. This means that every possible variant described in the previous section was entered in this analysis. In summary, the possible choices are the following:

- Forecast horizon: 1, 5 and 10
- Rolling Window size: 1000, 1500 and 2000
- α: 0,01 and 0,05
- GARCH model: GARCH, IGARCH, EGARCH, GJR-GARCH and RiskMetrics
- Distributions: Normal, Student-t, Generalized Error Distribution, skewed Normal, skewed Student-t and skewed Generalized Error Distribution

For each firm previously mentioned, we obtain a set of 450 time series with different VaR and ES estimates for future periods, *i.e.*, after the rolling window. Notice that the RiskMetrics methodology adopts the Normal distribution only and the value of α does not affect the GARCH modelling process. Naturally, all the outputs from these models were then subject to the battery of backtests that were mentioned in Sections 4.1 and 4.2. The standard 5% significance level was adopted for all the statistical tests carried out in this research.

Obtaining the models, corresponding forecasts and obtain the values of the test statistics from all the backtests is a computationally intensive task. All of the programming was done in the *R* environment using the *rugarch* package. Details about this package may be found in Ghalanios (2015), for example. This package was adopted because it is the only package that can accommodate such a wide combination of GARCH models and distributions. This package also contains direct implementations of some of the VaR and ES backtests from the analysis, namely the DBI and MFE tests. All remaining tests were coded and implemented manually from scratch.

As emphasized in the previous section, a rolling window scheme was carried out for the GARCH model estimation stage. This scheme is done in the following way: to obtain an estimate of the VaR at time t the candidate GARCH model is fitted using data interval [t-1-RW,t-1], where RW is the Rolling Window size. When the VaR estimate for t+1 is required the GARCH model is fitted using data interval [t-RW,t] and so on. The name Rolling Window is exactly due to the fact that the fitting interval moves in time.

At each point in time the conditional volatility estimate is saved and used to obtain the VaR and ES coupled with the distribution information. Since a zero conditional mean was assumed for the log-returns, the forecast is zero and does not play any role in the VaR/ES computation. Furthermore, the distribution parameters at each time period are recorded to enable the use of the *Monte Carlo* approach of the AS tests. The set of VaR and ES estimates is dubbed the evaluation set. The implementation of all these steps requires a substantial amount of computing time. In order to reduce the time to an acceptable level the evaluation interval was capped at a maximum of 2000 observations for each model. However due to data availability some companies had an evaluation interval smaller than this limit. In these situations, all of the available data was used. Furthermore, parallel computing was used, which enabled 3 possible cases to run simultaneously. All of these combined enabled the fitting procedure to be obtained in a feasible time frame.

7 Results

Once the simulations were concluded the end result are thousands of possible cases and corresponding p-values to the assortment of backtests. This makes data analysis nearly impossible to carry out by simple visual inspection of the listings. Therefore, the results are analyzed by counting the number of cases where the null hypothesis of the backtests was not rejected. These listings are too extensive to be a part of this Thesis. However, they are available upon request.

As previously mentioned the backtests are not equivalent insofar as their purpose is concerned. They can be divided into three groups: unconditional coverage, clustering and Expected Shortfall. These groups are not independent. The minimum requirement for a model is that it has unconditional coverage, meaning that the number of existing violations is similar to the number of expected violations. Second no clustering phenomena should be present in the violation data. It is expected that a good model fulfills both these requirements. Exceptional models will also provide accurate values for the Expected Shortfall. The backtests are grouped as follows:

- Unconditional coverage: POF and TUFF
- Clustering: IND, CC, DQ and DBI
- Expected Shortfall: MFE, AS1, AS2 and EKT

The data being analyzed contains information pertaining to firm stocks and a stock index. Two data sets are being considered, the firms' data set and the PSI20 data set. The number of cases that comply[1] with the aforementioned requirements are given on a per α and forecast horizon basis in Figure 1 to Figure 22.

In order for the data to be properly analyzed it must be known if the number of successfully[2] backtested cases pertaining to a given parameter is statistically similar to the ones pertaining to another parameter of the same type. For example, if one distribution has 400 successful cases and another distribution has 401 successful cases is the latter's performance better than or equal to the former? Therefore, pairwise t-tests were applied to the average of successfully backtested cases, the results of which can be found from Table X to Table XV. These t-tests are the basis for the performance comparison for the firms' data set. The corresponding information for the PSI20 data set can be found in Table XIII to Table XV. A significance level of 5% is used for all t-tests.

Additionally, one should highlight that not every choice from the simulation (see Section 6.4) has the same number of cases associated to it, the distribution function choice is an exception since the data pertaining to the RiskMetrics methodology is not considered in this case. The total number of cases per requirement and per α value and forecast horizon are as follows:

- Models:
  - GARCH, IGARCH, EGARCH and GJR-GARCH: 264 (firms' data set), 18 (PSI20 data set)
  - Riskmetrics: 44 (firms' data set), 3 (PSI20 data set)

[1] In this context, compliance with requirements means that the null hypothesis of all relevant backtests was not rejected
[2] In this context, successfully backtested means that the null hypothesis of all relevant backtests was not rejected

- Distributions: 176 (firms' data set), 12 (PSI20 data set)
- Rolling window size:
  - 1000 and 1500: 375 (firms' data set), 25 (PSI20 data set)
  - 2000: 350 (firms' data set), 25 (PSI20 data set)

7.1 Backtest Comparison

Given the information that this procedure has provided it is now possible to analyze the performance of the models in terms of backtesting. It can be argued that the results that have more interest are the number of times that two tests pertaining to the same requirement gave conflicting results, one test not rejecting the null hypothesis and the other one rejecting it.

The unconditional coverage requirement is evaluated through the use of two tests. The number of times that they gave conflicting results can be found in Table II. The data shows that with the exception of α=0,05 and a forecast horizon of 1, the POF test rejects many more cases accepted by the TUFF test than the inverse. Nevertheless, the number of cases accepted by the POF test and rejected by the TUFF test is still substantial.

In the case of the clustering backtests the data can be found in Table III. The DQ and DBI tests reject many more cases that were approved by the IND and CC tests than the inverse, which means that complex clustering phenomena that cannot be addressed by the IND test are fairly common in the data. Generally speaking, the cases that are accepted by DQ are not rejected by its counterparts, seeing as the number of rejections is very low. However, the DQ test does reject a large number of cases that were accepted by other tests. This behavior is more apparent in the case of the DBI test. Given the subjectivity surrounding the implementation of the DQ test it might be possible to improve the test so as to make the remaining tests redundant.

The results pertaining to the Expected Shortfall backtests can be found in Table IV. It is apparent that the cases accepted by AS1 are rarely rejected by other tests, while it does reject many cases that were accepted by other tests. It should be pointed out that in this context, where the ES backtests are only applied after the unconditional coverage backtests, the AS2 test is not expected to reject many more cases than AS1. The EKT test accepts many casess that are then rejected by the AS1 test, which probably means that the four points used to implement the test are not sufficient to accurately model the tail of the distribution. The MFE test does not reject a single case that has been accepted by AS1, however a large number of cases accepted by the MFE test are rejected by its counterparts. Therefore, the MFE test is redundant in regards to AS1.

7.2 Unconditional coverage

The tests that evaluate the unconditional coverage are the POF and TUFF tests. The breakdown of the number of cases where the null hypothesis was not rejected for both tests can be found in Figure 1 to Figure 6 for the firms' data set and from Figure 18 to Figure 20 for the PSI20 data set.

For α=0,01 all cases pertaining to the PSI20 data set were rejected by the backtests. After close inspection of the data it was apparent that all of the cases overestimated the value of VaR causing the number of realized violations to be much larger than the expected violations. This occurs for every forecast horizon.

In the case of the firms' data set not all companies have successfully backtested cases. Financial and Construction firms, as well as firms involved in financial scandals (Pharol, formerly known as PortugalTelecom) had very few successfully backtested cases, if any.

The model that leads to the most cases being successfully backtested depends on the forecast horizon. If the horizon is 1 period the EGARCH model type is clearly superior to the alternatives, even though IGARCH also displays a good performance. For a forecast horizon of 5 periods, the best results are obtained with the IGARCH model type. RiskMetrics has clear inferior results while the performance of the remaining models is not statistically different from each other. In the case of a forecast horizon of 10 periods, the performance of the various model types is not statistically significantly different, with the exception of RiskMetrics, which has the worst performance.

In terms of the choice of the distribution function there are also some differences between forecast horizons. In the case of the 1-period forecast horizon, the Student-t and the Generalized Error Distribution, as well as their skewed counterparts, had statistically similar performances, which were better than the remaining distributions. For the 5-period forecast horizon the results are similar with the exception that the skewed Generalized Error Distribution now has a worse performance, which is similar to the Normal and skewed Normal distributions. In the case of the 10-period forecast horizon the Student-t and skewed Student-t show the best performance, but is not statistically different from the performance of the Generalized Error distribution.

The Rolling window size does not produce a statistically different performance in any forecast horizon.

The distribution of accepted cases for every parameter, except the firms, can be found in Table V. By analyzing this data no major differences from what was previously stated, such an affinity of one model for a specific distribution or Rolling window size, was found.

When α=0,05 a larger number of cases are now deemed acceptable by the backtests, which was entirely expected. The models are also affected by the forecast horizon, but less so than for α=0,01. At a forecast horizon of 1 the best performance is achieved by GJR-GARCH and IGARCH, their performance is nonetheless statistically similar to all remaining models, except for RiskMetrics. If the forecast horizon is 5 periods, the IGARCH model has the best performance but it is not statistically different from GARCH and GJR-GARCH. For a forecast horizon of 10 periods, EGARCH provides the best results although they are not statistically different from RiskMetrics and IGARCH.

The distributions that provide a larger number of cases that are deemed acceptable by the backtests also varies with the forecast horizon. In the 1-period forecast horizon the Normal and skewed Normal distributions have the best performance, and is statistically different from its counterparts. When the forecasting horizon is 5 periods the Normal and skewed Normal distributions also have the best performance but are not statistically different from the Generalized Error Distribution and skewed Normal distributions. In the case of a forecast horizon of 10 periods, the Normal and the Generalized Error distribution, as well as their skewed counterparts have the best performance, which are not statistically significantly different from one another.

Insofar as the rolling window size is concerned there is also an influence from the forecast horizon. For the forecast horizon of 1 period the best performance is attained by the rolling window size of 1500, although not statistically different from the size 2000. For a 5-period forecast horizon the best performance was achieved by the rolling window size 1500. In the case of a forecast horizon of 10 periods the best performance is provided by the 1500 rolling window size, which is not statistically different from the 2000 rolling window size.

The analysis of Table V does not show any combinations of parameters that do not reflect previous statements.

In the case of the PSI20 data set there were no models whose performance is statistically significantly different from each other or zero for a forecast horizon of 1 period. In the forecast horizon of 5 periods, IGARCH clearly provides the best performance as well as in the forecast horizon of 10 periods. No statistically different performance was present in the case of the statistical distributions in any forecast horizon. The rolling window size of 2000 clearly leads to the best performance in the 5-period and 10-period forecast horizons. The results distribution can be found in Table VIII. Analysis of this data does not show any combinations of interest.

In all of the forecast horizons and α values combinations, the performances of the skewed statistical distributions are not statistically different from their non-skewed counterparts.

7.3 Clustering

When the clustering backtests, IND, CC, DQ and DBI, are added, the number of accepted cases falls quite sharply. This was to be expected given the presence of volatility clustering in the data. The most affected companies are the ones already mentioned in the unconditional coverage results. Roughly half of the companies do not have cases that have acceptable backtest results. The results can be found from Figure 7 to Figure 12 for the firms' data set and from Figure 21 to Figure 22 for the PSI20 data set.

In the case of the firms' data set, and for α=0,01 the best performing model also depends on the forecasting horizon. When it is 1-period the best model is EGARCH although it is not statistically different from IGARCH. In the case of a 5 periods forecast horizon IGARCH provides the best results. For a forecast horizon of 10 periods the models performance is not statistically different from one another, with the exception of IGARCH and RiskMetrics being statistically different between themselves.

The Student-t distribution and skewed Student-t distribution have the best performance for the forecast horizon of 5 periods although they are not statistically different from the Normal and skewed Normal distributions. In the remaining forecast horizons, the performance of the various distributions is not statistically different. Even so the Student-t and skewed Student-t have the best performance.

The rolling window size of 1500 has the best performance for the forecast horizon of 1 period, while for the forecast horizon of 10 periods the 2000 rolling window size provides the best results. For the forecast horizon of 5 periods there are no statistically different results.

For α=0,05 the model performance also depends of the forecasting horizon. For the 1-period case, the best performance is attained by the EGARCH model, although it is not statistically different from the GJR-GARCH model. For all other forecast horizons, the performance of the various models is not statistically different. Insofar as the distributions are concerned the Normal and skewed Normal have the best performances for 1-period forecast horizon, although they are not statistically different from the Student-t and skewed Student-t distributions. For the remaining forecast horizons no statistically different results are present. The rolling window size of 2000 has the best performance in all forecast horizons, although in 10-period forecast horizon it is not significantly different from a rolling window of size 1500.

The distribution between parameters for this case is present in Table VI. The analysis of this table does not provide any information that can add to the statements already made. No combination of model, distribution or rolling window size that provides better results has been detected.

In the case of the PSI20 data set there are too few data to provide any conclusion. However, the few cases that were successful had a rolling window size of 2000. For the forecast horizon of 1 period no cases were successfully backtested. The results for this case can be found in Table IX.

Similarly, to the unconditional coverage requirement, for the clustering requirement, for all of the forecast horizon and α values combinations, the performances of the skewed statistical distributions are not statistically different from their non-skewed counterparts.

7.4 Expected Shortfall

When the full set of backtests is applied there are so few cases where the null hypothesis of at least one of the backtests is not rejected that the results do not have much relevance, except as to reinforce the difficulty of obtaining a model capable of accurately depicting a financial time series. The results can be found from Figure 13 to Figure 17 for the firms' data set.

In all of the cases for the firms' data set, except for α=0,05 and a forecast horizon of 10 periods, they are not even statistically different from zero. For α=0,01 and a forecast interval of 1 period there are no cases that were successfully backtested. Furthermore, the few cases that are successfully backtested pertain to only three firms making the results extremely dependent on any particularity contained in the data. Even though there is little data it is apparent that the models that provided acceptable models were either the IGARCH and GJR-GARCH and the distributions the Student-t and the skewed Student-t.

For the case of α=0,05 and a forecast horizon of 10 periods there are significant numbers of acceptable cases. However, the performance of the models and distributions is not significantly different between components, except for the RiskMetrics model. The rolling window size of 1500 clearly provides the best performance. It should be pointed out that all of the data for these conditions pertains to only one company.

In the PSI20 data set there were no cases that were successfully backtested for the Expected Shortfall requirements.

8 Conclusions and Future Work

The work herein depicted provides a good insight into the difficulties behind the task of accurately modelling risk measurements. A model that will provide appropriate VaR and ES values is a very complex objective to achieve.

The combination of the type of model, statistical distribution and Rolling window size that provides the better results greatly depends on the intended value of α and forecast horizon, as well as to the data. Generally speaking, the IGARCH, EGARCH and GJR-GARCH provide the better results. The EGARCH model type seems to provide better results for a forecasting interval of 1 period and the IGARCH model seems to provide better results when the forecasting interval is 5 periods.

The differences of the performances of the various distributions depend greatly on the intended value of α and forecast horizon. However, for the most part, the skewed distributions did not have a significantly different performance compared to their non-skewed counterparts. This also occurs with the Rolling Window size, although a larger rolling window size seems to be preferable.

