ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Bachelor Thesis Econometrie en Operationele Research

A Realized Quantile Autoregressive Approach to Predict the Risk of Stock Returns

Name student: Casper Roelofs
Student ID number: 475609

Supervisor: Opschoor, P.A.
Second assessor: Koning, N.W.
Date final version: July 4, 2021

Abstract

This thesis introduces a new quantile autoregressive model, the Real-QAR, to forecast the downside financial risk of stock returns. This model extends the QAR model of Baur and Dimpfl (2019) by adding the lagged realized volatility as an additional term. The Real-QAR, QAR, EGARCH, and TGARCH models are evaluated by means of their capability to predict the (one-day-ahead) conditional density and two popular measures of financial risk (Value-at-Risk (VaR) and Expected Shortfall (ES)). Daily stock exchange data are used from January 1, 2000 to May 7, 2021 for nineteen stock indices. The findings show that the new proposed Real-QAR model is best at predicting conditional density, the VaR, and the VaR and ES jointly. However, the TGARCH model seems to be best at predicting the ES stand-alone.

Contents

1 Introduction

Forecasting the volatility of asset returns is an important task in financial econometrics as the volatility of stocks is more predictable than the level of returns. Thereby, the understanding of the behaviour of the volatility process very is important for practitioners in the financial industry since the behaviour of the volatility process has important applications in option pricing, portfolio optimisation, and risk management.

A well-known model for forecasting conditional volatility is the traditional GARCH model, which was introduced by Bollerslev (1986). It is based on the ARCH model that is presented by Engle (1982) and has been widely used to predict the volatility of stocks. Many variations of the traditional GARCH model have been proposed ever since (e.g., see the list of various GARCH models of Hansen and Lunde (2005)). Many of the more sophisticated GARCH models also accounted for the asymmetric response of the conditional variance to positive and negative lagged returns, e.g. the threshold-GARCH (TGARCH) model by Glosten et al. (1993). Recently, Baur and Dimpfl (2019) propose to estimate the volatility of financial returns in a quantile autoregressive (QAR) framework in which they are also able to identify the asymmetric response of volatility to lagged returns. In this QAR framework, the conditional variance is based on the predicted quantiles.

Nevertheless, in the last years, it seems that the focus of the financial literature has shifted to predicting the downside risk or the full density of stock returns rather than solely predicting volatility point forecasts (De Raaij and Raunig, 2005). The evaluation of the point forecasts depends on the choice of a proxy for the volatility, while the true value is unobservable. The choice of this proxy may be very time-consuming or costly to obtain when using realized variance, or very noisy when taking squared returns (Hansen et al., 2003; Andersen et al., 2003). Taking the latter case, this can thus lead to inaccurately predicting volatility point forecasts, or incorrectly treating forecasts as accurate since the uncertainty of the proxy is not taken into account.

To correctly handle uncertainty, Gneiting and Katzfuss (2014) argue that the forecasts should be probabilistic, which means that we should focus on forecasting the conditional probability distributions. Although financial returns are not predictable, Timmermann (2000) and Andersen et al. (2003) argue that we still need the knowledge of the underlying return distribution since the distributional characteristics are key ingredients for the pricing of financial instruments. The advantage of using the conditional density in a financial setting is that the Value-at-Risk (VaR) and Expected Shortfall (ES), both measures of downside financial risk, are easily derived from the conditional density since these are functions (of the lower quantiles) of the conditional distribution function. Therefore, predicting one-day-ahead conditional density‘s has become significant in modern finance and plays a key role in risk management nowadays.

This thesis builds on the work of Baur and Dimpfl (2019) by extending the evaluation of their QAR model with conditional density forecast and Expected Shortfall evaluation. Thereby, a new model is introduced: the Real-QAR model. This model extends the QAR model of Baur and Dimpfl (2019) by adding the lagged realized volatility as an additional term.

In this thesis, the following question is investigated: *Which model is best at predicting the risk of stock returns*? This question is answered by evaluating two different aspects of financial risk. First, the one-day-ahead conditional density forecasts of different models are evaluated, then the indirect volatility forecast ability of those models are measured. The conditional density forecasts are evaluated using the probability integral transform (PIT), the logarithmic scoring rule, and the continuous ranked probability score (CRPS) as is done in Gneiting and Katzfuss (2014). Additionally, it is checked whether the models are able to capture the downside risk of financial returns as the performance of the VaR and ES forecasts are evaluated. The results show that the new proposed Real-QAR model is best at predicting the risk of stock returns as it outperforms, the EGARCH, the TGARCH, and the QAR model by Baur and Dimpfl (2019).

This thesis proceeds as follows. In Section 2, the motivation about the QAR models is given and a short summary of previous literature about density forecasting is provided. In Section 3, an extensive description of the data is given. Section 4 describes the different models considered and introduces the evaluation methods that are used to measure the performance of the models. Section 5 presents the results, whereas Section 6 summarizes and concludes.

2 Literature

It is well-known that large negative lagged returns have a larger impact on the conditional variance than (large) positive lagged returns. Two popular models that are able to capture this potential difference of impact are the asymmetric GARCH models of Nelson (1991) and Glosten et al. (1993). Baur and Dimpfl (2019) show in their work, both theoretically and empirically, that the QAR model is also able to capture the asymmetric response of the conditional variance to (large) positive and negative lagged returns. This result is shown by the fact that the range of possible returns today is wider if the return of yesterday is negative than if it is positive. This range is determined by the area spanned by the (quantile) regression lines. Since the range of possible returns can be seen as a proxy for the conditional variance, they state that the QAR model is able to capture this leverage effect. Baur and Dimpfl (2019) measure the leverage effect by using the inter-quantile range of low and high QAR estimates.

This thesis introduces a new model, the Real-QAR model. This model extends the normal

QAR model by adding the lagged realized volatility as an additional term. Baur and Dimpfl (2019) suggested including such an additional covariate in the QAR model, but they were critical about the performance of such a model since it led to a deterioration of the model performance regarding their evaluation methods (volatility and Value-at-Risk). However, it seems that adding this additional term in the QAR equation significantly improves the model based on the criteria in this thesis, which covers conditional density, Value-at-Risk, and Expected Shortfall evaluation.

As the QAR models are able to fully identify the conditional density, this model could be a very good alternative for measuring the risk of financial returns since you could easily derive measures of financial risk as Value-at-Risk (VaR) and Expected Shortfall (ES) from this conditional density. Nowadays, VaR is still a popular risk measure. Unfortunately, this measure does not fully account for the real loss since it is defined as a quantile of the predicted returns. Therefore, the Basel Committee on Banking Supervision (BIS) approved the ES to replace the VaR to measure market risk for banking regulatory purposes (Du and Escanciano, 2017). The ES, also known as the conditional VaR, is defined as the conditional expectation of loss given that the loss is beyond a particular VaR level (Degiannakis and Potamia, 2017). This thesis extends on the findings of Baur and Dimpfl (2019) by adding this indirect measure of volatility. The predicted ES of the models are evaluated by the regression-based tests introduced by Bayer and Dimitriadis (2020). These tests are special since they include two tests that besides the realized returns only require ES forecasts, which is in contrast to all other existing ES backtests. Additionally, the ES and VaR forecasts are jointly evaluated using the loss function of Fissler and Ziegel (2016).

The evaluation of the predicted (conditional) densities is done by means of the Pearson's Chi-squared test for goodness-of-fit as suggested by Diebold et al. (1998) and by two scoring functions, the logarithmic score and the continuous ranked probability score (CRPS) proposed by Gneiting and Raftery (2007). Wang et al. (2016), Mazur (2017), and Opschoor et al. (2017) also use these scoring functions to evaluate the performance of the density forecasts of financial returns.

3 Data

This thesis uses nineteen of the twenty international stock market indices as included by Baur and Dimpfl (2019). The data is obtained from the Oxford-Man Institute of Quantitative Finance by Heber et al. (2009). The FTSE MIB is excluded from this thesis since this market only has data available from June 1, 2009. This data contains daily stock exchange data from January 1, 2000 to May 7, 2021 leaving the last 500 observations as the forecast period ($\mathsf{T} = 500$) and the prior data as the fitting sample as is done by Baur and Dimpfl (2019). The indices have different numbers of observations because days where either the true return or the realized variance is missing are removed for that particular stock index. Following Baur and Dimpfl (2019), this thesis uses open-to-close log returns. Hence, the realized variance corresponds to these particular returns.

The descriptive statistics of the data are given in Table 1. The returns are in general negatively skewed and heavy-tailed as they show a kurtosis higher than six consistently. There is varied evidence given for the presence of first-order autocorrelation in the log returns by the Box–Ljung statistics ($\mathcal{Q}_r(1)$) since the null hypothesis of independent returns is not rejected for slightly less than half of the stock indices. Nevertheless, for all stock indices, an AR(1) model seems to capture the time series properties of the returns since the $\mathcal{Q}_{\hat{\varepsilon}}(1)$ shows that the residuals of the AR(1) model do not display significant autocorrelations. At last, the Box–Ljung statistic $\mathcal{Q}_{\hat{\varepsilon}^2}(5)$ shows a strong signal for the presence of heteroskedasticity. When allowing for five lags (a week), the test statistic indicates that all squared residuals show significant autocorrelations.