It was much harder to obtain a model that provides acceptable VaR values for the PSI20 index than for the set of companies that are included in the index.

This work can be extended to include methods based on Extreme Value Theory and Markov Regime Switching strategies. The use of these strategies could lead to better results, especially the use of Extreme Value Theory.

The GARCH parameters of the models on which the analysis is based were obtained via maximum likelihood based algorithms, which select the parameters based on the whole distribution and not only the tail which is the area of focus for a Risk Measurement. Therefore, a strategy of obtaining GARCH parameters that optimize VaR (see, for example, Ranković et al, 2016), has the potential to provide better results.

Finally, given the results concerning the backtests it would be worth improving the poorly understood implementation of the DQ test so that it could make the IND, CC and DBI tests redundant. This would make future research on the subject of risk measurements easier to carry out.

References

Acerbi, C. & Szekely, B. (2014). *Backtesting Expected Shortfall*. Available at: https://www.msci.com/documents/10199/22aa9922-f874-4060-b77a-0f0e267a489b [Access in: 2016/06/06].

Angelidis, T., Benos, A. & Degiannakis, S. (2004). The use of GARCH models in VaR estimation. *Statistical Methodology* 1, 105-128.

Artzner, P., Delbaen, F., Eber, J. & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance* 9 (3), 203-228.

Bams, D., Lehnert, T. & Wolff, C. (2005). An evaluation framework for alternative VaR-models. *Journal of International Money and Finance* 24, 944-958.

Basel Committee on Banking Supervision (2013). *Fundamental review of the trading book: A revised market risk framework.* Available at: http://www.bis.org/publ/bcbs265.pdf [Access in: 2016/09/01].

Braione, M., & Scholtes, N. (2016). Forecasting value-at-risk under different distributional assumptions. *Econometrics* 4 (1), 3.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* 31 (3), 307-327.

Christofferssen, P. (1998). Evaluating Interval Forecasts. *International Economics Review* 39 (4), 841-862.

Christofferssen, P. & Pelletier, P. (2004). Backtesting value at risk: A duration based approach. *Journal of Financial Econometris* 2 (1), 84-108.

Dendramis, Y., Spungin, G. & Tzavalis, E. (2014). Forecasting VaR models under Different Volatility Processes and Distributions of Return Innovations. *Journal of Forecasting* 33, 515-531.

Emmer, S., Kratz, M. & Tasche, D. (2015). What is the best risk measure in practice? A comparison of standard measures. *Journal of Risk* 18 (2), 31-60.

Engle, R. & Manganelli, S. (2004). Caviar: Conditional autoregressive value-at-risk by regression quantiles. *Journal of Business and Economic Statistics* 22, 367-381.

Ghalanos, A. (2015). *Introduction to the rugarch package (Version 1.3-1).* Available at: https://cran.r-project.org/web/packages/rugarch/vignettes/Introduction_to_the_rugarch_package.pdf [Access in: 2016/03/02].

Glosten, L., Jagannathan, R. & Runkle, D. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance* 48 (5), 1779-1801.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives* 3, 73-84.

McNeil, A., Frey, R. & Embrechts, P. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance* 7, 271–300.

Nelson, D. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* 59, 347-370.

Nicolau, J. (2012). *Modelação de Séries Temporais Financeiras*, 1st Ed. Lisbon: Almedina.

Niepola, O. (2009). *Backtesting Value at Risk Models.* Masters Thesis in Economics, Helsinki School of economics, Available at: http://epub.lib.aalto.fi/fi/ethesis/pdf/12049/hse_ethesis_12049.pdf [Access in: 2016/04/23].

Ranković, V., Drenovak, M., Urosevic, B. & Jelic, R. (2016). Mean-univariate GARCH VaR portfolio optimization: Actual portfolio approach. *Computers & Operations Research* 72, 83–92.

Roccioletti, S. (2015). *Backtesting Value at Risk and Expected Shortfall*, 1st Ed. Wiesbaden: Springer Gabler.

Tsay, R. (2005). *Analysis of Financial Time Series*, 2nd Ed. Hoboken: J. Wiley & Sons.

Wimmerstedt, L. (2015). *Backtesting Expected Shortfall: the design and implementation of different backtests.* Master's Thesis in Mathematical Statistics, Swedish Royal Institute of Technology, School of Engineering Sciences, Available at: http://www.math.kth.se/matstat/seminarier/reports/M-exjobb15/150826a.pdf [Access in: 2016/06/02].

ANNEX A – Figures

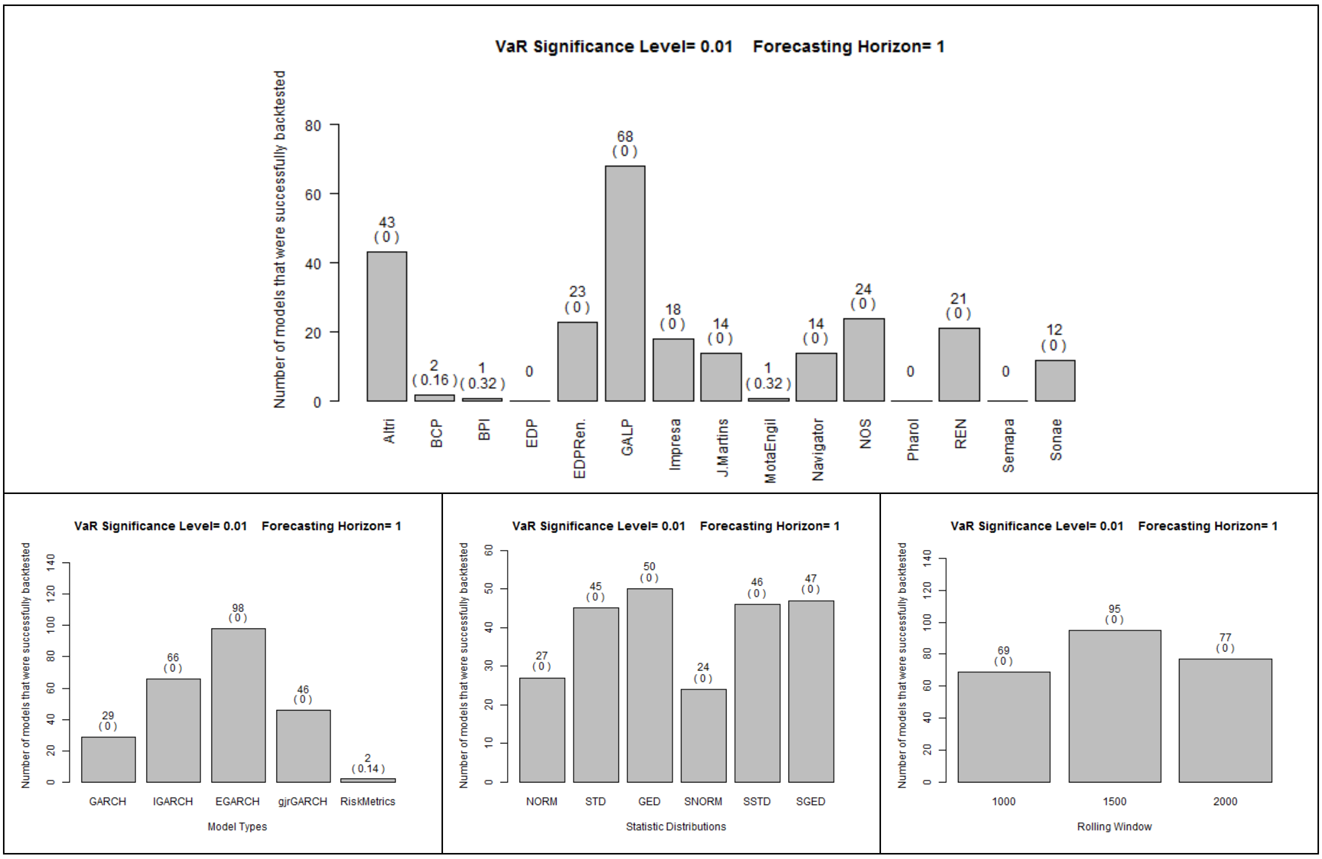

Figure 1 - Number of cases where the null hypothesis of the POF and TUFF tests was not rejected for the firms' data set. Forecast horizon=1 and α=0,01. (t-test p value in parenthesis)

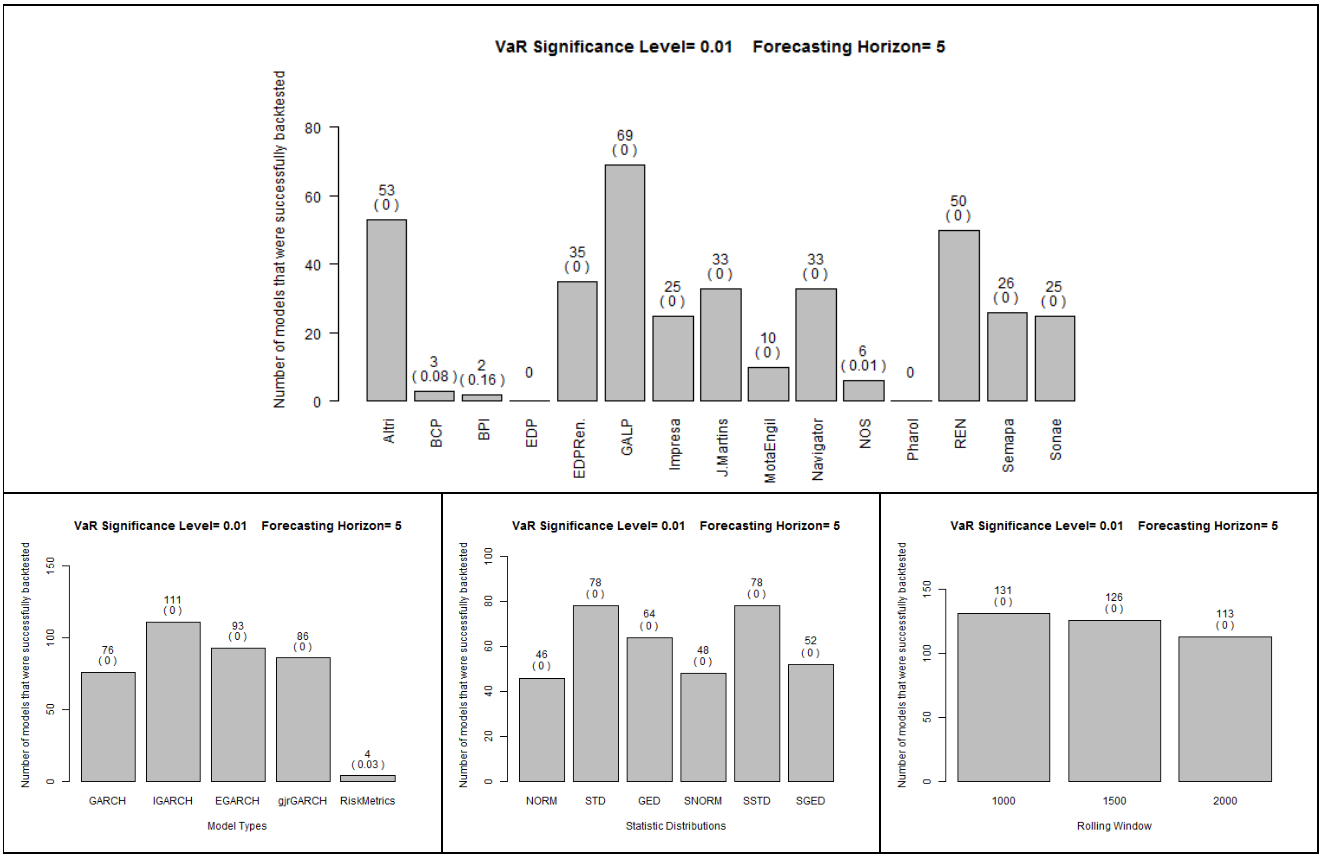

Figure 2 - Number of cases where the null hypothesis of the POF and TUFF tests was not rejected for the firms' data set. Forecast horizon=5 and α=0,01. (t-test p value in parenthesis)

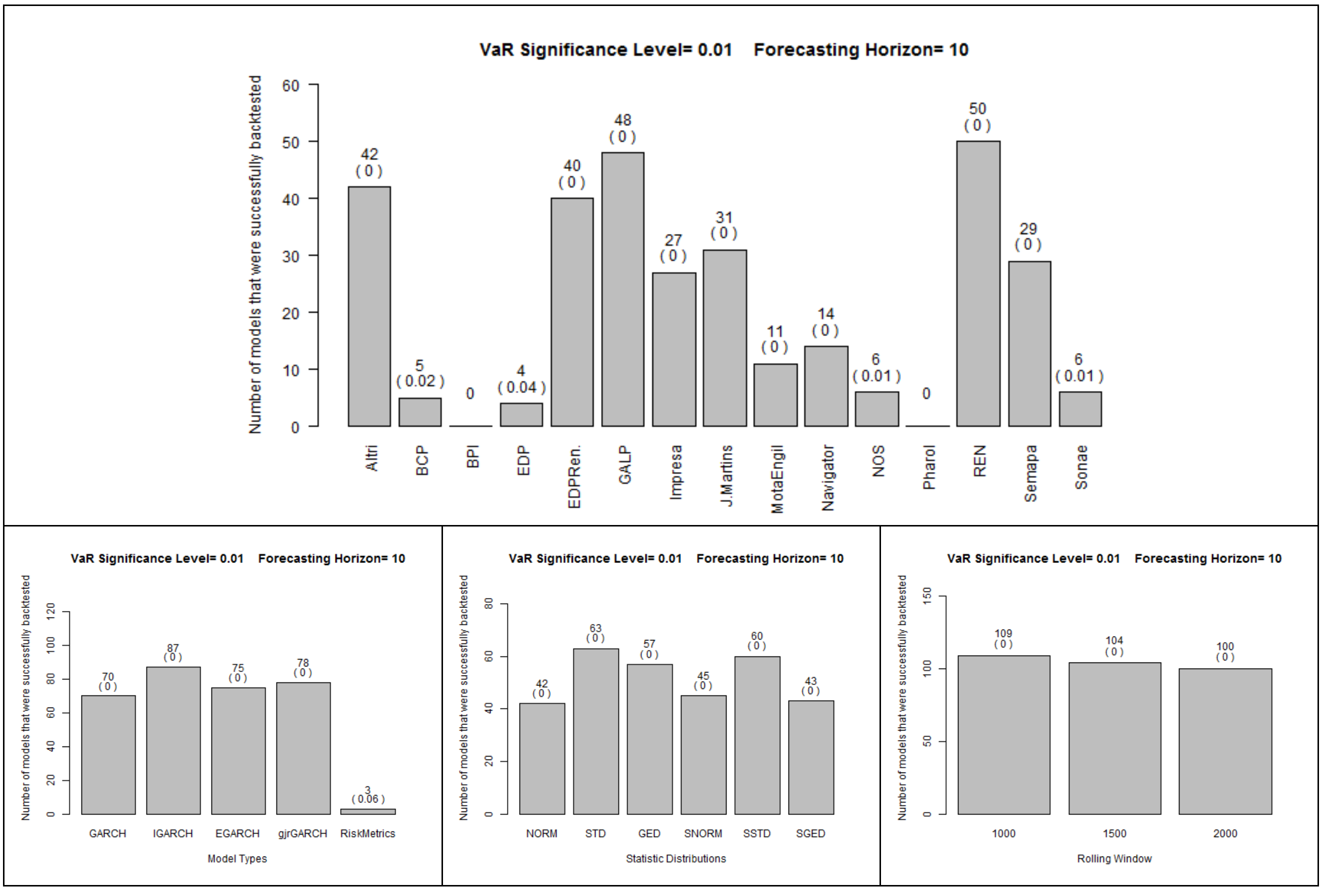

Figure 3 - Number of cases where the null hypothesis of the POF and TUFF tests was not rejected for the firms' data set. Forecast horizon=10 and α=0,01. (t-test p value in parenthesis)