Table 1 *Descriptive statistics*

Market	Obs	Mean	SD	Min	Max	ζ	κ	$\mathcal{Q}_r(1)$	$\mathcal{Q}_{\hat{\varepsilon}}(1)$	$\mathcal{Q}_{\hat{\varepsilon}^2}(5)$
AEX	5439	-0.03	1.20	-8.42	9.24	-0.24	10.35	0.007	0.986	0.000
AORD	5390	0.00	0.85	-7.80	5.36	-0.75	9.55	0.000	0.938	0.000
BVSP	5250	-0.01	1.43	-9.80	11.92	0.02	7.67	0.000	0.893	0.000
DAX	5409	-0.03	1.25	-9.41	9.99	-0.13	8.29	0.535	0.917	0.000
DJI	5348	0.02	1.10	-8.41	10.75	-0.07	11.44	0.000	0.738	0.000
ES50	5349	-0.02	1.30	-11.50	8.27	-0.34	9.13	0.105	0.949	0.000
FCHI	5442	-0.03	1.15	-8.12	7.28	-0.19	7.86	0.001	0.878	0.000
FTSE	5382	0.00	1.16	-10.3	9.39	-0.36	10.23	0.004	0.831	0.000
HSI	5227	-0.04	1.04	-11.62	12.16	0.16	13.67	0.000	0.750	0.000
IBEX	5407	-0.05	1.21	-7.58	13.04	-0.06	8.96	0.525	0.887	0.000
KS11	5253	-0.05	1.14	-11.78	8.76	-0.44	13.26	0.000	0.665	0.000
MXX	5351	0.02	1.22	-8.26	9.95	-0.02	8.53	0.535	0.917	0.000
Nasdaq	5352	-0.01	1.32	-8.05	14.91	0.02	10.52	0.000	0.702	0.000
Nikkei	5186	-0.03	1.12	-10.56	11.66	-0.52	14.16	0.000	0.850	0.000
NSEI	5292	-0.05	1.18	-11.44	9.51	-0.56	11.26	0.129	0.846	0.000
Russel	5349	0.04	1.21	-12.38	8.06	-0.37	10.00	0.000	0.959	0.000
S&P500	5351	0.01	1.12	-9.35	10.22	-0.22	9.36	0.000	0.696	0.000
SSMI	5346	-0.02	0.97	-10.13	8.68	-0.50	11.54	0.972	0.997	0.000
STI	3411	-0.04	0.81	-4.88	4.99	-0.02	6.39	0.021	0.984	0.000

Note. The descriptive statistics of the open-to-close log returns in percentages ($100\% \cdot P_t^{close}/P_t^{open}$) for different stock indices. Obs denotes the number of observations per index. The standard deviation is measured by SD. The mean, standard deviation, minimum and maximum are described in terms of the return in percentages. The skewness and kurtosis are measured by ζ and κ, respectively. The last three columns correspond to p-values for Ljung–Box tests for autocorrelation. The autocorrelation is tested up to the lag given in parentheses and independent distribution is the null hypothesis for these tests. $\mathcal{Q}_r(1)$ tests the autocorrelation of the log return time series, whereas the residuals from an AR(1) model are tested by $\mathcal{Q}_{\hat{\varepsilon}}(1)$. $\mathcal{Q}_{\hat{\varepsilon}^2}(5)$ tests for heteroskedasticity in the returns as it uses the squared residuals of an AR(1) model.

4 Methodology

This section is structured as follows. Firstly, a description of all models used in this thesis is given. Secondly, an explanation of all evaluation methods is provided.

4.1 QAR(p) model

This thesis follows Baur and Dimpfl (2019) for the different models that are used. Their paper is mainly about the so-called quantile autoregressive (QAR) model. The idea of quantile regression is that it specifies a separate regression model for each quantile of a series instead of solely modelling the conditional mean of the distribution as is done with classic linear regression. Koenker and Xiao (2006) introduced the QAR model in a time series setting, where each conditional quantile of the time series y_t is expressed as linear function of its past observations. For the τ-th quantile of y_t, the QAR model is specified as follows:

$$Q_{y_t}(\tau|\mathcal{I}_{t-1}) = \theta_0(\tau) + \theta_1(\tau)y_{t-1} + \ \dots \ \theta(\tau)_p y_{t-p}, \tag{1}$$

where, $\mathcal{I}_{t-1}$ is the information set until time t, $\theta_j(\cdot)$ are functions mapping $\tau \in [0,\ 1] \rightarrow \mathbb{R}$. Each quantile τ can contain a different number of lags p in the linear function. Baur and Dimpfl (2019) show that under two assumptions of Koenker and Xiao (2006) the conditional moments of y_t are given as:

$$\mathbb{E}[y_t|\mathcal{I}_{t-1}] = \int_{-\infty}^{\infty} y_t \, \hat{f}(y_t|\mathcal{I}_{t-1}) \, dy_t = \mu_t, \tag{2}$$

$$Var[y_t|\mathcal{I}_{t-1}] = \int_{-\infty}^{\infty} (y_t - \mathbb{E}[y_t|\mathcal{I}_{t-1}])^2 \, \hat{f}(y_t|\mathcal{I}_{t-1}) \, dy_t = \sigma_t^2. \tag{3}$$

Here, $\hat{f}(y_t|\cdot)$ is the predicted conditional density and $\mathcal{I}_{t-1}$ is the information set until time t. The subscript t indicates that the conditional moment is time-varying. The derivation of conditional moments of the general QAR(p) model can be found in Appendix A of Baur and Dimpfl (2019).

To estimate the conditional variance of (financial) returns, Baur and Dimpfl (2019) propose a two-step guideline. In the first step, Equation (1) is used to get an estimator of the conditional distribution function as follows:

$$\hat{F}(y_t|\mathcal{I}_{t-1}) = \sup\{\tau \in [0,\ 1] \,|\, \hat{Q}_{y_t}(\tau|\mathcal{I}_{t-1}) \leq y_t\}. \tag{4}$$

Here, $\hat{Q}_{y_t}(\tau|y_{t-1}, \dots, y_{t-p})$ is the τ-th conditional quantile estimator suggested by Bassett and Koenker (1982). The estimation of the QAR models used in this thesis slightly differs from the QAR model of Baur and Dimpfl (2019). This thesis estimates 999 quantiles (from 0.001 up to

0.999) instead of only 99 quantiles (0.01 up to 99) since this creates even better predictions of downside financial risk. Since each quantile can have a different number of lags included, the Schwarz Bayesian Information Criterion (SIC) in its L1 form is used to determine the number of lags per quantile as suggested by Hurvich and Tsai (1990). The criterion is defined as follows:

$$\mathrm{SIC}(\tau) = n \log(\hat{s}^2) + p \cdot \log(n), \tag{5}$$

where $\hat{s}$ is the mean absolute distance of the data from the fitted values, n contains the number of observations, and p the number of lags included for the τ-th quantile.

From the estimated cdf $\hat{F}(y_t|y_{t-1}, \ldots, y_{t-p})$, a vector of predicted returns per quantile can be subtracted ($\hat{y}_{-t}$). From this vector $\hat{y}_{-t}$, which contains predicted returns one for each quantile, the empirical conditional probability density function ($\hat{f}(y_t|\mathcal{I}_{t-1})$) is constructed using the adaptive kernel method of Portnoy et al. (1989). This method is a non-parametric procedure to estimate the density of a random variable. Here, the vector of probabilities required coincides with the vector of quantiles used in the (QAR) model estimation, and the density is then calculated at the centres of the kernel, which are given by the predicted returns $\hat{y}_{-t}$. Based on the predicted density $\hat{f}(y_t|\mathcal{I}_{t-1})$, the conditional moments could be derived using Equations (2) and (3).

When using the QAR model, the one-day-ahead conditional density is subtracted using the adaptive kernel method which is used for evaluation.

4.2 Real-QAR(p) model

Besides the regular QAR model, this thesis introduces a new model: the Real-QAR model. As suggested by Baur and Dimpfl (2019) this model extends the regular QAR model by adding the lagged realized volatility as an additional term such that Equation (1) becomes:

$$Q_{y_t}(\tau|\mathcal{I}_{t-1}) = \theta_0(\tau) + \theta_1(\tau)y_{t-1} + \ldots \theta_p(\tau)y_{t-p} + \theta_{p+1}(\tau)\sigma_{RV,t-1}. \tag{6}$$

The Real-QAR(p) model also allows different quantiles to have a different number of lagged returns. However, this model also allows quantiles to exclude any lagged returns from the (quantile) regression equation, such that the equation only contains an intercept term and the lagged realized volatility. The reason for this is that it is plausible that the lagged return and the lagged realized volatility measure are highly correlated. An augmented version of the SIC in Equation (5) is used to determine the number of lagged returns per quantile. The following

SIC is used for the Real-QAR model:

$$\mathrm{SIC}(\tau) = n \log(\hat{s}^2) + (p+1) \cdot \log(n), \tag{7}$$

where p is the number of lagged returns included, $\hat{s}$ is the mean absolute distance of the data from the fitted values, and n contains the number of observations for the τ-th quantile. The criterion differs from its L1 form because it has a stronger penalty for adding lagged returns as it has the additional realized volatility term already included.

The average number of lags included for both models, the QAR and Real-QAR, are shown in Figure 1. It is shown that the average number of lags included for the tail quantiles is lower for the Real-QAR model due to the additional realized variance term. However, on average more lags are included for the median and central quantiles for the Real-QAR model, which suggests a higher persistent time series for the Real-QAR model. The same conclusions can be drawn from Figure 4 in the Appendix, where the median lag of all quantiles is shown for both models.