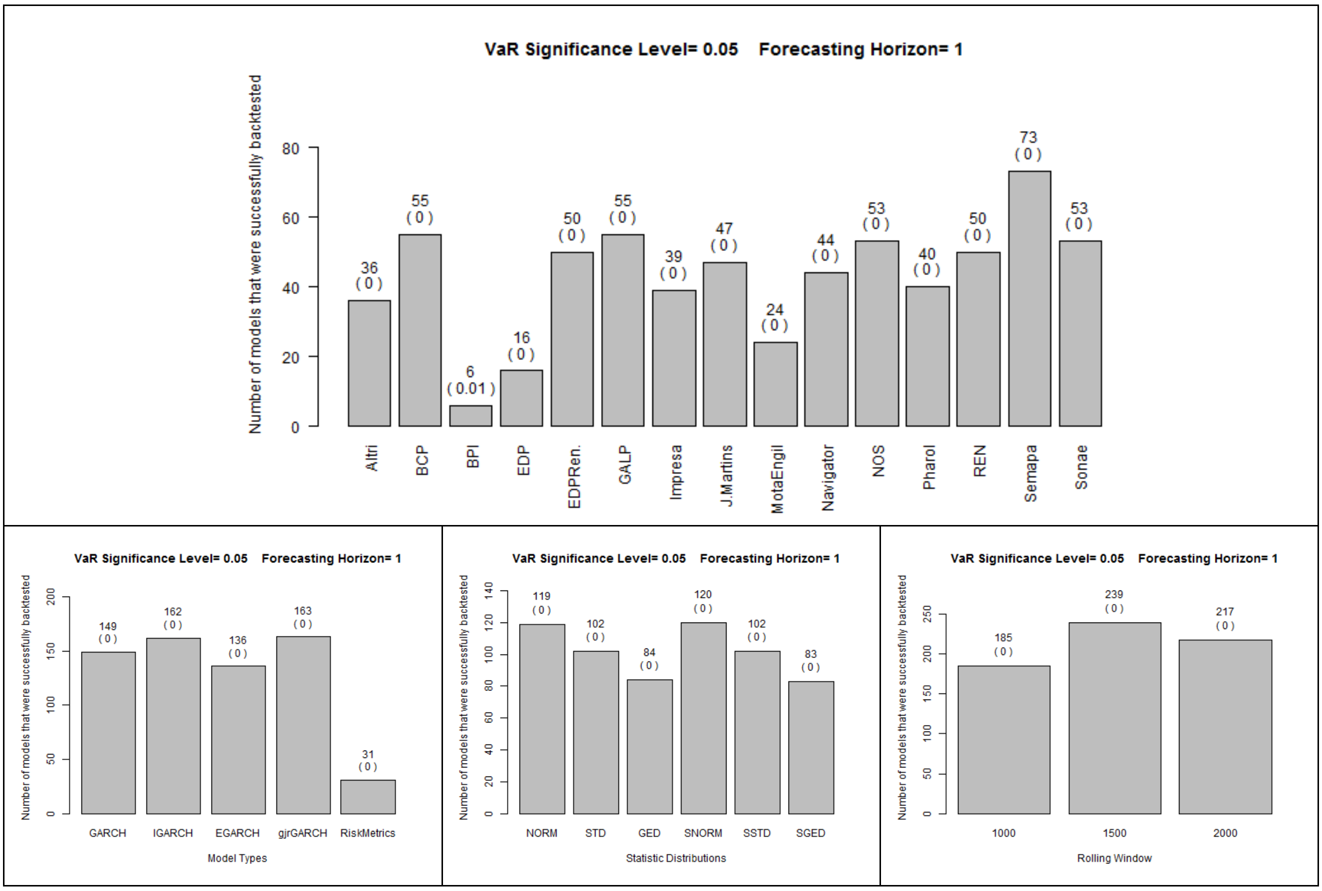

Figure 4 - Number of cases where the null hypothesis of the POF and TUFF tests was not rejected for the firms' data set. Forecast horizon=1 and α=0,05. (t-test p value in parenthesis)

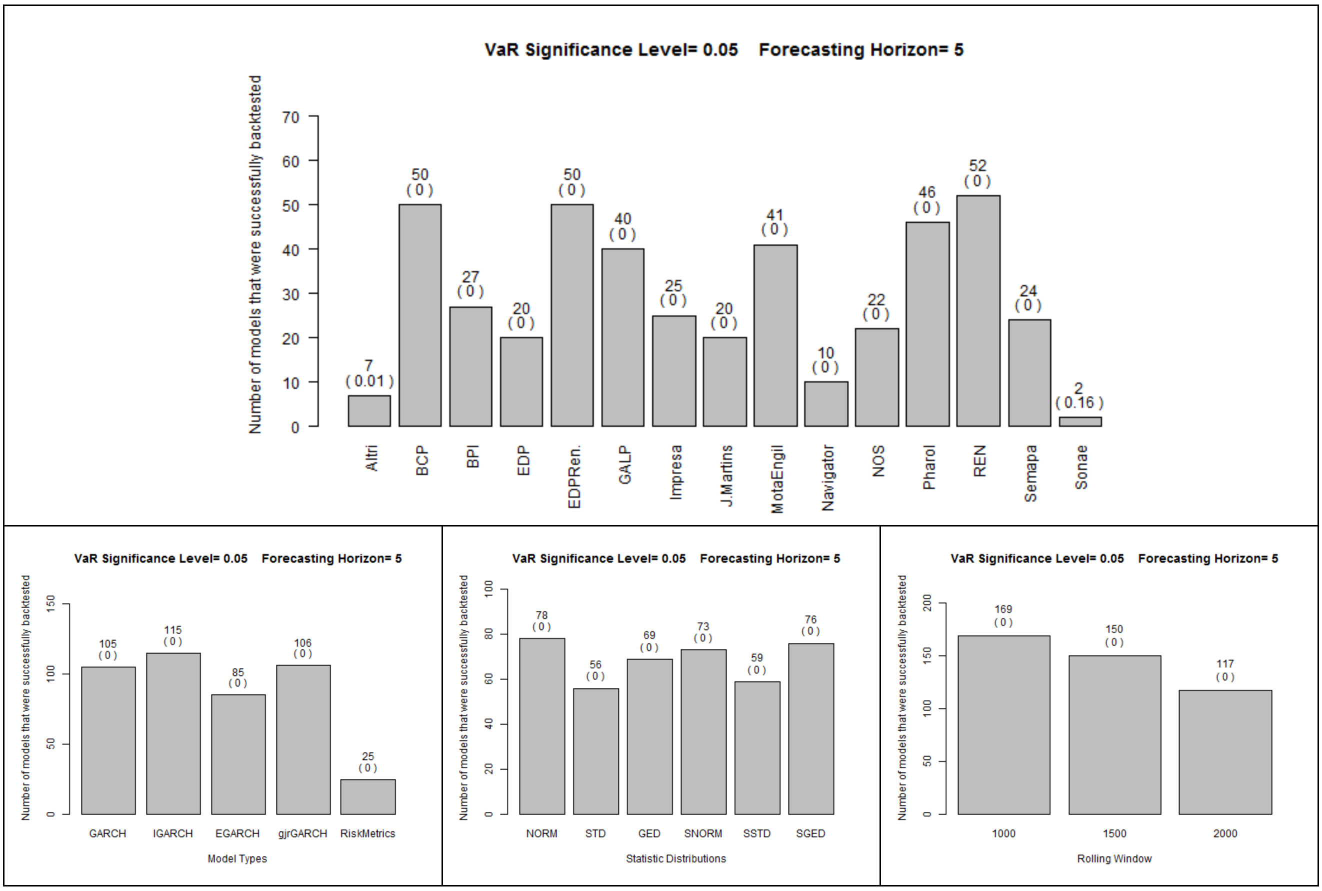

Figure 5 - Number of cases where the null hypothesis of the POF and TUFF tests was not rejected for the firms' data set. Forecast horizon=5 and α=0,05. (t-test p value in parenthesis)

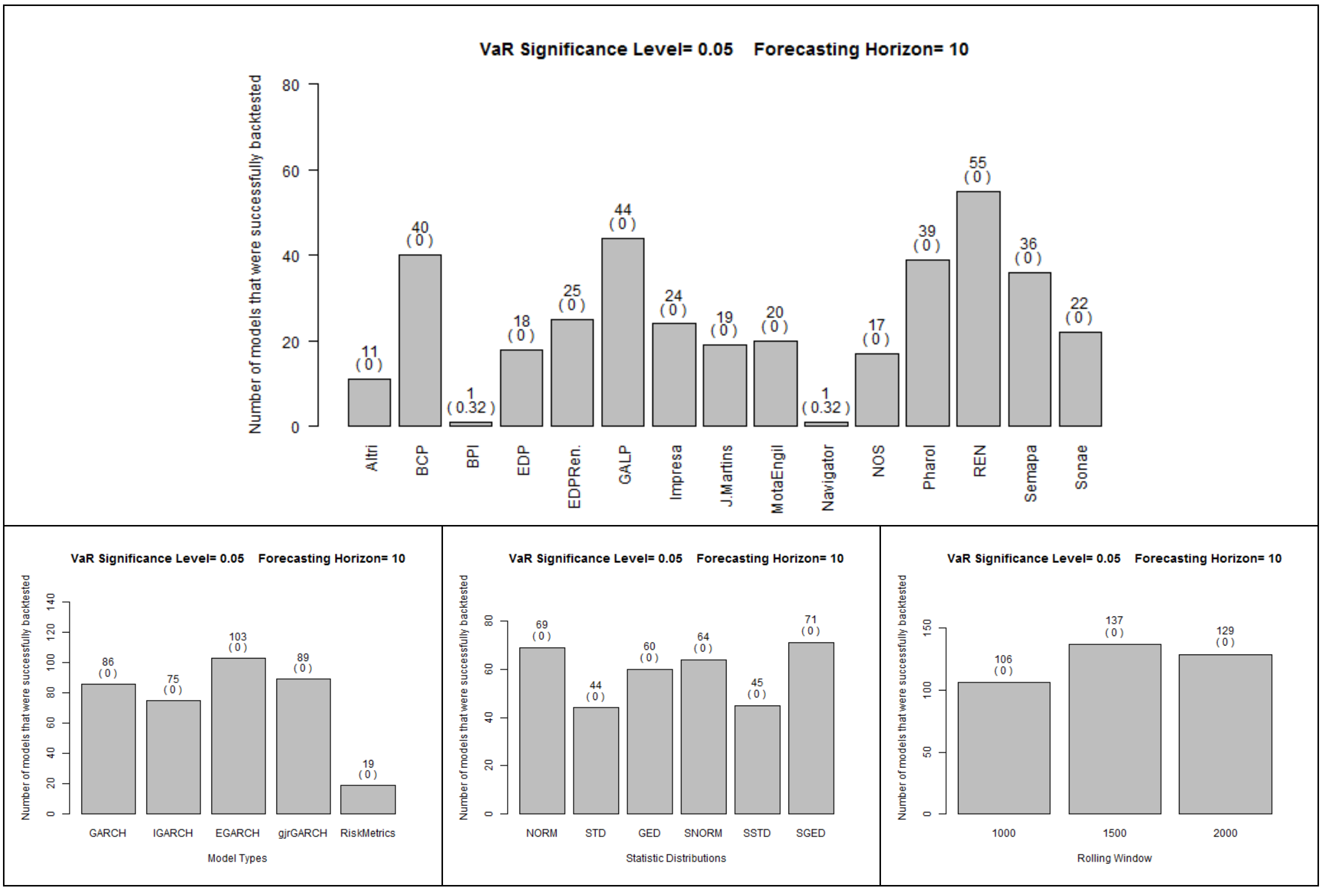

Figure 6 - Number of cases where the null hypothesis of the POF and TUFF tests was not rejected for the firms' data set. Forecast horizon=10 and α=0,05. (t-test p value in parenthesis)

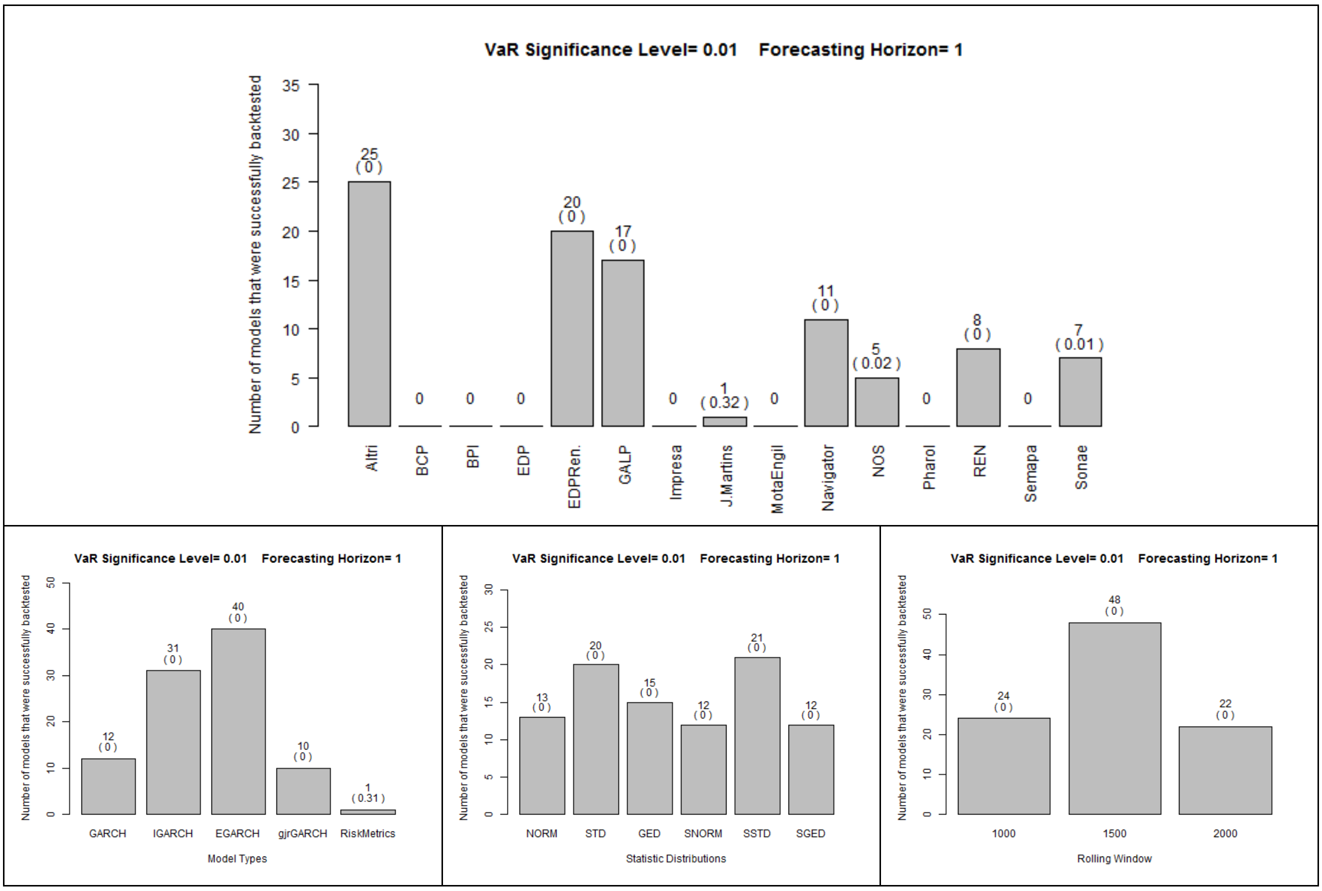

Figure 7 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ and DBI tests was not rejected for the firms' data set. Forecast horizon=1 and α=0,01. (t-test p value in parenthesis)