The Real-QAR model follows the same procedure in Equation (5) as the regular QAR model for obtaining a conditional distribution. From here, the same method is used for obtaining the conditional density function.

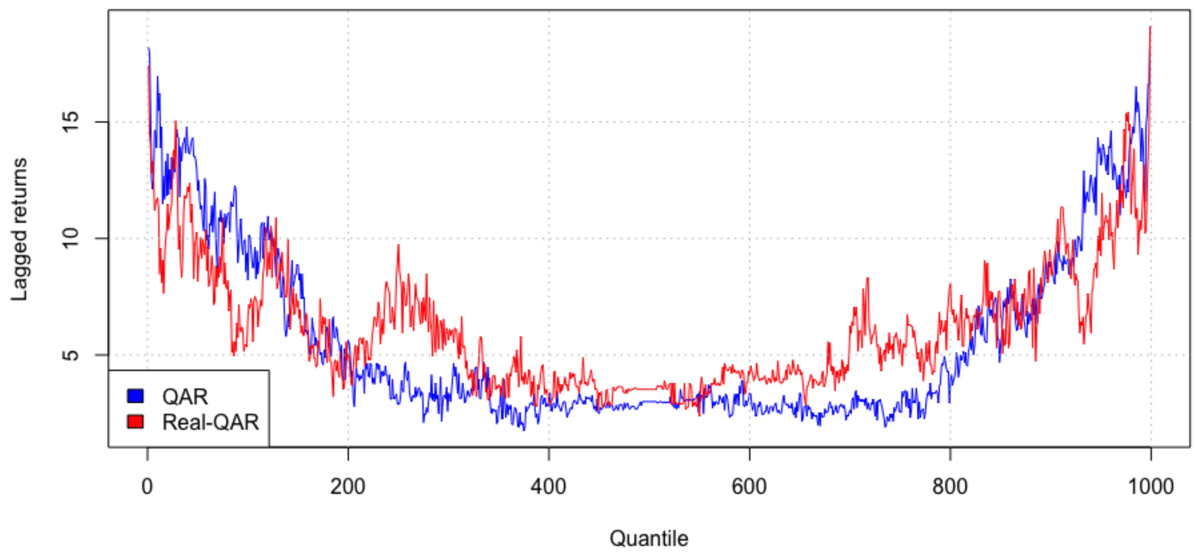


Figure 1. SIC lag selection

The average number of lags chosen for the nineteen stock indices for the QAR model (solid blue line) and for the Real-QAR model (solid red line) for different quantiles.

4.3 Asymmetric GARCH models

As a benchmark two GARCH models are included, which are known to capture the asymmetric response of lagged returns. The first GARCH model is the threshold-GARCH (TGARCH) model of Glosten et al. (1993), also known as the GJR-GARCH. Baur and Dimpfl (2019) allow the mean equation to have one AR term, which results in the following (T)GARCH equations:

$$r_t = \delta r_{t-1} + \epsilon_t, \tag{8}$$

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \gamma\epsilon_{t-1}I[\epsilon_{t-1} < 0] + \beta\sigma_{t-1}^2, \tag{9}$$

where I[$\cdot$] is an indicator function which has a value of 1 if $\epsilon_{t-1} < 0$ and 0 otherwise. To ensure stationarity, the restriction $1 - \alpha - \beta - \frac{\gamma}{2} \geq 0$ must hold. The conditions for a non-negative conditional variance are $\omega > 0$, $\alpha + \gamma/2 \geq 0$, and $\beta \geq 0$

The second GARCH model we include is the exponential-GARCH (EGARCH) model of Nelson (1991). The same mean equation is used as for the TGARCH (Equation (8)). The variance equation for the EGARCH models is as follows:

$$\ln(\sigma_t^2) = \omega + \alpha\left(\frac{\epsilon_{t-1}}{\sqrt{h_{t-1}}}\right) + \lambda\left[\frac{|\epsilon_{t-1}|}{\sqrt{h_{t-1}}} - \sqrt{\frac{2}{\pi}}\right]. \tag{10}$$

The main difference between the EGARCH and the TGARCH model is reflected in the variance equation, whereas the EGARCH model does not include the non-negative parameter restrictions. Since we know from the stylized facts of financial returns that returns are not normally distributed, we specify the ϵ_t in both models to follow Student's t-distribution which has the following density function:

$$p(y_t|\mu_t, \sigma_t, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1 + \frac{y_t^2}{\nu})^{-(\nu+1)/2}, \tag{11}$$

where $\Gamma(\cdot)$ denotes the gamma distribution, and ν corresponds to the degrees of freedom.

The focus of this thesis is on the one-day-ahead predicted densities. For the GARCH models, this thesis follows Hoogerheide et al. (2012) and Degiannakis and Potamia (2017), so that when the error terms are specified to follow the student's t-distribution the density forecast of the log returns can be expressed as:

$$y_{t+1}|y_t \sim \mathcal{S}\left(\mu_t, \sigma_{t+1}^2\frac{\nu - 2}{\nu}, \nu\right), \tag{12}$$

where, μ_t equals the time-varying location parameter (mean) in Equation (8), σ_{t+1}^2 equals the time-varying conditional variance, and ν corresponds to the degrees of freedom. For the GARCH models the parameters $(\mu_t, \sigma^2_{t+1}, \nu)$ are estimated within the GARCH equation, where $\nu > 2$ to guarantee the existence of the conditional variance.

4.4 Conditional density evaluation

When evaluating the conditional density forecasts, three methods are used. The first evaluation method is the probability integral transform (PIT) suggested by Diebold et al. (1998). Let y_t be a time series of (log) return observations. These observations are from an unknown data generating process which is characterized by the density series $\{f_t(y_t|\mathcal{I}_{t-1})\}_{t=1}^n$, where $\mathcal{I}_{t-1}$ denotes the information set available at time t. The constructed sequence of one-day-ahead (conditional) density forecasts are defined as $\{\hat{f}_t(y_t|\mathcal{I}_{t-1})\}_{t=1}^n$. When the conditional density forecasts are perfect, it should be that: $\{\hat{f}_t(y_t|\mathcal{I}_{t-1})\}_{t=1}^n = \{f_t(y_t|\mathcal{I}_{t-1})\}_{t=1}^n$. Testing whether this statement is true seems difficult since we cannot observe the true data generating process. The properties of PIT provide the solutions for this problem (Bauwens et al., 2004). The probability integral transform z_t is defined as the cumulative density function corresponding to the density $\hat{F}_t(y_t)$ evaluated at y_t:

$$z_t = \int_{-\infty}^{y_t} \hat{f}_t(s)\, ds = \hat{F}_t(y_t). \tag{13}$$

Diebold et al. (1998) stated that under the null hypothesis the PIT sequence $\{z_t\}_{n=1}^n$ should be independent and identically distributed (iid) as a uniform distribution on the interval (0,1) ($\{z_t\}_{n=1}^n \overset{iid}{\sim} U(0,1)$). Therefore, histograms could be used to check uniformity of this series. However, this thesis uses more formal testing by including the Pearson's Chi-squared test for goodness-of-fit as suggested by Diebold et al. (1998). The test statistic is:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{14}$$

where n is the number of bins used, E_i and O_i are the expected and observed frequencies in bin i respectively. When this test statistic is higher than the critical value from the chi-squared distribution with $n-1$ degrees of freedom and the 5% confidence level, the null hypothesis is rejected. The bins represent the bars of the histogram. This thesis uses five equally sized bins ($n = 5$) following Mitchell and Wallis (2011), Ravazzolo and Vahey (2014), and Cobb (2020).

As a second evaluation measure, the conditional density forecasts are evaluated by means two of scoring functions suggested by Gneiting and Raftery (2007). Firstly, the one-day-ahead density forecasts are evaluated by the logarithmic scoring function. Here, the expected logarithmic score ($\ln(S)$), the average of the sample logarithmic scores, is calculated for evaluation.

$$\ln(S) = -\frac{1}{T}\sum_{n=1}^{T}\ln(\hat{f}_t(y_t)) \tag{15}$$

Here, $\hat{f}_t(\cdot)$ is the one-day-ahead density forecast and y_t is the realized return for day t.

Secondly, the one-day-ahead density forecasts are evaluated by the mean of continuous ranked probability score (CRPS) statistics. The CRPS is a quadratic measure for the average absolute deviation between the empirical cumulative distribution function (CDF) of y_t and the forecasted CDF (Gneiting and Raftery, 2007):

$$\mathsf{CRPS}(F, y) = \frac{1}{T}\sum_{n=1}^{T}\int_{-\infty}^{\infty}(\hat{F}_t(s) - \mathbb{I}\{y_t \leq s\})^2)\,ds. \tag{16}$$

Here, $\hat{F}_t$ is the predicted CDF and $\mathbb{I}(\cdot)$ is an indicator function which has a value of 1 when $y_t \leq s$. An advantage of the CRPS is that it is less sensitive to outliers and better accounts a model when the values from the conditional density forecast are near to the realization (Gneiting and Raftery, 2007). For both scoring functions, it implies that the corresponding model with a higher score has less accurate density forecasts.

To test whether the density forecast significantly outperforms another model, a Diebold and Mariano (1995) test is used as suggested by Gneiting and Katzfuss (2014). The test is implemented as a two-sided test with the null hypothesis that the two models produce equal forecasts, while the alternative hypothesis is that the two models have different forecast accuracy. The QAR model by Baur and Dimpfl (2019) is set as the benchmark model for this purpose. For the CRPS, the test is implemented with a squared error loss function as this is a quadratic measure.

4.5 Value-at-Risk

The first evaluated indirect forecast measure of financial risk is the Value-at-Risk (VaR). The one-day-ahead VaR ($VaR_{\tau,t}(1-\tau, 1)$) is defined as the τ-th quantile of the conditional distribution of the daily return y_{t+1}:

$$P[y_{t+1} \leq VaR_{\tau,t}(1-\tau,1)|\mathcal{I}_t], = \hat{F}_{t+1|t}(VaR_{\tau,t}(1-\tau,1)) = \tau \tag{17}$$

where $\mathcal{I}_t$ is the information given on day t. To evaluate the VaR performance, the same tests are done as Baur and Dimpfl (2019) except for the DQ test proposed by Engle and Manganelli (2004). This test statistic did not seem to have any added value in their paper since it was not able to highlight any differences between the models. Therefore, this thesis only includes the unconditional and conditional coverage test proposed by Kupiec (1995) and Christoffersen (1998), respectively and the test for independence.

4.6 Expected Shortfall

As a second indirect forecast measure of financial risk, the expected shortfall (ES) is evaluated. The one-day-ahead ES $\mathsf{ES}_{\tau,t}$ is defined as the average loss given that the VaR is violated. So, the average loss when the loss is greater than the τ-th quantile of the conditional distribution of the daily return y_{t+1} (i.e. $\mathsf{VaR}_{\tau,t}$):

$$\mathsf{ES}_{\tau}(y_{t+1}|\mathcal{I}_t) = \mathbb{E}[y_{t+1}|y_{t+1} < \mathsf{VaR}_{\tau,t}(1-\tau, 1)], \tag{18}$$

where $\mathcal{I}_t$ is the information given on day t. The Expected Shortfall thus describes the size of the expected loss given a violation of VaR and is, consequently, a measure of the tail risk.

To check whether the ES is correctly specified, the ES is backtested by the regression-based tests (ESR Backtest) introduced by Bayer and Dimitriadis (2020). These tests are special since they include two tests (of the three) that in addition to the true returns only require ES forecasts, which is in contrast to all other existing ES backtests. This thesis implements the (two-sided) Auxiliary and the one-sided Intercept ESR Backtest, where the Auxiliary test also includes the VaR forecasts and the latter only the ES forecasts. The tests follow the classic backtesting idea of Mincer and Zarnowitz (1969), such that the daily returns (y_t) are regressed on an intercept term and the ES forecasts $\hat{e}_t$ for $t \in \{1, \dots, T\}$:

$$y_t = \beta_1 + \beta_2 \hat{e}_t + u_t^e, \tag{19}$$

where (β_1 , β_2) should equal (0,1) when the ES forecasts are correctly specified since it holds that $\hat{e}_t = \mathsf{ES}_{\tau}(Y_t|\mathcal{I}_{t-1})$ almost surely (Bayer and Dimitriadis, 2020). However, they state from earlier literature that estimating the β parameters individually is impossible since there are no strictly consistent identification and loss functions for the functional expected shortfall. Therefore, they suggest doing a joint regression from which they are able to estimate and test the β parameters. For the Auxiliary test, they use the following system to test $\mathbb{H}_0 : (\beta_1, \beta_2) = (0, 1)$ against $\mathbb{H}_a : (\beta_1, \beta_2) \neq (0, 1)$:

$$y_t = \alpha_1 + \alpha_2 v_t + u_t^q \quad \text{and} \quad y_t = \beta_1 + \beta_2 \hat{e}_t + u_t^e, \tag{20}$$

where ν_t is the vector of predicted VaR forecasts of the corresponding risk level. For the (one-sided) Intercept test the regression Equations (20) are replaced by:

$$y_t - \hat{e}_t = \alpha_1 + \alpha_2 \hat{e}_t + u_t^q \quad \text{and} \quad y_t - \hat{e}_t = \beta_1 + u_t^e, \tag{21}$$

to test $\mathbb{H}_0 : \beta_1 \geq 0$ against $\mathbb{H}_a : \beta_1 \leq 0$. The derivation of the ESR Backtests can be found in the paper of Bayer and Dimitriadis (2020).

The idea behind the one-sided Intercept ESR Backtest is that the regulators should prevent and fine the underestimation of financial risks. The Basel Committee (2016) already used one-sided backtests for example for the hit ratios of VaR forecasts. Hence, Bayer and Dimitriadis (2020) proposed a backtest for ES forecasts that allows for one-sided testing.

Additionally, the Expected Shortfall of each model is evaluated by means of the loss function of Fissler and Ziegel (2016). This loss function allows for a joint evaluation of the VaR and ES, where a lower statistic is preferred. The loss function is defined as follows:

$$L_{FZ_t}(y_t, VaR_{\tau,t}, ES_{\tau,t}; \tau) = -\frac{1}{\tau ES_{\tau,t}} \cdot \mathbb{1}\{y_t \leq VaR_{\tau,t}\}(VaR_{\tau,t} - y_t) + \frac{VaR_{\tau,t}}{ES_{\tau,t}} + \log(-ES_{\tau,t}) - 1. \tag{22}$$

Consequently, each model is evaluated by taking the mean of the forecast period. To test whether the test statistic is significantly better than the statistic obtained from the QAR model, a two-sided Diebold-Mariano test in included.

5 Results

As stated in Section 4, this thesis estimates 999 quantiles (from 0.001 up to 0.999) instead of only 99 quantiles (0.01 up to 99) since this creates even better predictions of downside financial risk regarding the criteria in this section. The section proceeds as follows. Firstly, the density forecasts are evaluated and secondly, the Value-at-Risk and Expected Shortfall forecasts are evaluated.

5.1 Conditional density forecasts

When comparing the four models based on their predictive density performance in Table 2, it is shown that the normal QAR model is outperformed by the other three models based on the CRPS statistic as nearly all statistics are significantly better by the other models. The logarithmic score ($Ln(S)$) shows a similar result since the QAR model is outperformed more than half of the cases by the EGARCH and the TGARCH model and for a quarter of the cases by the Real-QAR model. Based on the CRPS, the QAR model seems to be outperformed even when the difference in score is small. The reason for this is that a Diebold-Mariano test implemented with a squared error loss function is used since the CRPS a quadratic measure. There is no clear pattern presented when plotting the logarithmic score and CRPS statistics of the indices over the evaluation period. Therefore, it can not be concluded that one model is preferred during volatile periods when predicting the conditional density.

However, when we look at the Pearson's Chi-squared test for goodness-of-fit, the results show that the two GARCH models are outperformed by the normal QAR model. For the QAR model we do not reject the null hypothesis of the PIT sequence ($\{z_t\}_{n=1}^{n} \overset{iid}{\sim} U(0,1)$) for ten stock indices, while this is only for one stock index for both the GARCH models. This number of cases is low for the GARCH models since for most of the indices the first, second, and last bin of the PIT sequence is rather empty comparing to the other two bins, which indicates that the

Table 2 *Density forecast evaluation*

	QAR			Real-QAR			EGARCH			TGARCH		
Market	χ^2	$Ln(S)$	CRPS	χ^2	$Ln(S)$	CRPS	χ^2	$Ln(S)$	CRPS	χ^2	$Ln(S)$	CRPS
AEX	0.000	1.206	0.460	0.000	1.129*	0.450*	0.000	1.121*	0.446*	0.000	1.136*	0.449*
AORD	0.000	2.305	0.628	0.003	1.386	0.597*	0.032	1.363	0.595*	0.031	1.369	0.596*
BVSP	0.000	1.615	0.647	0.001	1.575	0.625*	0.001	1.467*	0.627*	0.001	1.467*	0.626*
DAX	0.043	1.367	0.540	0.418	1.303*	0.525	0.000	1.311*	0.531	0.000	1.321*	0.540
DJI	0.841	1.513	0.543	0.891	1.182	0.514*	0.000	1.237	0.523*	0.000	1.240	0.523*
ES50	0.064	1.578	0.641	0.130	1.502	0.621*	0.000	1.389*	0.619*	0.000	1.411*	0.627*
FCHI	0.004	1.377	0.513	0.030	1.269	0.496*	0.000	1.214*	0.497*	0.000	1.234*	0.501*
FTSE	0.027	1.667	0.691	0.056	1.608	0.670*	0.018	1.490*	0.664*	0.017	1.497*	0.666*
HSI	0.449	1.403	0.516	0.273	1.349	0.511	0.011	1.331	0.514	0.010	1.327	0.513
IBEX	0.003	1.611	0.541	0.007	1.348	0.523*	0.006	1.288	0.528*	0.006	1.290	0.529*
KS11	0.102	1.487	0.524	0.338	1.301	0.506*	0.001	1.265	0.514*	0.000	1.270	0.518
MXX	0.124	1.408	0.540	0.038	1.387*	0.537	0.028	1.371*	0.537	0.027	1.363*	0.535
Nasdaq	0.668	1.772	0.603	0.001	1.374	0.583*	0.000	1.418	0.591*	0.000	1.417	0.590*
Nikkei	0.000	1.203	0.459	0.032	1.085*	0.437*	0.000	1.085*	0.438*	0.000	1.100*	0.440*
NSEI	0.143	1.806	0.607	0.020	1.381	0.594*	0.000	1.426	0.600*	0.000	1.440	0.603
Russel	0.010	1.864	0.769	0.100	1.574	0.729*	0.127	1.593	0.736*	0.123	1.600	0.738*
S&P500	0.235	1.301	0.526	0.034	1.137*	0.495*	0.000	1.213*	0.508*	0.000	1.218*	0.509*
SSMI	0.663	1.493	0.505	0.649	1.155	0.487*	0.003	1.171	0.489*	0.003	1.179	0.493*
STI	0.157	1.112	0.409	0.590	1.126	0.392*	0.004	0.984	0.394*	0.003	0.985*	0.395*
	9	1.531	0.561	11	1.323	0.542	18	1.302	0.545	18	1.256	0.547

Note. The evaluation of the conditional density forecasts. The column corresponding to χ^2 denotes the p-value of the goodness-of-fit test, where a significance level of 5% is used for rejecting the null hypothesis. $Ln(s)$ and CRPS denote the logarithmic score and the continuous ranked probability score, respectively. An asterisk (*) for the Real-QAR, EGARCH and TGARCH indicates this score is significantly different from the normal QAR model tested by a two-sided Diebold-Mariano test. The last line denotes the summary of all tests. The columns corresponding to χ^2 show the number of times the null hypothesis of the Pearson's Chi-squared test for goodness-of-fit is rejected for each model. The columns corresponding to $Ln(s)$ and CRPS denote the mean of this statistic for each model.