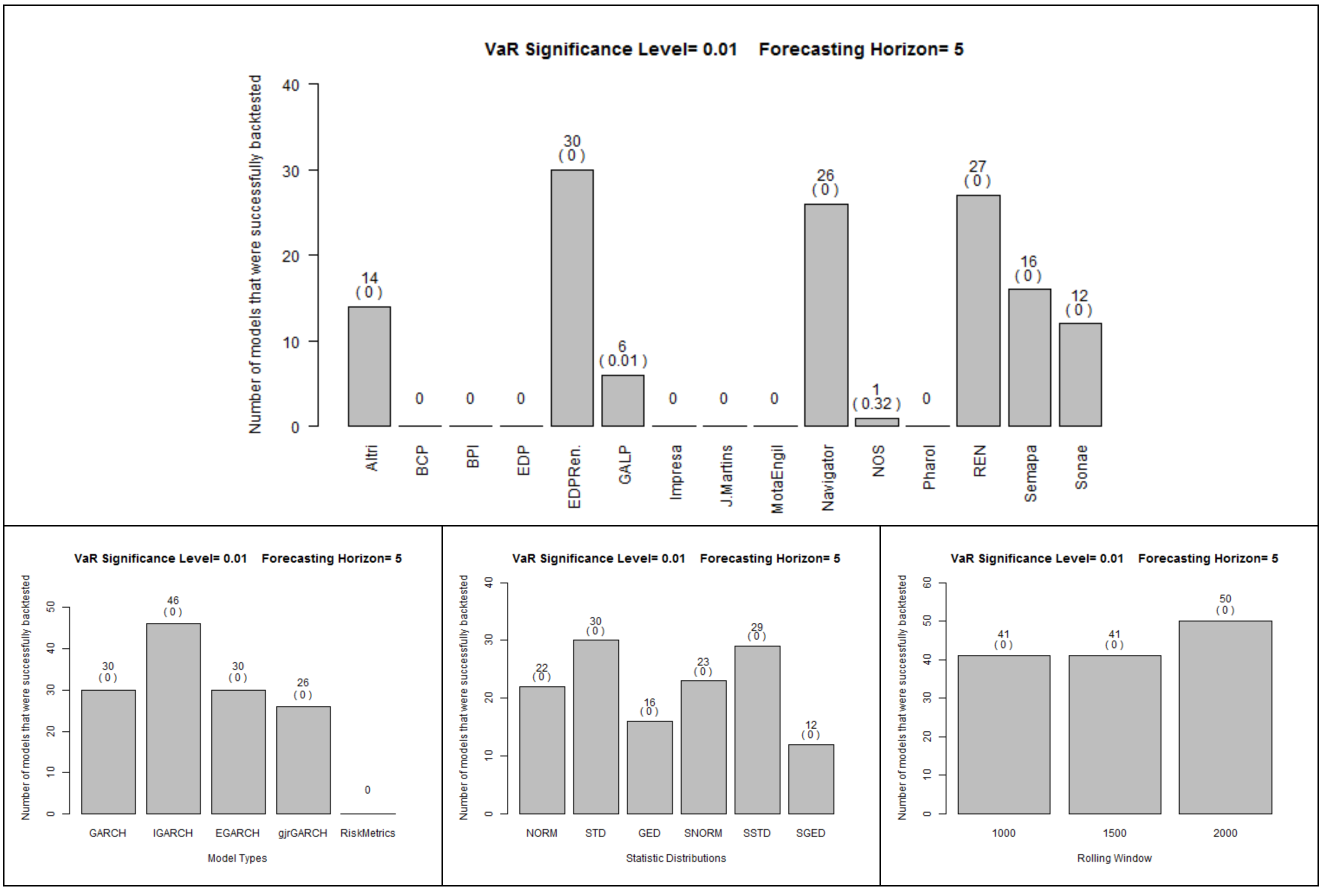

Figure 8 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ and DBI tests was not rejected for the firms' data set. Forecast horizon=5 and α=0,01. (t-test p value in parenthesis)

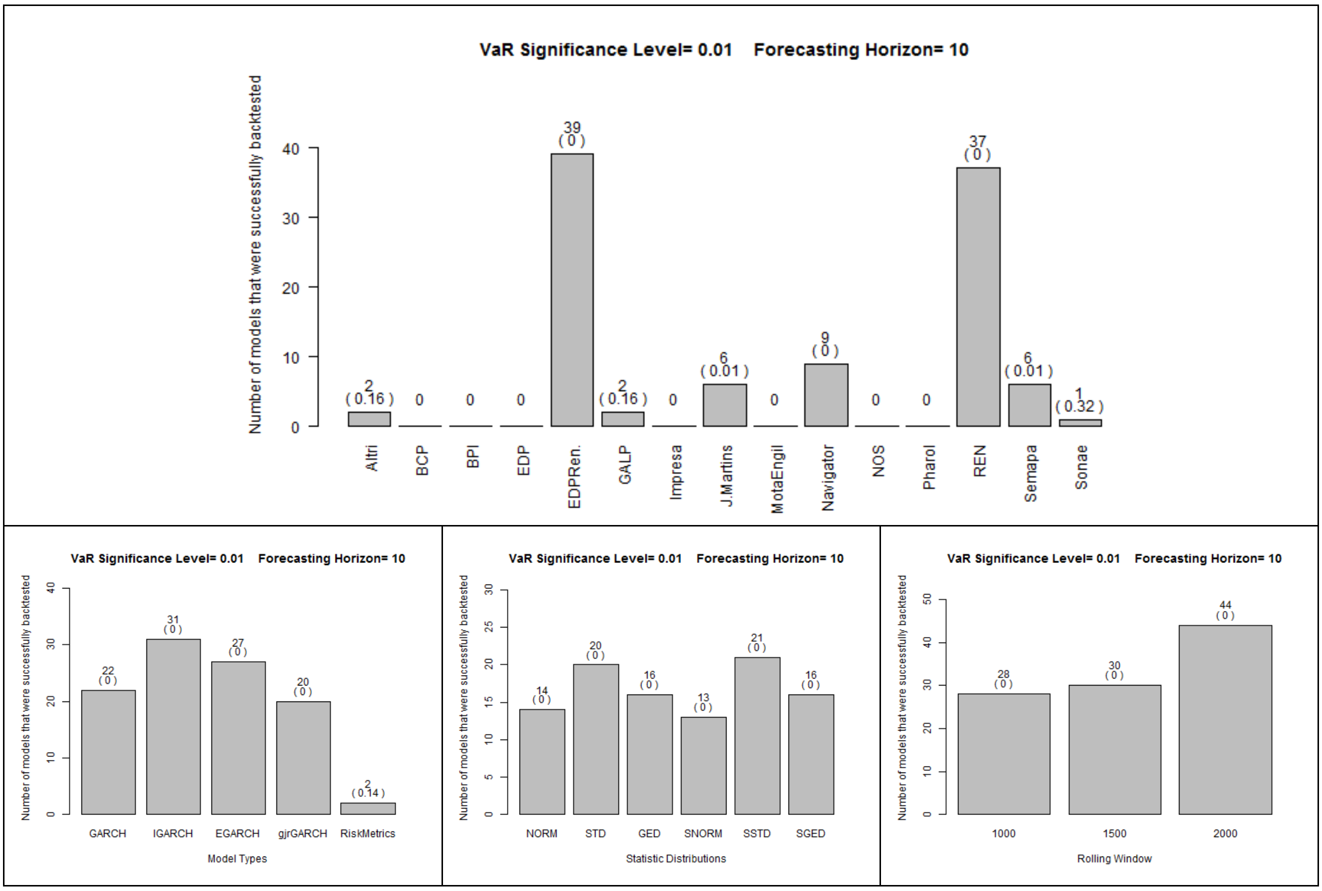

Figure 9 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ and DBI tests was not rejected for the firms' data set. Forecast horizon=10 and α=0,01. (t-test p value in parenthesis)

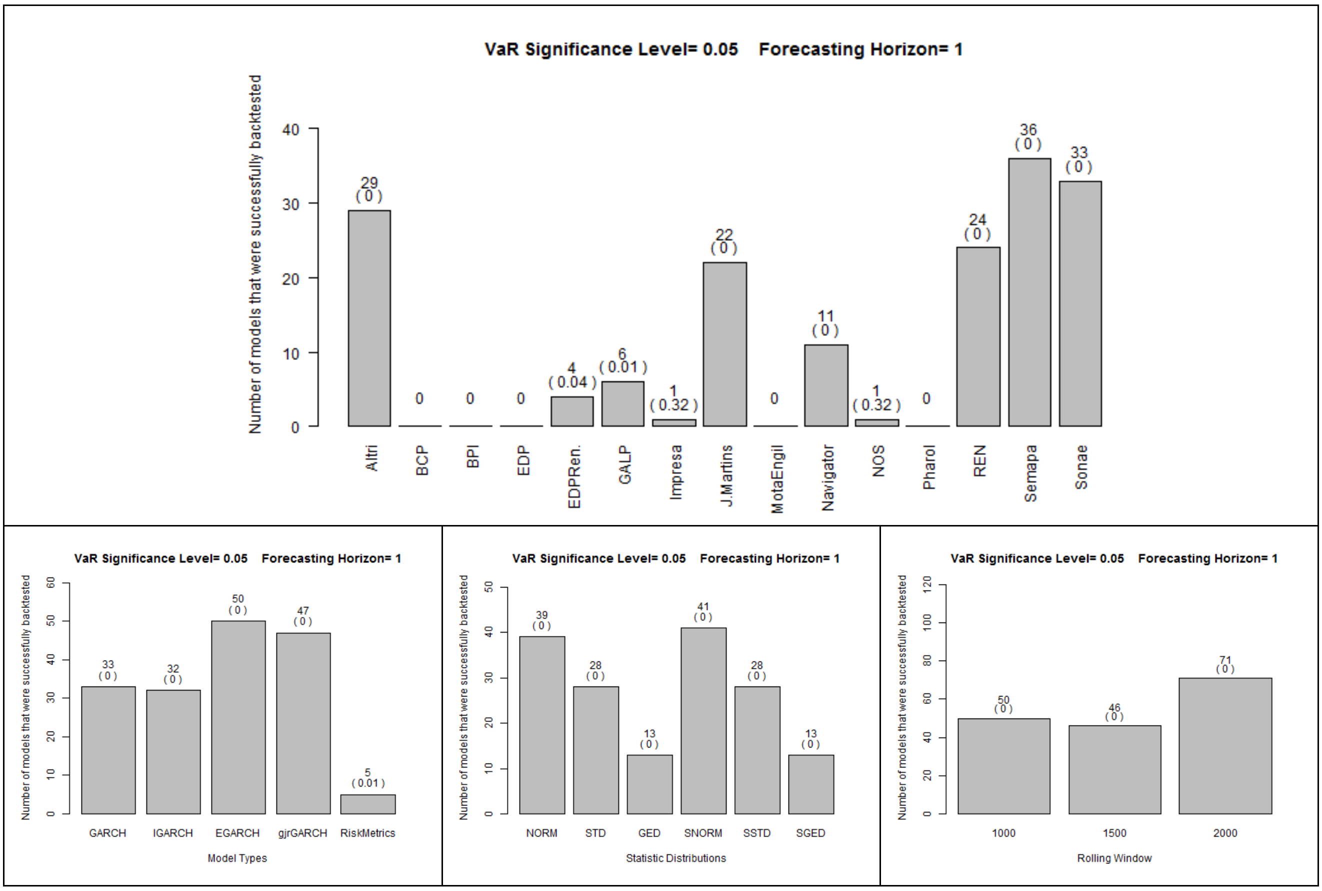

Figure 10 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ and DBI tests was not rejected for the firms' data set. Forecast horizon=1 and α=0,05. (t-test p value in parenthesis)

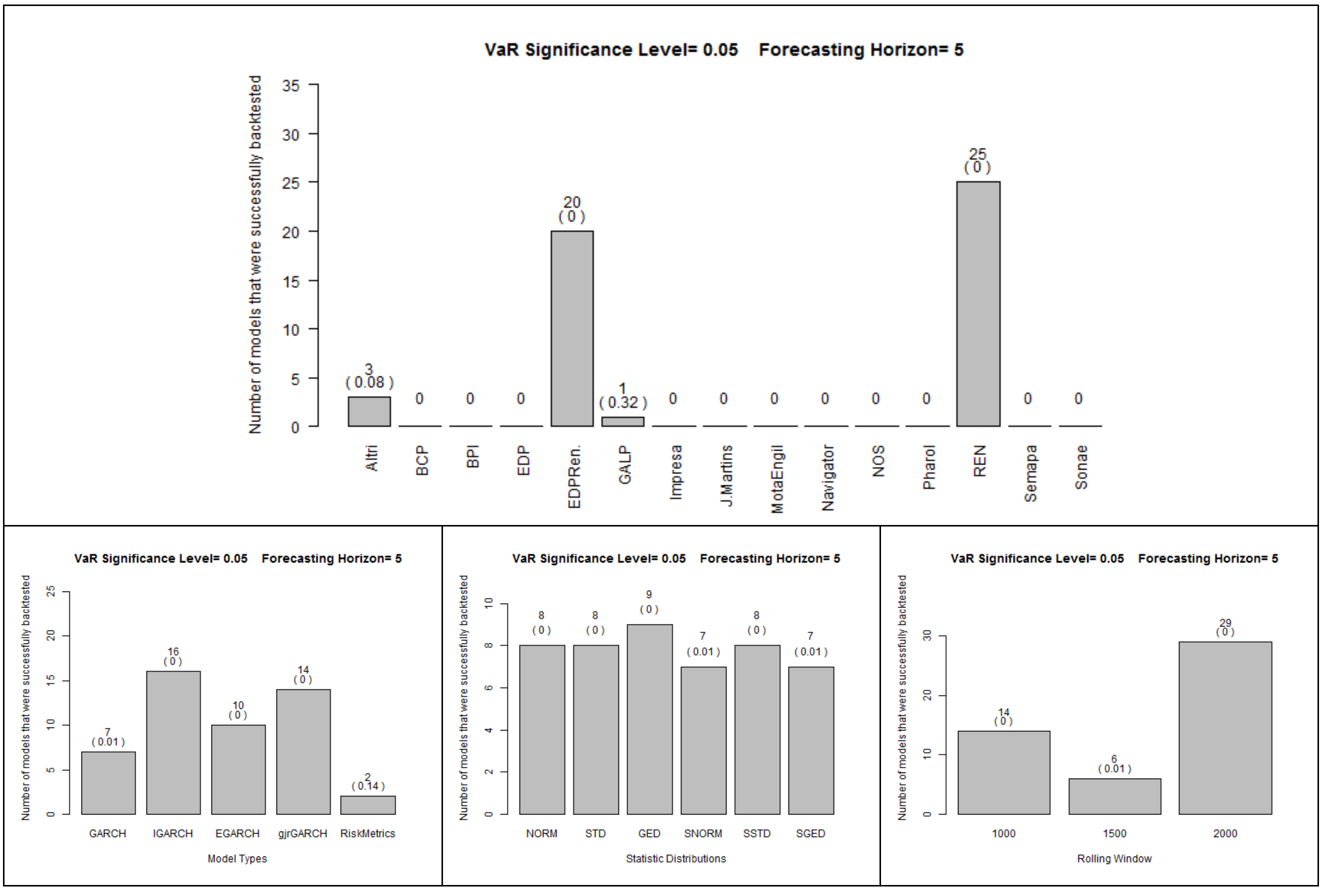

Figure 11 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ and DBI tests was not rejected for the firms' data set. Forecast horizon=5 and α=0,05. (t-test p value in parenthesis)

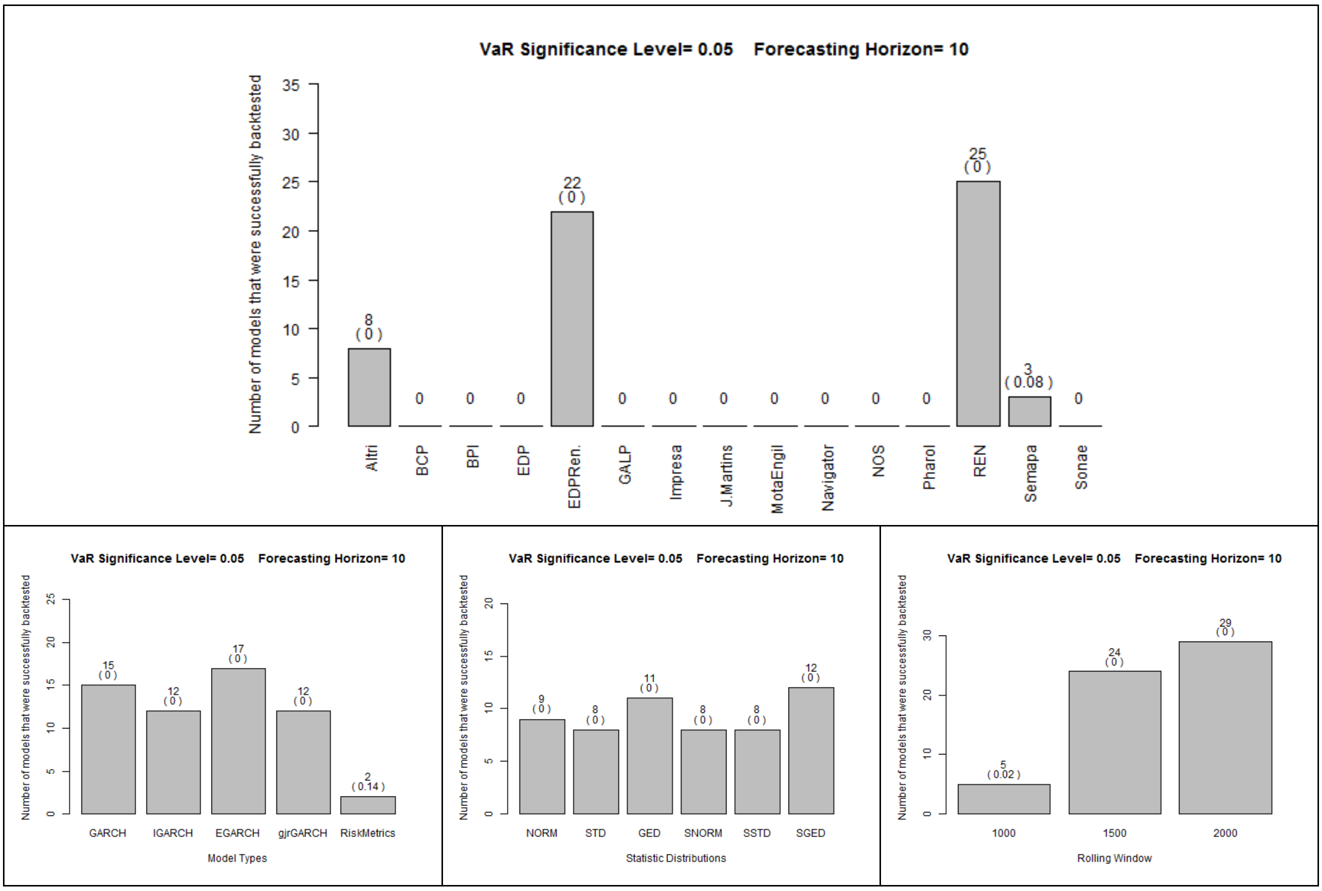

Figure 12 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ and DBI tests was not rejected for the firms' data set. Forecast horizon=10 and α=0,05. (t-test p value in parenthesis)

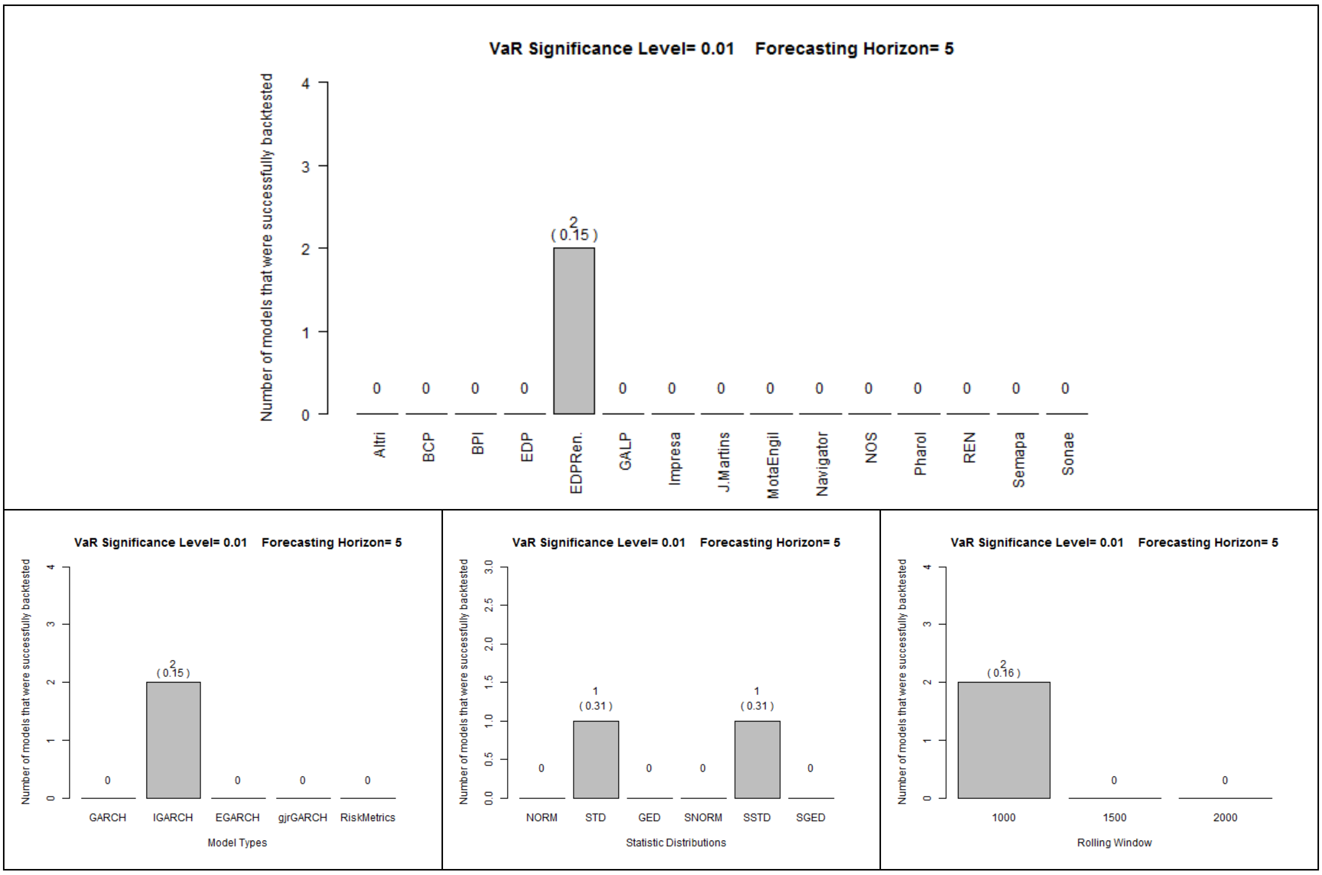

Figure 13 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ, DBI, MFE, AS1, AS2 and EKT tests was not rejected for the firms' data set. Forecast horizon=5 and α=0,01. (t-test p value in parenthesis)

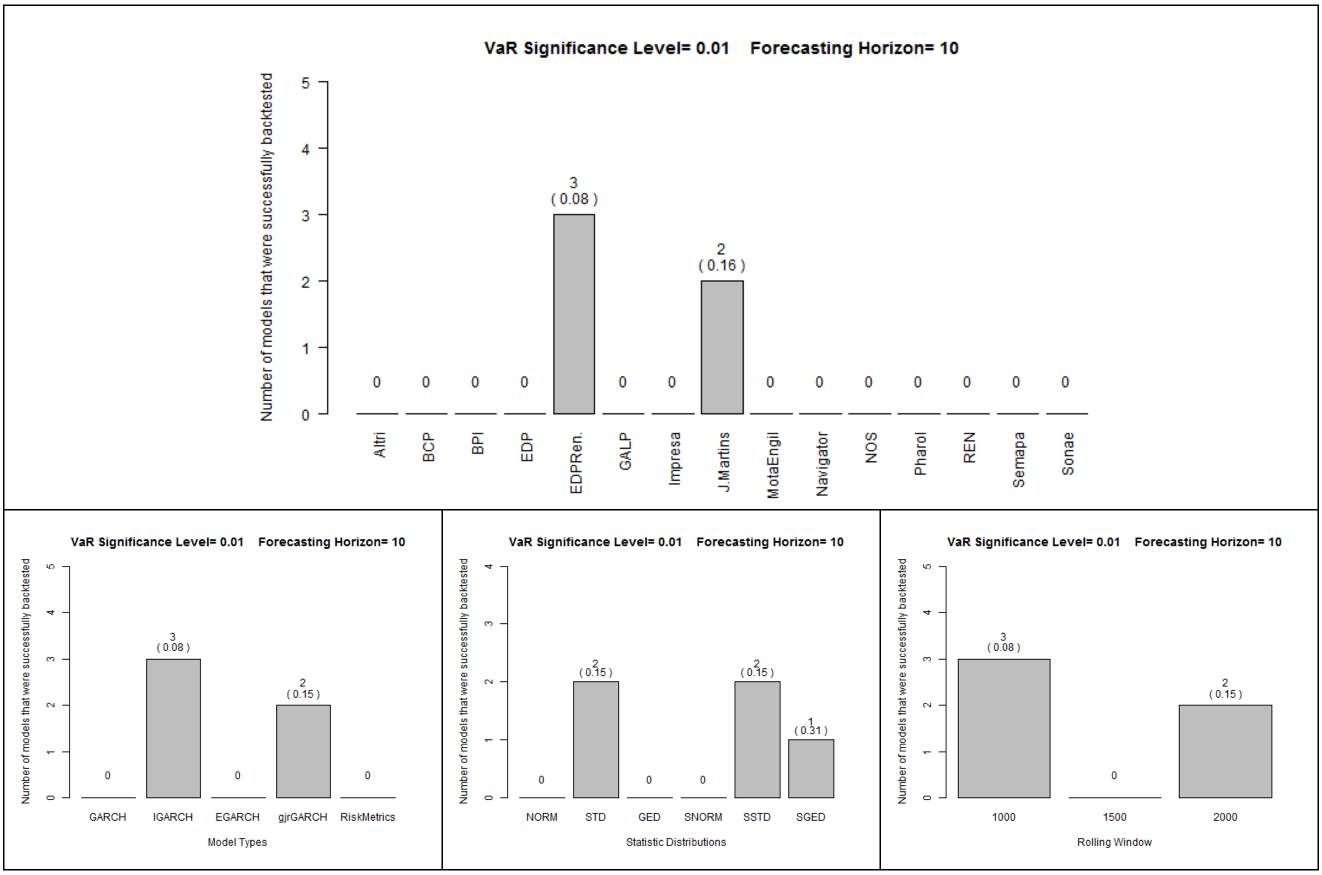

Figure 14 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ, DBI, MFE, AS1, AS2 and EKT tests was not rejected for the firms' data set. Forecast horizon=10 and α=0,01. (t-test p value in parenthesis)

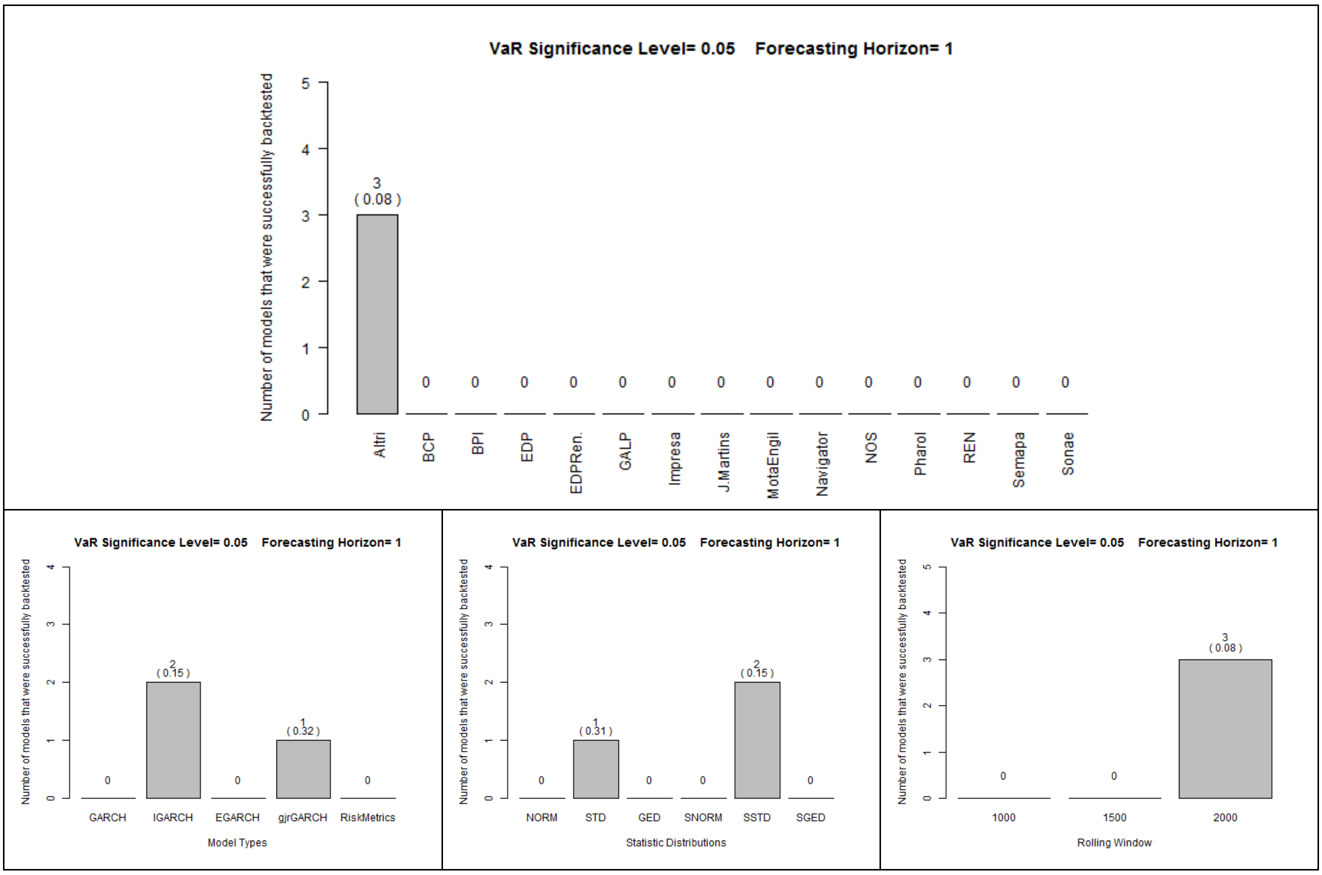

Figure 15 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ, DBI, MFE, AS1, AS2 and EKT tests was not rejected for the firms' data set. Forecast horizon=1 and α=0,05. (t-test p value in parenthesis)