PIT sequence is not independent and identically distributed (iid) as a uniform distribution on the interval (0,1).

Contrarily, the Real-QAR model shows similar results as the QAR model by Baur and Dimpfl (2019) as the null hypothesis is rejected eleven times. For this reason, the Real-QAR model is the most favourable model when comparing the performance based on their conditional density forecast. The general picture of the PIT sequences is displayed in Figure 2. Here, the PIT sequences of the four models for the Dow Jones Industrial are shown.

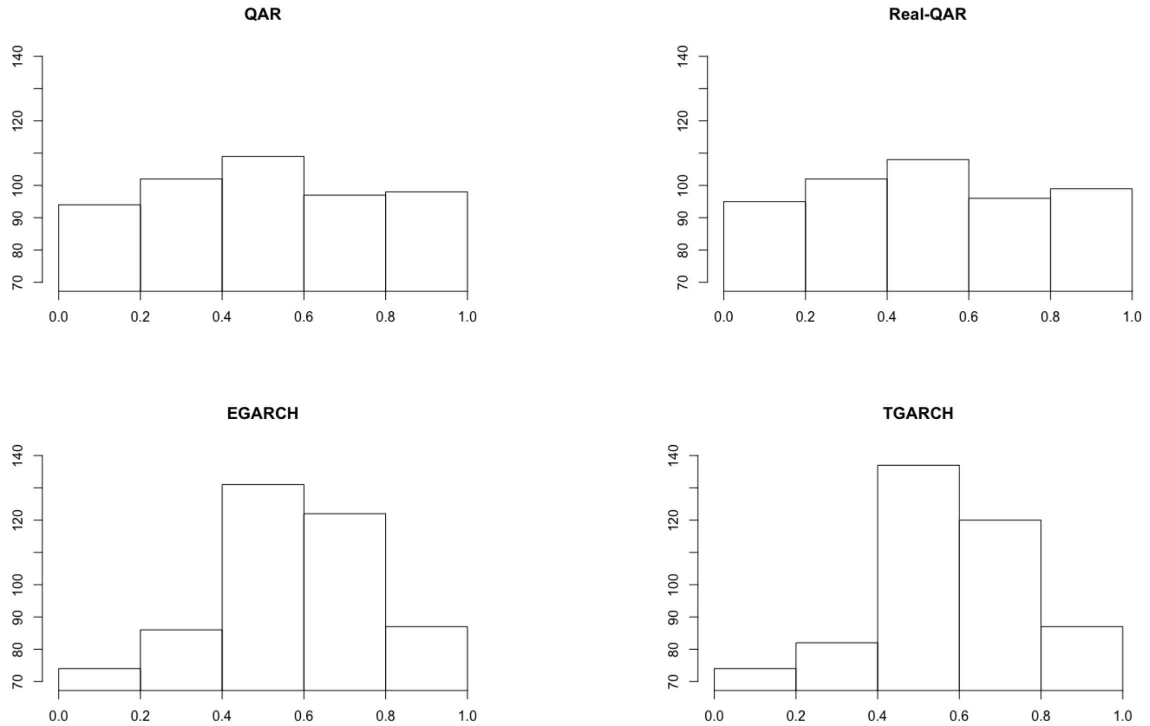


Figure 2. PIT histogram for DJI

This Figure shows the general image of PIT sequences. It can been seen that the QAR models seem independent and identically distributed (iid) as a uniform distribution on the interval (0,1), while this is not the case for the GARCH models as the first, second, and last bin of the PIT sequence is empty compared to the two bins.

5.2 Value-at-Risk forecasts

The results of the VaR forecasts are summarized on the left side of Table 3, while a more detailed evaluation is given in Table 4. When comparing the QAR models with each other, the results show similar results for the unconditional coverage test. The null hypothesis of correct fraction of violation is rejected ten and eleven times (out of 38) for the QAR and Real-QAR model respectively. Nevertheless, the results of the independence test vary between those models. The QAR model counts eight cases for which the null hypothesis of independent VaR violations is rejected, whereas the Real-QAR model only counts one case for which the null hypothesis is rejected. This results in a favourable position for the Real-QAR model when comparing those models with the conditional coverage test since this test statistic is a combination of the two aforementioned test statistics. Here, the QAR model counts eleven cases where the null hypothesis is rejected, while the Real-QAR model only counts seven of these cases. Both QAR models do not show clear differences in performance between the two risk levels.

The results in Table 3 and 4 for the GARCH models show that both models, EGARCH and TGARCH, are likely to underestimate the risk as the number of violations is most of the time larger than the target number for both risk levels. However, since the rate of violations is not that much above the target rate, those models perform slightly less than the QAR models. The null hypotheses for the unconditional and conditional coverage tests are rejected fifteen and fourteen times respectively for the EGARCH, while this is twelve and nine times for the TGARCH. For both models, it seems that, based on the unconditional coverage test, they are much better at predicting the rate of VaR violations for the 5% risk level than for the 1%. Both models only count three and two cases respectively in which the hypothesis of dependent VaR violation is not rejected. So, based on these results it can be concluded that the Real-QAR model is slightly better than the other models in forecasting Value-at-Risk.

Table 3 *Summary VaR and ES forecast evaluation*

		Value-at-Risk				Expected Shortfall		
Model	q	%	$\mathsf{LR}_{\mathsf{uc}}$	$\mathsf{LR}_{\mathsf{ind}}$	$\mathsf{LR}_{\mathsf{cc}}$	$\mathsf{P}_{\mathsf{aux}}$	$\mathsf{P}_{\mathsf{int}}$	FZL
QAR	1.0	1.25	6	3	5	4	2	1.57
	5.0	5.02	4	5	6	9	7	1.03
Real-QAR	1.0	1.46	6	0	4	2	3	1.28 (7)
	5.0	5.10	5	1	3	9	6	0.86 (8)
EGARCH	1.0	1.96	11	1	8	2	6	1.39 (4)
	5.0	6.21	4	2	6	3	2	0.85 (5)
TGARCH	1.0	1.91	11	0	5	2	2	1.35 (5)
	5.0	5.80	1	2	4	2	3	0.80 (7)

Note. The summary of all VaR and ES test statistics for all four models. The risk level is given by q. The average number of VaR violations is denoted by %. The columns $\mathsf{LR}_{\mathsf{uc}}$, $\mathsf{LR}_{\mathsf{ind}}$, $\mathsf{LR}_{\mathsf{cc}}$, $\mathsf{P}_{\mathsf{aux}}$, and $\mathsf{P}_{\mathsf{int}}$ denote the number of times the null hypothesis if the corresponding test is rejected for each model and risk level. The FZL column gives the average value of the FZ Loss statistic. For the Real-QAR and the two GARCH models, the number in parantheses indicates the amount of times the FZ Loss statistic is significantly better than the QAR model.

Figure 3 shows the 5% VaR forecasts of the four models during 2020. Here, it is shown that VaR forecasts by the Real-QAR model (dotted blue line) react stronger after a day of high volatility than the normal QAR model (solid red line) since the VaR line contains more and bigger spikes. This could be explained by the additional realized variance term included in the Real-QAR model. As this model reacts stronger than the other models after a highly volatile day, the VaR forecasts are less smooth than in particular the QAR model proposed by Baur and

Dimpfl (2019). The Figure shows the general image of all stock indices as the pattern is similar.

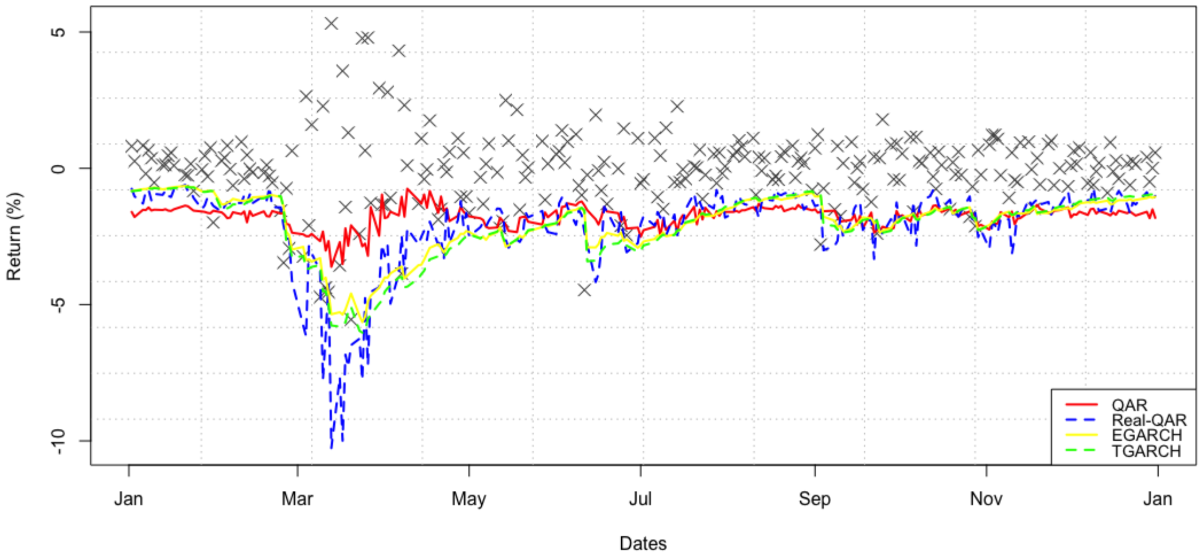


Figure 3. 5% VaR forecasts for the DAX index during 2020

This Figure shows the general image of the VaR forecasts at a 5% risk level. It can been seen that the Real-QAR model is less smooth than the QAR model since it reacts stronger after a day of high volatility due to the additional realized variance term.

5.3 Expected Shortfall forecasts

The results of the ES forecasts are summarized on the right side of Table 3, while a more detailed evaluation is given in Table 5. The results show that the normal QAR model is slightly outperformed by the Real-QAR model based on the Auxiliary and the one-sided Intercept test. The QAR model shows thirteen cases for which the ES is misspecified based on the Auxiliary test and for nine cases based on the Intercept test, while the Real-QAR model shows eleven and nine cases, respectively. Thereby, both QAR models perform significantly worse at the 5% risk level.

Based on the regression-based backtests, it is shown that the GARCH models both outperform the two QAR models. The forecasted ES of the EGARCH model is five and eight times misspecified based on the Auxiliary and Intercept test, respectively. The ES forecasted by the TGARCH model is misspecified only four and five times for the Auxiliary and Intercept test, respectively. So, based on ESR Backtest by Bayer and Dimitriadis (2020), it can be concluded that the TGARCH model is best at predicting the ES solely compared to the other three models.

Additionally, the Expected Shortfall and Value-at-Risk are jointly evaluated by means of the loss function of Fissler and Ziegel (2016). These results are also shown in Tables 3 and 5. Here, it is shown that the normal QAR model is outperformed by the other three models as in almost all cases its score statistics are higher than the score statistics of the other models. The Real-QAR model has fifteen statistics which are significantly lower based on a two-sided Diebold-Mariano test, while the EGARCH and TGARCH models have nine and twelve significantly better statistics. This could be explained by the fact that the Real-QAR model has more favourable results for the VaR forecasts given in Table 4. So, based on this loss statistic it can be concluded that one should favour the Real-QAR model to forecast the VaR and ES jointly since this model has the most significantly better scores.

6 Conclusion

This thesis extends on the findings of Baur and Dimpfl (2019) by introducing a new quantile autoregressive model, the Real-QAR. Thereby, it incorporates conditional density and expected shortfall evaluation, which has not been done before either. Especially looking into the (one-day-ahead) conditional densities seems interesting since this practice plays a key role in risk management nowadays. This is because popular measures of financial risk, such as the Value-at-Risk (VaR) and the Expected Shortfall (ES), are easily derived from the conditional density.

Therefore, the research question of this thesis is: *Which model is best at predicting the risk of stock returns*? This question is answered by looking into the predictive performance of three different aspects of risk: conditional density, VaR, and ES. The normal QAR model is clearly outperformed based on the predicted density and the ES, while it has similar results as the GARCH models for the VaR. The results of the conditional density and the VaR show that the Real-QAR should have the preference when predicting the risk of stock returns. The evaluation methods of the ES stand-alone tend to point out the TGARCH model as the best tail-risk predicting model. However, for the joint valuation of the VaR and ES the Real-QAR model has the best results in predicting the tail-risk of stock returns. All in all, it can be concluded that the new proposed Real-QAR model is best at predicting the risk of stock returns and should therefore be used by investors.

For further research, it could be interesting to look at the performance of combined density forecasts. Clemen (1989) summarizes the evidence in the previous literature that combining forecasts increases forecast accuracy. Thereby, Opschoor et al. (2017) argue that combined density forecasts are also able to significantly improve single density forecasts. This suggests that combining the Real-QAR model with other models could even provide better density forecasts and therefore also better VaR and ES forecasts.

Table 4 *Value-at-Risk forecast evaluation*

		Violation (%)				QAR			Real-QAR			EGARCH			TGARCH		
Market	q_{VaR}	Q	RQ	E	T	LR_{uc}	LR_{ind}	LR_{cc}	LR_{uc}	LR_{ind}	LR_{cc}	LR_{uc}	LR_{ind}	LR_{cc}	LR_{uc}	LR_{ind}	LR_{cc}
AEX	1.0	0.8	1.2	3.0	2.4	0.641	0.799	0.869	0.663	0.701	0.845	0.000	0.454	0.001	0.008	0.439	0.021
	5.0	3.8	3.4	7.2	7.0	0.199	0.743	0.417	0.082	0.594	0.193	0.034	0.113	0.034	0.052	0.089	0.041
AORD	1.0	4.2	3.2	3.6	3.0	0.000	0.267	0.000	0.000	0.784	0.000	0.000	0.144	0.000	0.000	0.065	0.000
	5.0	10.2	7.2	7.0	6.8	0.000	0.030	0.000	0.033	0.784	0.101	0.052	0.101	0.045	0.080	0.255	0.118
BVSP	1.0	1.0	0.4	0.4	0.4	0.999	0.750	0.951	0.125	0.901	0.311	0.125	0.899	0.306	0.125	0.899	0.306
	5.0	3.6	3.0	3.8	3.8	0.131	0.144	0.117	0.027	0.333	0.054	0.199	0.743	0.417	0.199	0.743	0.417
DAX	1.0	0.6	1.4	2.6	2.2	0.331	0.848	0.613	0.397	0.654	0.632	0.002	0.034	0.001	0.020	0.479	0.052
	5.0	3.4	4.4	5.4	5.2	0.082	0.113	0.067	0.530	0.974	0.821	0.685	0.606	0.813	0.838	0.571	0.842
DJI	1.0	1.0	0.6	1.8	1.6	0.999	0.750	0.950	0.331	0.847	0.612	0.106	0.563	0.230	0.215	0.608	0.407
	5.0	6.6	5.4	6.4	5.8	0.117	0.211	0.142	0.643	0.837	0.833	0.168	0.488	0.309	0.423	0.796	0.703
ES50	1.0	1.0	2.6	3.0	3.0	0.999	0.750	0.951	0.003	0.401	0.008	0.000	0.335	0.001	0.000	0.331	0.001
	5.0	5.2	5.0	6.2	5.2	0.838	0.175	0.416	0.999	0.141	0.365	0.235	0.139	0.180	0.838	0.038	0.134
FCHI	1.0	0.6	1.4	2.2	2.0	0.331	0.849	0.612	0.397	0.654	0.632	0.020	0.223	0.033	0.048	0.177	0.059
	5.0	3.4	2.6	5.6	5.2	0.082	0.269	0.121	0.007	0.329	0.016	0.546	0.013	0.049	0.838	0.000	0.031
FTSE	1.0	2.0	2.8	2.6	2.6	0.048	0.177	0.059	0.001	0.365	0.003	0.003	0.329	0.007	0.003	0.401	0.401
	5.0	6.4	5.6	5.8	5.4	0.168	0.046	0.063	0.546	0.598	0.730	0.430	0.308	0.448	0.685	0.643	0.833
HSI	1.0	0.8	1.4	1.2	1.2	0.641	0.799	0.869	0.396	0.654	0.632	0.663	0.701	0.845	0.663	0.701	0.845
	5.0	4.8	6.4	6.4	6.0	0.836	0.875	0.967	0.168	0.376	0.267	0.168	0.173	0.164	0.319	0.364	0.415
IBEX	1.0	0.2	0.8	2.0	2.2	0.028	0.949	0.090	0.641	0.799	0.869	0.048	0.179	0.059	0.020	0.223	0.033
	5.0	3.6	3.4	4.8	4.6	0.131	0.241	0.163	0.082	0.594	0.193	0.836	0.435	0.735	0.678	0.375	0.633