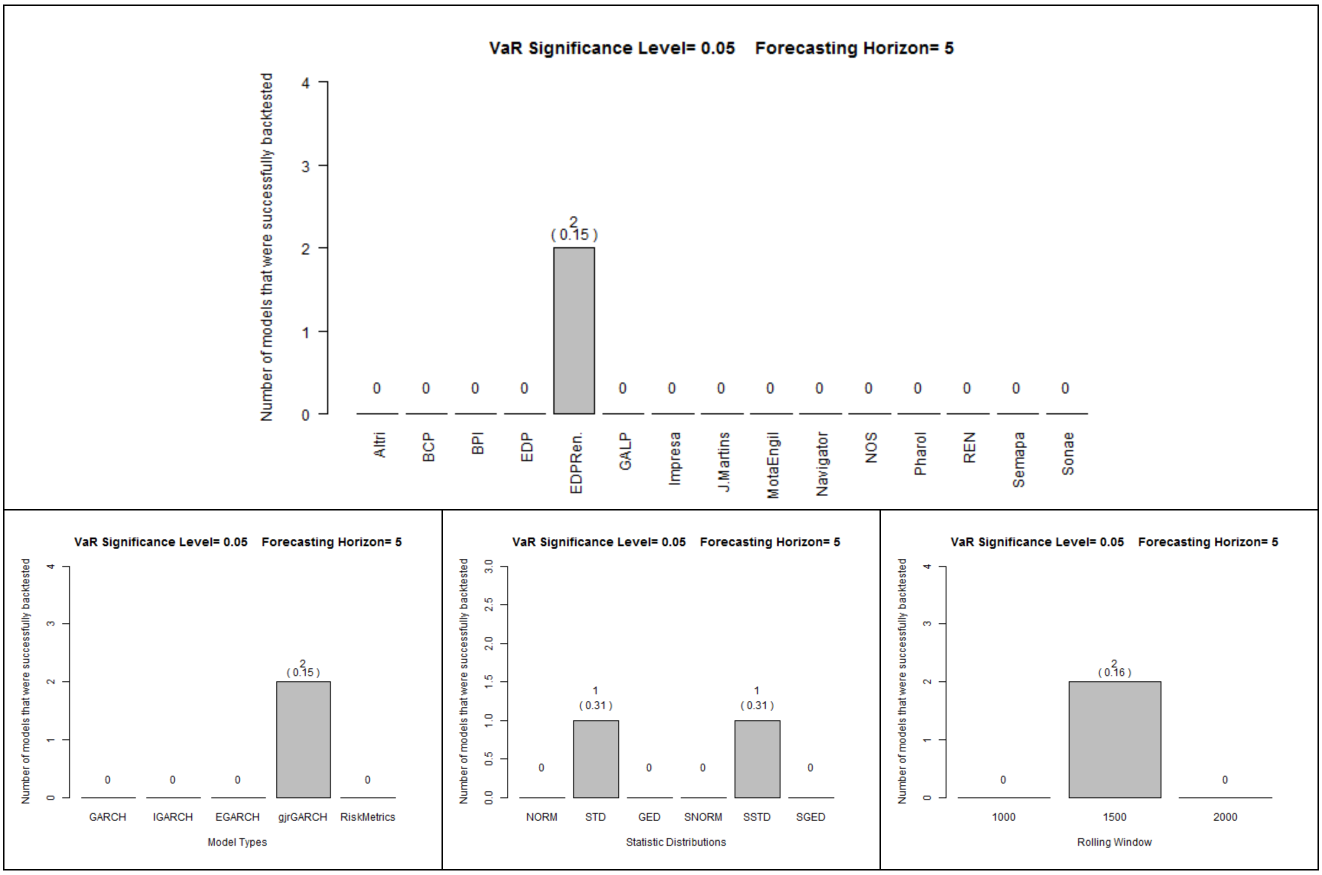

Figure 16 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ, DBI, MFE, AS1, AS2 and EKT tests was not rejected for the firms' data set. Forecast horizon=5 and α=0,05. (t-test p value in parenthesis)

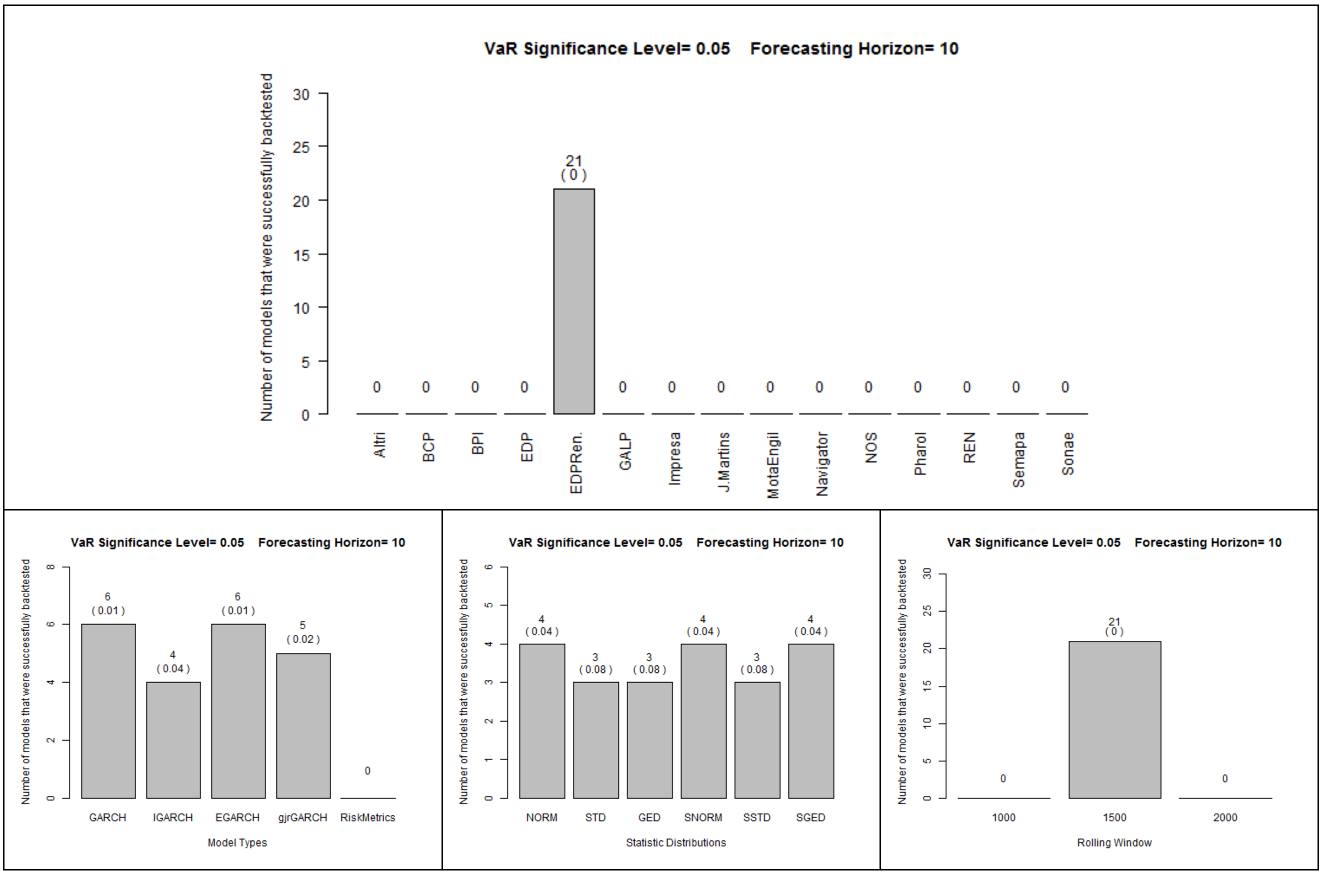

Figure 17 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ, DBI, MFE, AS1, AS2 and EKT tests was not rejected for the firms' data set. Forecast horizon=10 and α=0,05. (t-test p value in parenthesis)

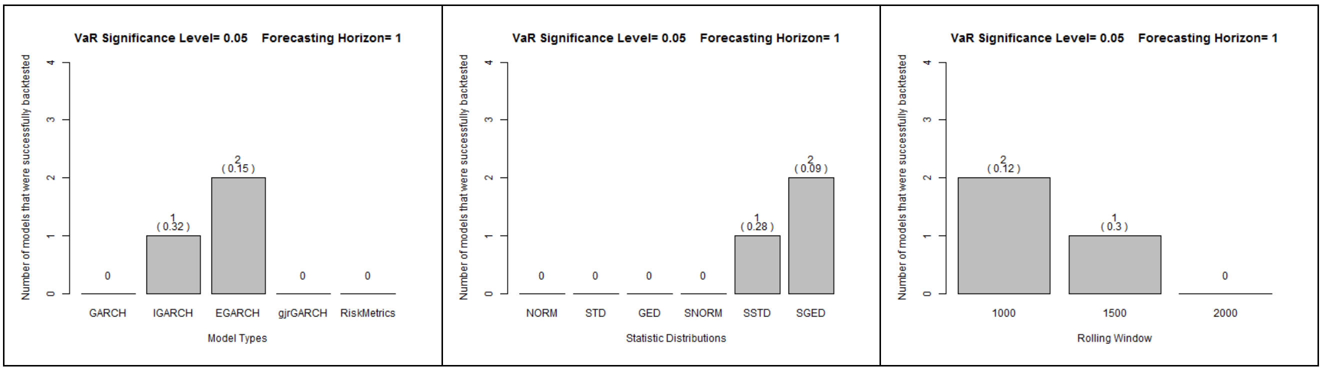

Figure 18 - Number of cases where the null hypothesis of the POF and TUFF tests was not rejected for the PSI20 data set. Forecast horizon=1 and α=0,05. (t-test p value in parenthesis)

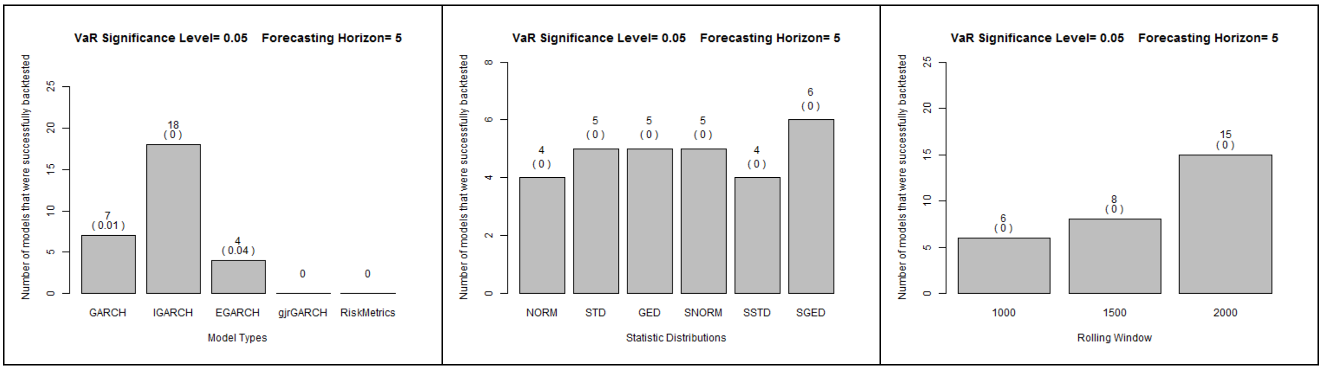

Figure 19 - Number of cases where the null hypothesis of the POF and TUFF tests was not rejected for the PSI20 data set. Forecast horizon=5 and α=0,05. (t-test p value in parenthesis)

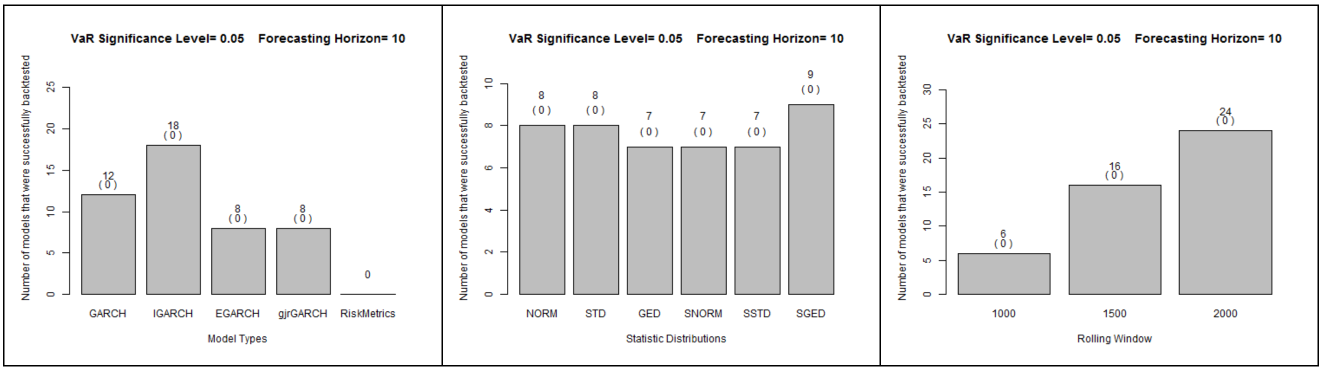

Figure 20 - Number of cases where the null hypothesis of the POF and TUFF tests was not rejected for the PSI20 data set. Forecast horizon=10 and α=0,05. (t-test p value in parenthesis)

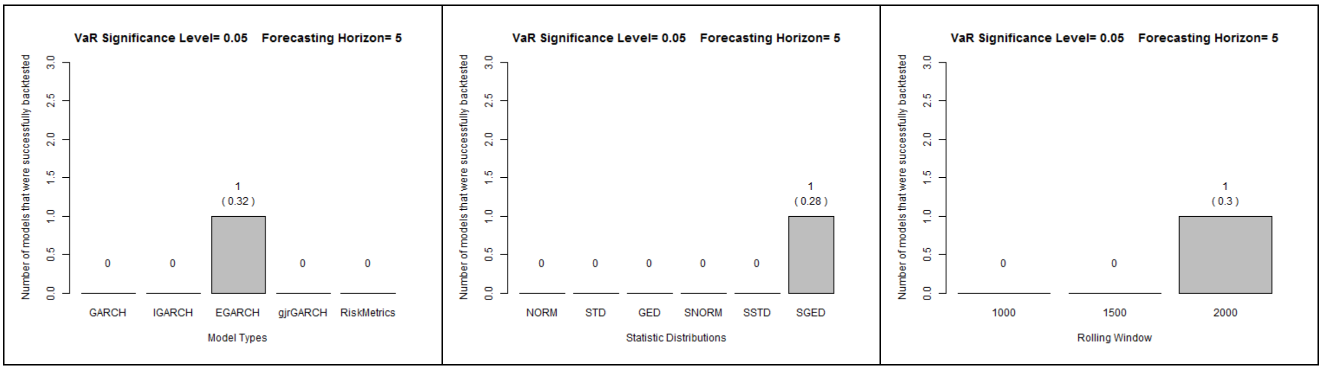

Figure 21 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ and DBI tests was not rejected for the PSI20 data set. Forecast horizon=5 and α=0,05. (t-test p value in parenthesis)