Continues on the next page

Table 4 *Value-at-Risk forecast evaluation (continued.)*

		Violation (%)				QAR			Real-QAR			EGARCH			TGARCH		
Market	q_{VaR}	Q	RQ	E	T	LR_{uc}	LR_{ind}	LR_{cc}	LR_{uc}	LR_{ind}	LR_{cc}	LR_{uc}	LR_{ind}	LR_{cc}	LR_{uc}	LR_{ind}	LR_{cc}
KS11	1.0	0.6	1.0	1.4	1.4	0.331	0.007	0.021	0.999	0.750	0.951	0.396	0.654	0.632	0.397	0.654	0.632
	5.0	3.2	4.1	7.0	6.8	0.049	0.087	0.036	0.399	0.897	0.695	0.052	0.706	0.142	0.079	0.630	0.192
MXX	1.0	0.0	0.2	0.4	0.8	0.001	–	0.007	0.028	0.949	0.090	0.125	0.899	0.306	0.641	0.799	0.869
	5.0	3.4	2.8	5.0	5.4	0.082	0.113	0.067	0.014	0.365	0.033	0.999	0.802	0.970	0.685	0.075	0.196
Nasdaq	1.0	1.4	0.8	2.6	2.8	0.397	0.654	0.632	0.641	0.799	0.689	0.003	0.401	0.008	0.001	0.365	0.002
	5.0	5.6	6.8	7.2	6.6	0.546	0.598	0.730	0.079	0.289	0.126	0.034	0.016	0.006	0.117	0.028	0.028
Nikkei	1.0	0.4	1.0	1.4	1.2	0.125	0.890	0.306	0.999	0.750	0.951	0.397	0.654	0.632	0.663	0.701	0.845
	5.0	3.2	4.0	6.4	5.0	0.048	0.000	0.001	0.288	0.191	0.247	0.168	0.172	0.164	0.999	0.141	0.365
NSEI	1.0	1.2	2.4	2.0	2.2	0.663	0.049	0.142	0.007	0.439	0.021	0.048	0.520	0.115	0.020	0.478	0.052
	5.0	5.4	8.8	8.0	8.0	0.685	0.214	0.450	0.000	0.597	0.001	0.004	0.895	0.017	0.004	0.419	0.013
Russel	1.0	2.8	2.0	2.0	2.2	0.001	0.390	0.002	0.048	0.176	0.059	0.048	0.520	0.115	0.020	0.479	0.052
	5.0	8.0	5.6	6.0	6.6	0.004	0.284	0.010	0.546	0.259	0.461	0.319	0.364	0.415	0.116	0.211	0.142
S&P500	1.0	1.6	1.4	1.6	1.6	0.215	0.608	0.407	0.397	0.654	0.632	0.215	0.608	0.407	0.215	0.608	0.407
	5.0	5.6	6.2	6.8	6.2	0.546	0.718	0.784	0.235	0.039	0.063	0.079	0.289	0.126	0.235	0.426	0.365
SSMI	1.0	2.0	1.6	2.2	2.0	0.048	0.009	0.006	0.215	0.608	0.407	0.020	0.479	0.052	0.048	0.520	0.115
	5.0	5.0	6.6	7.2	6.4	0.999	0.003	0.022	0.116	0.211	0.142	0.034	0.030	0.013	0.168	0.155	0.152
STI	1.0	1.6	1.6	1.2	1.4	0.215	0.608	0.407	0.215	0.608	0.407	0.663	0.701	0.845	0.397	0.654	0.632
	5.0	5.0	5.6	4.8	4.2	0.999	0.004	0.022	0.546	0.718	0.784	0.836	0.436	0.735	0.399	0.897	0.695

Note. The results of the VaR forecast evaluation based on the out-of-sample forecasts (T=500). The target value of violations is given by q_{VaR}. The p-values corresponding to the unconditional coverage test, the test for independence, and the conditional coverage test are given in columns 7-18. Q (RQ, T, E) represent the QAR (Real-QAR, TGARCH, EGARCH) model respectively. For the cases without violations, the test for independence is not possible. This is shown by a bar (–).

Table 5 *Expected Shortfall forecast evaluation*

		QAR			Real-QAR			EGARCH			TGARCH		
Market	q_{ES}	p_{aux}	p_{int}	FZL	p_{aux}	p_{int}	FZL	p_{aux}	p_{int}	FZL	p_{aux}	p_{int}	FZL
AEX	1.0	0.054	0.001	1.302	0.550	0.494	1.061	0.054	0.121	1.310	0.143	0.169	1.244
	5.0	0.716	0.258	0.777	0.526	0.234	0.640	0.020	0.012	0.763	0.088	0.056	0.727
AORD	1.0	0.001	0.334	2.971	0.005	0.006	1.552	0.017	0.233	1.682	0.057	0.409	1.565
	5.0	0.005	0.031	1.552	0.000	0.004	1.052*	0.007	0.008	0.997*	0.012	0.018	0.983*
BVSP	1.0	0.856	0.371	1.615	0.812	0.391	1.472	0.984	0.435	1.373	0.913	0.448	1.412
	5.0	0.772	0.200	1.200	0.783	0.437	0.932*	0.956	0.592	0.844*	0.894	0.641	0.849*
DAX	1.0	0.592	0.379	1.330	0.168	0.129	1.321	0.453	0.310	1.478	0.145	0.402	1.435
	5.0	0.942	0.313	0.943	0.157	0.414	0.847	0.122	0.010	0.904	0.118	0.160	0.897
DJI	1.0	0.793	0.448	1.689	0.930	0.406	0.967*	0.411	0.437	1.401	0.233	0.394	1.397
	5.0	0.024	0.244	1.030	0.032	0.016	0.746	0.150	0.156	0.825	0.224	0.321	0.807
ES50	1.0	0.549	0.462	1.561	0.349	0.311	1.527	0.077	0.191	1.883	0.133	0.095	1.753
	5.0	0.012	0.047	1.099	0.027	0.006	1.031	0.341	0.179	1.052	0.116	0.048	1.042
FCHI	1.0	0.976	0.349	1.352	0.794	0.345	1.143	0.255	0.152	1.416	0.451	0.418	1.325
	5.0	0.763	0.341	0.858	0.822	0.290	0.716	0.385	0.125	0.776	0.307	0.151	0.774
FTSE	1.0	0.040	0.085	1.674	0.127	0.365	1.593	0.007	0.000	1.699	0.030	0.021	1.731
	5.0	0.014	0.004	1.314	0.008	0.001	1.068*	0.308	0.113	1.039*	0.223	0.094	1.048*
HSI	1.0	0.842	0.635	1.141	0.300	0.378	1.089	0.907	0.591	1.008*	0.934	0.582	0.992*
	5.0	0.208	0.134	0.781	0.060	0.063	0.703	0.810	0.346	0.732	0.746	0.312	0.726
IBEX	1.0	0.014	0.761	1.248	0.693	0.619	1.047*	0.977	0.419	1.151	0.778	0.283	1.190
	5.0	0.241	0.405	0.868	0.977	0.593	0.733*	0.575	0.200	0.768	0.568	0.269	0.767
KS11	1.0	0.676	0.467	1.471	0.854	0.473	1.161*	0.647	0.357	1.136*	0.642	0.394	1.099*
	5.0	0.709	0.228	0.982	0.483	0.139	0.816	0.323	0.037	0.744*	0.483	0.072	0.750*
MXX	1.0	0.000	0.987	1.217	0.310	0.827	1.186	0.963	0.557	1.217	0.114	0.974	0.956*
	5.0	0.059	0.912	0.809	0.062	0.953	0.752*	0.261	0.332	0.796	0.419	0.753	0.700*
Nasdaq	1.0	0.332	0.405	1.810	0.940	0.585	1.052*	0.298	0.173	1.601	0.230	0.291	1.616
	5.0	0.033	0.044	1.204	0.000	0.024	0.956	0.024	0.025	1.060	0.020	0.034	1.055
Nikkei	1.0	0.129	0.385	1.162	0.649	0.455	1.008	0.973	0.619	0.837*	0.981	0.613	0.885*
	5.0	0.401	0.455	0.785	0.418	0.246	0.562	0.700	0.309	0.536	0.803	0.456	0.552*
NSEI	1.0	0.701	0.346	1.455	0.167	0.079	2.575	0.371	0.279	1.687	0.291	0.393	1.672
	5.0	0.020	0.013	1.191	0.193	0.097	1.661	0.223	0.192	1.085	0.072	0.284	1.098
Russel	1.0	0.207	0.167	2.066	0.092	0.021	1.431*	0.361	0.244	1.662	0.023	0.008	1.657
	5.0	0.000	0.000	1.474	0.041	0.032	1.071*	0.118	0.016	1.123*	0.145	0.054	1.132*
S&P500	1.0	0.842	0.308	1.944	0.609	0.289	1.091*	0.194	0.175	1.611	0.239	0.227	1.579
	5.0	0.032	0.222	1.087	0.005	0.181	0.817	0.281	0.154	0.879	0.234	0.227	0.866
SSMI	1.0	0.222	0.203	1.523	0.236	0.108	1.141*	0.894	0.436	1.235	0.801	0.358	1.183
	5.0	0.028	0.090	0.942	0.016	0.005	0.715*	0.211	0.097	0.785	0.235	0.073	0.734
STI	1.0	0.196	0.025	1.228	0.363	0.258	1.044	0.890	0.258	0.890*	0.672	0.319	0.898*
	5.0	0.068	0.009	0.749	0.159	0.065	0.501	0.777	0.359	0.403	0.806	0.392	0.404

Note. The results of the ES forecast evaluation based on out-of-sample forecasts (T=500). The risk level is given by q_{ES}. The p-values corresponding to the Auxiliary test and the Intercept test are given by p_{aux} and p_{int} respectively. Thereby, the joint VaR and ES evaluation using the FZL averages are shown, where an asterisk (*) indicates a significantly different value than the normal QAR model tested by a two-sided Diebold-Mariano test.

References

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003): "Modeling and forecasting realized volatility," *Econometrica*, 71, 579–625.

Basel Committee, B. (2016): "Implementing Basel III in Europe - European Banking Authority." *Handbook of Basel III Capital*, 11–38.

Bassett, G. and R. Koenker (1982): "An empirical quantile function for linear models with iid errors," *Journal of the American Statistical Association*, 77, 407–415.

Baur, D. G. and T. Dimpfl (2019): "A Quantile Regression Approach to Estimate the Variance of Financial Returns," *Journal of Financial Econometrics*, 17, 616–644.

Bauwens, L., P. Giot, J. Grammig, and D. Veredas (2004): "A comparison of financial duration models via density forecasts," *International Journal of Forecasting*, 20, 589–609.

Bayer, S. and T. Dimitriadis (2020): "Regression-Based Expected Shortfall Backtesting," *Journal of Financial Econometrics*, nbaa013.

Bollerslev, T. (1986): "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, 31, 307–327.

Christoffersen, P. F. (1998): "Evaluating interval forecasts," *International economic review*, 841–862.

Clemen, R. T. (1989): "Combining forecasts: A review and annotated bibliography," *International journal of forecasting*, 5, 559–583.

Cobb, M. P. (2020): "Aggregate density forecasting from disaggregate components using Bayesian VARs," *Empirical Economics*, 58, 287–312.

De Raaij, G. and B. Raunig (2005): "Evaluating density forecasts from models of stock market returns," *European Journal of Finance*, 11, 151–166.

Degiannakis, S. and A. Potamia (2017): "Multiple-days-ahead value-at-risk and expected shortfall forecasting for stock indices, commodities and exchange rates: Inter-day versus intraday data," *International Review of Financial Analysis*, 49, 176–190.

Diebold, F. and R. Mariano (1995): "Comparing predictive accuracy. Journal of Business and Economics Statistics, v. 13," .

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998): "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39, 863–883.

Du, Z. and J. C. Escanciano (2017): "Backtesting expected shortfall: accounting for tail risk," *Management Science*, 63, 940–958.

Engle, R. F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50, 987.

Engle, R. F. and S. Manganelli (2004): "CAViaR: Conditional autoregressive value at risk by regression quantiles," *Journal of business & economic statistics*, 22, 367–381.

Fissler, T. and J. F. Ziegel (2016): "Higher order elicitability and Osband's principle," *The Annals of Statistics*, 44.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993): "On the relation between the expected value and the volatility of the nominal excess return on stocks," *The journal of finance*, 48, 1779–1801.

Gneiting, T. and M. Katzfuss (2014): "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, 1, 125–151.

Gneiting, T. and A. E. Raftery (2007): "Strictly proper scoring rules, prediction, and estimation," *Journal of the American statistical Association*, 102, 359–378.

Hansen, P. R. and A. Lunde (2005): "A forecast comparison of volatility models: does anything beat a GARCH (1, 1)?" *Journal of applied econometrics*, 20, 873–889.

Hansen, P. R., A. Lunde, and J. M. Nason (2003): "Choosing the best volatility models: the model confidence set approach," *Oxford Bulletin of Economics and Statistics*, 65, 839–861.

Heber, G., A. Lunde, and N. Shephard (2009): "Oxford-Mann institute's realized library, Version 0.3," .

Hoogerheide, L. F., D. Ardia, and N. Corré (2012): "Stock index returns' density prediction using GARCH models: Frequentist or Bayesian estimation?" *Economics Letters*, 116, 322–325.

Hurvich, C. M. and C.-L. Tsai (1990): "Model selection for least absolute deviations regression in small samples," *Statistics & Probability Letters*, 9, 259–265.

Koenker, R. and Z. Xiao (2006): "Quantile autoregression," *Journal of the American statistical association*, 101, 980–990.

Kupiec, P. (1995): "Techniques for verifying the accuracy of risk measurement models," *The J. of Derivatives*, 3.

Mazur, B. (2017): "Density forecasting performance of alternative GARCH and DCS models for daily financial returns," .

Mincer, J. A. and V. Zarnowitz (1969): "The evaluation of economic forecasts," in *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, NBER, 3–46.

Mitchell, J. and K. F. Wallis (2011): "Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness," *Journal of Applied Econometrics*, 26, 1023–1040.

Nelson, D. B. (1991): "Conditional heteroskedasticity in asset returns: A new approach," *Econometrica: Journal of the Econometric Society*, 347–370.

Opschoor, A., D. Van Dijk, and M. van der Wel (2017): "Combining density forecasts using focused scoring rules," *Journal of Applied Econometrics*, 32, 1298–1313.

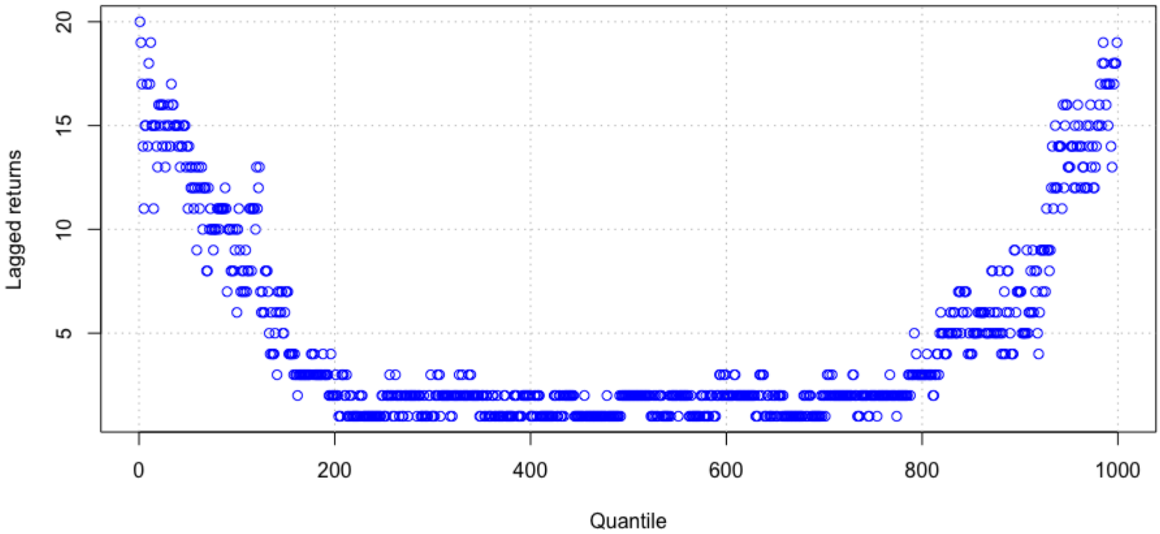
Portnoy, S., R. Koenker, et al. (1989): "Adaptive L-estimation for linear models," *Annals of Statistics*, 17, 362–381.

Ravazzolo, F. and S. P. Vahey (2014): "Forecast densities for economic aggregates from disaggregate ensembles," *Studies in Nonlinear Dynamics & Econometrics*, 18, 367–381.

Timmermann, A. (2000): "Density forecasting in economics and finance," *Journal of Forecasting*, 19, 231.

Wang, Y., F. Ma, Y. Wei, and C. Wu (2016): "Forecasting realized volatility in a changing world: A dynamic model averaging approach," *Journal of Banking & Finance*, 64, 136–149.

Appendix



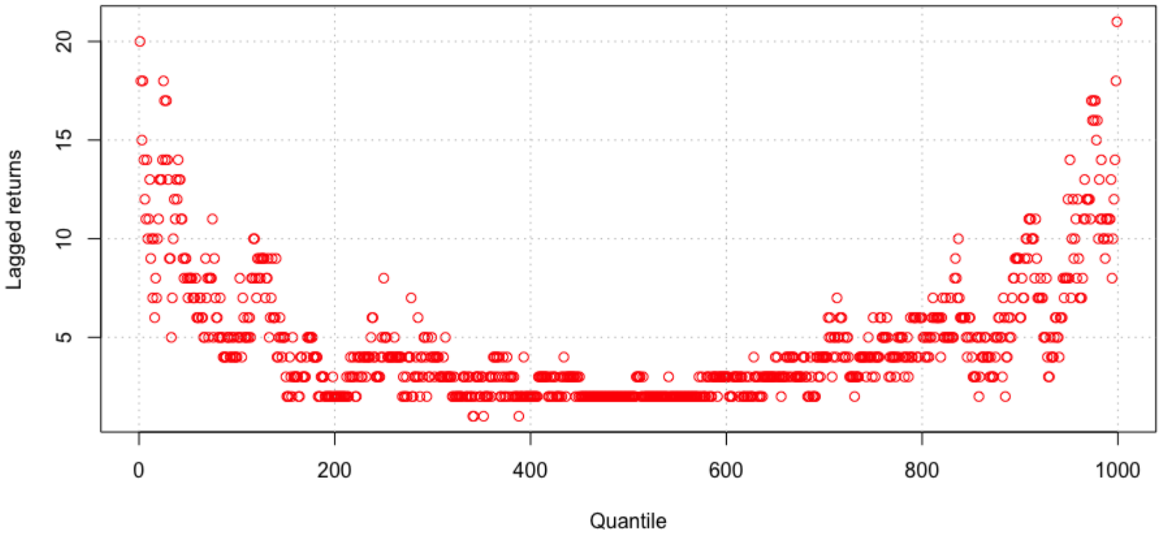


Figure 4. Median of SIC lag selection

The median of lags chosen for the nineteen stock indices for the QAR model (upper) and for the Real-QAR model (lower) for different quantiles.