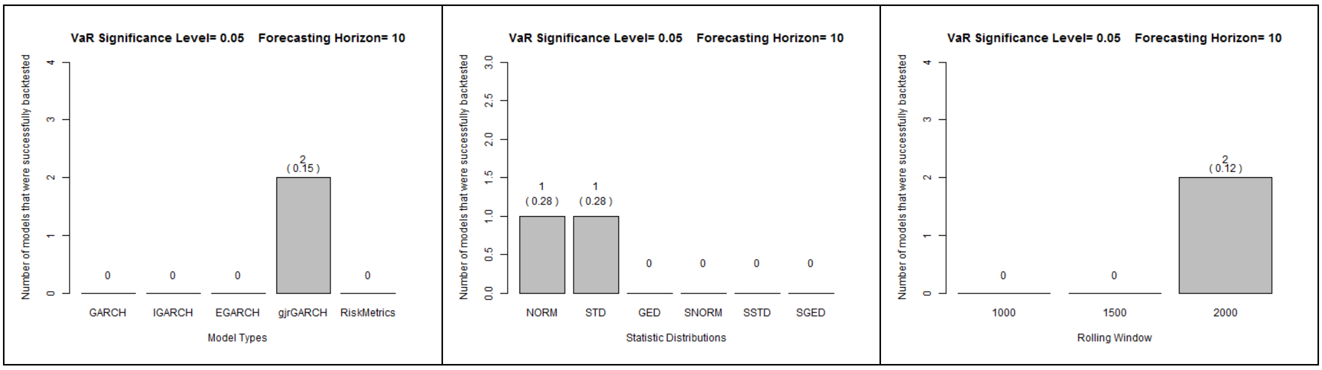

Figure 22 - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ and DBI tests was not rejected for the PSI20 data set. Forecast horizon=10 and α=0,05. (t-test p value in parenthesis)

Note: The following abbreviations are used in the figures:

- NORM – Normal Distribution
- STD – Student-t Distribution
- GED – Generalized Error Distribution
- SNORM – Skewed Normal Distribution
- SSTD – Skewed Student-t Distribution
- SGED – Skewed Generalized Error Distribution

ANNEX B – Tables

Table I - ***Basel Committee Traffic Light Approach***

Zone	Green Light					Yellow Light					Red Light
Nº of Violations	0	1	2	3	4	5	6	7	8	9	≥10
F	0	0	0	0	0	0.4	0.5	0.65	0.75	0.85	1

Table II - Breakdown of the unconditional coverage backtests null hypothesis rejection. (t-test p value in parenthesis)

			Null Hypothesis Rejected By					
			Forecast Horizon = 1		Forecast Horizon = 5		Forecast Horizon = 10	
			POF	TUFF	POF	TUFF	POF	TUFF
Null Hypothesis Not Rejected By	α=0.01	POF		61 (0)		108 (0)		172 (0)
		TUFF	616 (0)		455 (0)		341 (0)	
	α=0.05	POF		198 (0)		128 (0)		122 (0)
		TUFF	259 (0)		497 (0)		545 (0)	

Table III - Breakdown of the clustering backtests null hypothesis rejection. (t-test p value in parenthesis)

			Null Hypothesis Rejected By											
			Forecast Horizon = 1				Forecast Horizon = 5				Forecast Horizon = 10			
			IND	CC	DQ	DBI	IND	CC	DQ	DBI	IND	CC	DQ	DBI
Null Hypothesis Not Rejected By	α=0.01	IND		4 (0.04)	59 (0)	31 (0)		7 (0.01)	107 (0)	62 (0)		4 (0.04)	119 (0)	49 (0)
		CC	14 (0)		69 (0)	37 (0)	13 (0)		113 (0)	62 (0)	8 (0)		123 (0)	50 (0)
		DQ	0	0		13 (0)	0	0		11 (0)	0	0		4 (0.04)
		DBI	61 (0)	57 (0)	102 (0)		70 (0)	64 (0)	126 (0)		40 (0)	37 (0)	114 (0)	
	α=0.05	IND		15 (0)	86 (0)	6 (0.01)		4 (0.04)	55 (0)	19 (0)		1 (0.32)	78 (0)	39 (0)
		CC	46 (0)		104 (0)	10 (0)	40 (0)		80 (0)	28 (0)	33 (0)		108 (0)	60 (0)
		DQ	47 (0)	34 (0)		0	17 (0)	6 (0.01)		0	2 (0.16)	0		8 (0)
		DBI	319 (0)	292 (0)	352 (0)		99 (0)	72 (0)	118 (0)		58 (0)	47 (0)	103 (0)	

Table IV - Breakdown of the Expected Shortfall backtests null hypothesis rejection. (t-test p value in parenthesis)

			Null Hypothesis Rejected By											
			Forecast Horizon = 1				Forecast Horizon = 5				Forecast Horizon = 10			
			MFE	AS1	AS2	EKT	MFE	AS1	AS2	EKT	MFE	AS1	AS2	EKT
Null Hypothesis Not Rejected By	α=0.01	MFE		60 (0)	55 (0)	63 (0)		39 (0)	31 (0)	25 (0)		34 (0)	32 (0)	41 (0)
		AS1	0		2 (0.15)	5 (0.02)	0		4 (0.04)	4 (0.04)	0		8 (0)	16 (0)
		AS2	18 (0)	25 (0)		26 (0)	21 (0)	33 (0)		28 (0)	6 (0.01)	16 (0)		17 (0)
		EKT	2 (0.15)	4 (0.04)	2 (0.15)		2 (0.15)	20 (0)	15 (0)		3 (0.07)	12 (0)	5 (0.02)	
	α=0.05	MFE		40 (0)	5 (0.02)	26 (0)		31 (0)	3 (0.07)	23 (0)		26 (0)	2 (0.14)	8 (0)
		AS1	0		0	0	0		0	2 (0.14)	0		0	0
		AS2	87 (0)	122 (0)		110 (0)	13 (0)	41 (0)		27 (0)	12 (0)	36 (0)		19 (0)
		EKT	4 (0.04)	18 (0)	6 (0.01)		7 (0)	17 (0)	1 (0.31)		0	18 (0)	1 (0.31)	

Table V - Number of cases where the POF and TUFF tests did not reject the null hypothesis for the firms' data set. Verification of unconditional VaR coverage

			α = 0.01					α = 0.05				
			GARCH	IGARCH	EGARCH	GJR-GARCH	RiskMetrics	GARCH	IGARCH	EGARCH	GJR-GARCH	RiskMetrics
Forecast Horizon = 1	Rolling Window = 1000	Normal	1	3	1	2	0	9	9	8	9	9
		Student-t	1	3	6	2		8	5	5	9	
		GED	1	5	5	5		5	7	7	9	
		Skewed Normal	1	2	1	1		8	7	9	9	
		Skewed Student-t	1	4	6	3		8	5	5	9	
		Skewed GED	1	4	6	4		6	8	6	6	
	Rolling Window = 1500	Normal	2	4	3	2	0	10	11	12	12	11
		Student-t	3	4	8	3		11	9	7	11	
		GED	2	6	7	4		7	11	6	7	
		Skewed Normal	2	4	3	2		11	11	12	12	
		Skewed Student-t	3	4	8	3		11	9	7	11	
		Skewed GED	2	5	7	4		7	12	6	5	
	Rolling Window = 2000	Normal	2	2	3	2	2	9	10	10	10	11
		Student-t	2	3	8	2		10	10	7	10	
		GED	1	4	8	2		4	7	7	7	
		Skewed Normal	2	2	2	2		10	10	11	10	
		Skewed Student-t	1	3	8	2		10	10	7	10	
		Skewed GED	1	4	8	1		5	11	4	7	
Forecast Horizon = 5	Rolling Window = 1000	Normal	3	6	3	3	1	8	7	6	8	8
		Student-t	8	9	7	6		7	5	5	6	
		GED	3	8	5	6		7	8	6	6	
		Skewed Normal	3	7	3	3		8	7	7	6	
		Skewed Student-t	8	9	8	6		7	5	6	7	
		Skewed GED	2	6	5	3		8	9	6	6	
	Rolling Window = 1500	Normal	3	6	3	4	1	7	8	6	8	9
		Student-t	6	8	5	7		5	4	4	5	
		GED	3	7	7	6		5	7	5	6	
		Skewed Normal	4	5	4	4		6	7	5	8	
		Skewed Student-t	6	8	5	6		5	4	4	6	
		Skewed GED	2	4	7	5		6	10	6	4	
	Rolling Window = 2000	Normal	4	4	3	4	2	6	5	4	5	8
		Student-t	6	5	5	6		4	3	3	5	
		GED	3	6	6	4		3	8	3	5	
		Skewed Normal	4	4	4	3		5	5	4	5	
		Skewed Student-t	6	5	5	6		4	3	2	6	
		Skewed GED	2	4	8	4		4	10	3	4	
Forecast Horizon = 10	Rolling Window = 1000	Normal	4	5	3	3	1	6	3	6	6	5
		Student-t	6	6	4	6		2	1	5	2	
		GED	2	8	4	6		6	2	6	6	
		Skewed Normal	4	6	2	3		4	2	7	5	
		Skewed Student-t	6	5	4	6		2	1	5	2	
		Skewed GED	2	5	4	4		7	4	6	5	
	Rolling Window = 1500	Normal	3	4	3	4	1	7	5	6	8	8
		Student-t	6	4	4	5		4	4	5	5	
		GED	5	6	6	4		4	5	6	5	
		Skewed Normal	3	4	4	4		7	3	4	8	
		Skewed Student-t	6	4	4	5		4	4	5	5	
		Skewed GED	2	3	5	5		5	7	8	5	
	Rolling Window = 2000	Normal	3	4	3	3	1	5	4	7	6	6
		Student-t	6	5	5	6		3	5	5	3	
		GED	2	5	7	2		4	7	5	4	
		Skewed Normal	3	4	4	4		7	4	6	7	
		Skewed Student-t	6	5	4	5		4	5	5	3	
		Skewed GED	1	4	5	3		5	9	6	4	

Table VI - Number of cases where the POF, TUFF, IND, CC, DQ and DBI tests did not reject the null hypothesis for the firms' data set. Verification of unconditional VaR coverage and clustering.

			α = 0.01					α = 0.05				
			GARCH	IGARCH	EGARCH	GJR-GARCH	RiskMetrics	GARCH	IGARCH	EGARCH	GJR-GARCH	RiskMetrics
Forecast Horizon = 1	Rolling Window = 1000	Normal	0	1	0	1	0	4	1	2	4	1
		Student-t	0	1	3	1		1	1	3	3	
		GED	0	2	3	1		0	0	3	1	
		Skewed Normal	0	1	0	0		3	2	4	3	
		Skewed Student-t	0	3	3	1		1	1	3	3	
		Skewed GED	0	0	3	0		0	2	3	1	
	Rolling Window = 1500	Normal	1	2	3	1	0	3	3	4	4	1
		Student-t	1	3	4	1		2	1	1	3	
		GED	2	2	3	0		0	1	1	0	
		Skewed Normal	2	2	3	1		3	3	3	4	
		Skewed Student-t	2	3	4	1		2	1	1	3	
		Skewed GED	1	2	4	0		0	1	1	0	
	Rolling Window = 2000	Normal	1	1	1	1	1	3	2	5	4	3
		Student-t	1	3	2	0		3	3	3	4	
		GED	0	1	1	0		1	2	3	1	
		Skewed Normal	1	1	0	1		3	3	5	5	
		Skewed Student-t	0	2	2	0		3	3	3	4	
		Skewed GED	0	1	1	0		1	2	2	0	
Forecast Horizon = 5	Rolling Window = 1000	Normal	2	1	1	2	0	0	1	1	1	0
		Student-t	3	4	2	1		0	1	0	1	
		GED	0	3	3	0		1	1	0	1	
		Skewed Normal	2	2	1	2		0	1	1	1	
		Skewed Student-t	2	3	3	1		0	1	0	1	
		Skewed GED	0	2	1	0		0	1	0	0	
	Rolling Window = 1500	Normal	1	3	1	1	0	0	0	0	0	0
		Student-t	1	3	2	3		0	1	0	1	
		GED	1	2	1	0		0	0	1	0	
		Skewed Normal	2	3	1	1		0	0	0	0	
		Skewed Student-t	1	3	2	3		0	1	0	1	
		Skewed GED	1	2	3	0		0	0	1	0	
	Rolling Window = 2000	Normal	3	3	1	3	0	1	1	1	2	2
		Student-t	3	3	2	3		1	1	1	1	
		GED	1	2	2	1		1	2	1	1	
		Skewed Normal	3	3	1	2		1	1	1	1	
		Skewed Student-t	3	3	2	3		1	1	1	1	
		Skewed GED	1	1	1	0		1	2	1	1	
Forecast Horizon = 10	Rolling Window = 1000	Normal	1	2	1	1	0	1	0	0	0	0
		Student-t	2	1	1	1		0	0	1	0	
		GED	1	1	3	1		0	0	0	0	
		Skewed Normal	1	1	1	1		1	0	0	0	
		Skewed Student-t	2	1	1	1		0	0	1	0	
		Skewed GED	1	1	1	0		1	0	0	0	
	Rolling Window = 1500	Normal	1	1	1	1	1	1	1	1	1	0
		Student-t	1	2	1	1		1	0	1	1	
		GED	1	2	1	0		1	2	1	1	
		Skewed Normal	1	1	1	1		1	1	1	1	
		Skewed Student-t	1	2	2	1		1	0	1	1	
		Skewed GED	1	2	2	1		1	2	1	1	
	Rolling Window = 2000	Normal	1	2	1	1	1	1	0	2	1	2
		Student-t	2	3	2	3		1	1	1	1	
		GED	1	2	2	1		1	2	2	1	
		Skewed Normal	1	2	1	1		1	0	1	1	
		Skewed Student-t	2	3	2	3		1	1	1	1	
		Skewed GED	1	2	3	1		1	2	2	1	

Table VII - Number of cases where the POF, TUFF, IND, CC, DQ, DBI, MFE, AS and EKT tests did not reject the null hypothesis for the firms' data set. Verification of unconditional VaR coverage, VaR clustering and ES.

			α = 0.01					α = 0.05				
			GARCH	IGARCH	EGARCH	GJR-GARCH	RiskMetrics	GARCH	IGARCH	EGARCH	GJR-GARCH	RiskMetrics
Forecast Horizon = 1	Rolling Window = 1000	Normal	0	0	0	0	0	0	0	0	0	0
		Student-t	0	0	0	0		0	0	0	0	
		GED	0	0	0	0		0	0	0	0	
		Skewed Normal	0	0	0	0		0	0	0	0	
		Skewed Student-t	0	0	0	0		0	0	0	0	
		Skewed GED	0	0	0	0		0	0	0	0	
	Rolling Window = 1500	Normal	0	0	0	0	0	0	0	0	0	0
		Student-t	0	0	0	0		0	0	0	0	
		GED	0	0	0	0		0	0	0	0	
		Skewed Normal	0	0	0	0		0	0	0	0	
		Skewed Student-t	0	0	0	0		0	0	0	0	
		Skewed GED	0	0	0	0		0	0	0	0	
	Rolling Window = 2000	Normal	0	0	0	0	0	0	0	0	0	0
		Student-t	0	0	0	0		0	1	0	0	
		GED	0	0	0	0		0	0	0	0	
		Skewed Normal	0	0	0	0		0	0	0	0	
		Skewed Student-t	0	0	0	0		0	1	0	1	
		Skewed GED	0	0	0	0		0	0	0	0	
Forecast Horizon = 5	Rolling Window = 1000	Normal	0	0	0	0	0	0	0	0	0	0
		Student-t	0	1	0	0		0	0	0	0	
		GED	0	0	0	0		0	0	0	0	
		Skewed Normal	0	0	0	0		0	0	0	0	
		Skewed Student-t	0	1	0	0		0	0	0	0	
		Skewed GED	0	0	0	0		0	0	0	0	
	Rolling Window = 1500	Normal	0	0	0	0	0	0	0	0	0	0
		Student-t	0	0	0	0		0	0	0	1	
		GED	0	0	0	0		0	0	0	0	
		Skewed Normal	0	0	0	0		0	0	0	0	
		Skewed Student-t	0	0	0	0		0	0	0	1	
		Skewed GED	0	0	0	0		0	0	0	0	
	Rolling Window = 2000	Normal	0	0	0	0	0	0	0	0	0	0
		Student-t	0	0	0	0		0	0	0	0	
		GED	0	0	0	0		0	0	0	0	
		Skewed Normal	0	0	0	0		0	0	0	0	
		Skewed Student-t	0	0	0	0		0	0	0	0	
		Skewed GED	0	0	0	0		0	0	0	0	
Forecast Horizon = 10	Rolling Window = 1000	Normal	0	0	0	0	0	0	0	0	0	0
		Student-t	0	1	0	0		0	0	0	0	
		GED	0	0	0	0		0	0	0	0	
		Skewed Normal	0	0	0	0		0	0	0	0	
		Skewed Student-t	0	1	0	0		0	0	0	0	
		Skewed GED	0	1	0	0		0	0	0	0	
	Rolling Window = 1500	Normal	0	0	0	0	0	1	1	1	1	0
		Student-t	0	0	0	0		1	0	1	1	
		GED	0	0	0	0		1	1	1	0	
		Skewed Normal	0	0	0	0		1	1	1	1	
		Skewed Student-t	0	0	0	0		1	0	1	1	
		Skewed GED	0	0	0	0		1	1	1	1	
	Rolling Window = 2000	Normal	0	0	0	0	0	0	0	0	0	0
		Student-t	0	0	0	1		0	0	0	0	
		GED	0	0	0	0		0	0	0	0	
		Skewed Normal	0	0	0	0		0	0	0	0	
		Skewed Student-t	0	0	0	1		0	0	0	0	
		Skewed GED	0	0	0	0		0	0	0	0	

Table VIII - Number of cases where the null hypothesis of the POF and TUFF was not rejected for the PSI20 data set.

			α = 0.05				
			GARCH	IGARCH	EGARCH	GJR-GARCH	RiskMetrics
Forecast Horizon = 1	Rolling Window = 1000	Normal	0	0	0	0	0
		Student-t	0	0	0	0	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	1	0	0	
		Skewed GED	0	0	1	0	
	Rolling Window = 1500	Normal	0	0	0	0	0
		Student-t	0	0	0	0	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	0	0	0	
		Skewed GED	0	0	1	0	
	Rolling Window = 2000	Normal	0	0	0	0	0
		Student-t	0	0	0	0	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	0	0	0	
		Skewed GED	0	0	0	0	
Forecast Horizon = 5	Rolling Window = 1000	Normal	0	1	0	0	0
		Student-t	0	1	0	0	
		GED	0	1	0	0	
		Skewed Normal	0	1	0	0	
		Skewed Student-t	0	1	0	0	
		Skewed GED	0	1	0	0	
	Rolling Window = 1500	Normal	0	1	0	0	0
		Student-t	1	1	0	0	
		GED	0	1	0	0	
		Skewed Normal	0	1	0	0	
		Skewed Student-t	0	1	0	0	
		Skewed GED	0	1	1	0	
	Rolling Window = 2000	Normal	1	1	0	0	0
		Student-t	1	1	0	0	
		GED	1	1	1	0	
		Skewed Normal	1	1	1	0	
		Skewed Student-t	1	1	0	0	
		Skewed GED	1	1	1	0	
Forecast Horizon = 10	Rolling Window = 1000	Normal	0	1	0	0	0
		Student-t	0	1	0	0	
		GED	0	1	0	0	
		Skewed Normal	0	1	0	0	
		Skewed Student-t	0	1	0	0	
		Skewed GED	0	1	0	0	
	Rolling Window = 1500	Normal	1	1	1	0	0
		Student-t	1	1	0	1	
		GED	1	1	0	0	
		Skewed Normal	1	1	0	0	
		Skewed Student-t	1	1	0	0	
		Skewed GED	1	1	1	1	
	Rolling Window = 2000	Normal	1	1	1	1	0
		Student-t	1	1	1	1	
		GED	1	1	1	1	
		Skewed Normal	1	1	1	1	
		Skewed Student-t	1	1	1	1	
		Skewed GED	1	1	1	1	

Table IX - Number of cases where the null hypothesis of the POF, TUFF, IND, CC, DQ and DBI tests was not rejected for the PSI20 data set.

			α = 0.05				
			GARCH	IGARCH	EGARCH	GJR-GARCH	RiskMetrics
Forecast Horizon = 1	Rolling Window = 1000	Normal	0	0	0	0	0
		Student-t	0	0	0	0	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	0	0	0	
		Skewed GED	0	0	0	0	
	Rolling Window = 1500	Normal	0	0	0	0	0
		Student-t	0	0	0	0	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	0	0	0	
		Skewed GED	0	0	0	0	
	Rolling Window = 2000	Normal	0	0	0	0	0
		Student-t	0	0	0	0	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	0	0	0	
		Skewed GED	0	0	0	0	
Forecast Horizon = 5	Rolling Window = 1000	Normal	0	0	0	0	0
		Student-t	0	0	0	0	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	0	0	0	
		Skewed GED	0	0	0	0	
	Rolling Window = 1500	Normal	0	0	0	0	0
		Student-t	0	0	0	0	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	0	0	0	
		Skewed GED	0	0	0	0	
	Rolling Window = 2000	Normal	0	0	0	0	0
		Student-t	0	0	0	0	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	0	0	0	
		Skewed GED	0	0	1	0	
Forecast Horizon = 10	Rolling Window = 1000	Normal	0	0	0	0	0
		Student-t	0	0	0	0	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	0	0	0	
		Skewed GED	0	0	0	0	
	Rolling Window = 1500	Normal	0	0	0	0	0
		Student-t	0	0	0	0	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	0	0	0	
		Skewed GED	0	0	0	0	
	Rolling Window = 2000	Normal	0	0	0	1	0
		Student-t	0	0	0	1	
		GED	0	0	0	0	
		Skewed Normal	0	0	0	0	
		Skewed Student-t	0	0	0	0	
		Skewed GED	0	0	0	0	

Table X - Results of t-tests for the comparison of the number of successfully backtested cases per model for the firms' data set

BackTests	α	Forecast Horizon	GARCH IGARCH	GARCH EGARCH	GARCH GJR-GARCH	GARCH RiskMetrics	IGARCH EGARCH	IGARCH GJR-GARCH	IGARCH RiskMetrics	EGARCH GJR-GARCH	EGARCH RiskMetrics	GJR-GARCH RiskMetrics
Unconditional Coverage	0.01	1	0	0	0.02	0.07	0	0.02	0	0	0	0
		5	0	0.05	0.26	0	0.04	0	0	0.43	0	0
		10	0.05	0.57	0.36	0	0.17	0.31	0	0.73	0	0
	0.05	1	0.07	0.09	0.05	0	0	0.89	0.03	0	0	0.04
		5	0.25	0.02	0.91	0	0	0.3	0.01	0.02	0	0
		10	0.21	0.05	0.73	0.08	0	0.11	0.02	0.11	0.49	0.12
Clustering	0.01	1	0	0	0.66	0.37	0.22	0	0	0	0	0.54
		5	0.03	1	0.55	0	0.03	0.01	0	0.55	0	0
		10	0.17	0.43	0.74	0.28	0.55	0.09	0.04	0.26	0.11	0.38
	0.05	1	0.89	0.03	0.06	0.82	0.02	0.05	0.88	0.71	0.13	0.2
		5	0.05	0.45	0.11	0.56	0.21	0.7	0.65	0.39	0.82	0.82
		10	0.54	0.71	0.54	0.74	0.33	1	1	0.33	0.58	1
Expected Shortfall	0.01	5	0.15	NaN	NaN	NaN	0.15	0.15	0.15	NaN	NaN	NaN
		10	0.08	NaN	0.15	NaN	0.08	0.65	0.08	0.15	NaN	0.15
	0.05	1	0.15	NaN	0.32	NaN	0.15	0.56	0.15	0.32	NaN	0.32
		5	NaN	NaN	0.15	NaN	NaN	0.15	NaN	0.15	NaN	0.15
		10	0.52	1	0.76	0.01	0.52	0.73	0.04	0.76	0.01	0.02

Table XI - Results of t-tests for the comparison of the number of successfully backtested cases per distribution for the firms' data set

BackTests	α	Forecast Horizon	NORM STD	NORM GED	NORM SNORM	NORM SSTD	NORM SGED	STD GED	STD SNORM	STD SSTD	STD SGED	GED SNORM	GED SSTD	GED SGED	SNORM SSTD	SNORM SGED	SSTD SGED
Unconditional Coverage	0.01	1	0.01	0	0.62	0	0	0.5	0	0.89	0.78	0	0.57	0.67	0	0	0.89
		5	0	0.01	0.78	0	0.4	0.1	0	1	0	0.03	0.05	0.1	0	0.58	0
		10	0	0.03	0.67	0.01	0.89	0.4	0.01	0.68	0	0.09	0.68	0.05	0.04	0.78	0.02
	0.05	1	0	0	0.84	0	0	0	0	1	0	0	0	0.88	0	0	0
		5	0	0.2	0.48	0.01	0.77	0.1	0.02	0.68	0	0.57	0.16	0.32	0.05	0.67	0.02
		10	0	0.21	0.49	0	0.78	0	0	0.89	0	0.58	0.04	0.12	0.01	0.33	0
Clustering	0.01	1	0.18	0.68	0.83	0.13	0.83	0.4	0.12	0.86	0.12	0.53	0.26	0.53	0.08	1	0.08
		5	0.19	0.27	0.86	0.25	0.06	0	0.26	0.88	0	0.21	0.02	0.41	0.33	0.04	0
		10	0.25	0.69	0.83	0.19	0.69	0.5	0.18	0.86	0.46	0.54	0.36	1	0.13	0.54	0.36
	0.05	1	0.1	0	0.77	0.1	0	0	0.05	1	0.01	0	0.01	1	0.05	0	0.01
		5	1	0.8	0.79	1	0.79	0.8	0.79	1	0.79	0.6	0.8	0.6	0.79	1	0.79
		10	0.8	0.64	0.8	0.8	0.49	0.5	1	1	0.34	0.47	0.47	0.82	1	0.34	0.34
Expected Shortfall	0.01	5	0.31	NaN	NaN	0.31	NaN	0.3	0.31	1	0.31	NaN	0.31	NaN	0.31	NaN	0.31
		10	0.15	NaN	NaN	0.15	0.31	0.2	0.15	1	0.56	NaN	0.15	0.31	0.15	0.31	0.56
	0.05	1	0.31	NaN	NaN	0.15	NaN	0.3	0.31	0.56	0.31	NaN	0.15	NaN	0.15	NaN	0.15
		5	0.31	NaN	NaN	0.31	NaN	0.3	0.31	1	0.31	NaN	0.31	NaN	0.31	NaN	0.31
		10	0.7	0.7	1	0.7	1	1	0.7	1	0.7	0.7	1	0.7	0.7	1	0.7

Table XII - Results of t-tests for the comparison of the number of successfully backtested cases per rolling window size for the firms' data set

BackTests	α	Forecast Horizon	1000 1500	1000 2000	1500 2000
Unconditional Coverage	0.01	1	0.01	0.18	0.23
		5	0.64	0.35	0.65
		10	0.63	0.86	0.77
	0.05	1	0	0	0.43
		5	0.06	0	0.02
		10	0	0	0.91
Clustering	0.01	1	0	0.95	0
		5	1	0.15	0.15
		10	0.78	0.02	0.03
	0.05	1	0.64	0.01	0
		5	0.06	0.01	0
		10	0	0	0.31
Expected Shortfall	0.01	5	0.16	0.16	NaN
		10	0.08	0.71	0.15
	0.05	1	NaN	0.08	0.08
		5	0.16	NaN	0.16
		10	0	NaN	0

Table XIII - Results of t-tests for the comparison of the number of successfully backtested cases per model for the PSI20 data set

BackTests	α	Forecast Horizon	GARCH IGARCH	GARCH EGARCH	GARCH GJR-GARCH	GARCH RiskMetrics	IGARCH EGARCH	IGARCH GJR-GARCH	IGARCH RiskMetrics	EGARCH GJR-GARCH	EGARCH RiskMetrics	GJR-GARCH RiskMetrics
Unconditional Coverage	0.05	1	0.29	0.11	NaN	NaN	0.52	0.29	0.29	0.11	0.11	NaN
		5	0	0.18	0	0	0	0	0	0.01	0.01	NaN
		10	0	0.04	0.04	0	0	0	0	1	0	0
Clustering	0.05	10	NaN	NaN	0.11	NaN	NaN	0.11	NaN	0.11	NaN	0.11

Table XIV - Results of t-tests for the comparison of the number of successfully backtested cases per distribution for the PSI20 data set

BackTests	α	Forecast Horizon	NORM STD	NORM GED	NORM SNORM	NORM SSTD	NORM SGED	STD GED	STD SNORM	STD SSTD	STD SGED	GED SNORM	GED SSTD	GED SGED	SNORM SSTD	SNORM SGED	SSTD SGED
Unconditional Coverage	0.05	1	NaN	NaN	NaN	0.28	0.09	NaN	NaN	0.28	0.09	NaN	0.28	0.09	0.28	0.09	0.5
		5	0.59	0.59	0.59	1	0.27	1	1	0.59	0.58	1	0.59	0.58	0.59	0.58	0.27
		10	1	0.49	0.49	0.49	0.41	0.5	0.49	0.49	0.41	1	1	0.13	1	0.13	0.13
Clustering	0.05	10	1	0.28	0.28	0.28	0.28	0.3	0.28	0.28	0.28	NaN	NaN	NaN	NaN	NaN	NaN

Table XV - Results of t-tests for the comparison of the number of successfully backtested cases per Rolling Window size for the PSI20 data set

BackTests	α	Forecast Horizon	1000 1500	1000 2000	1500 2000
Unconditional Coverage	0.05	1	0.54	0.12	0.3
		5	0.45	0	0
		10	0	0	0
Clustering	0.05	10	NaN	0.12	0.12