University of Warsaw
Faculty of Economic Sciences

Working Papers
No. 11/2021 (359)

Comparison of the Accuracy in VaR Forecasting for Commodities Using Different Methods of Combining Forecasts

Szymon Lis
Marcin Chlebus

Warsaw 2021

WORKING PAPERS 11/2021 (359)

UNIVERSITY OF WARSAW
FACULTY OF ECONOMIC SCIENCES
WORKING PAPERS

Comparison of the accuracy in VaR forecasting for commodities using different methods of combining forecasts

Szymon Lis*, **Marcin Chlebus**

University of Warsaw, Faculty of Economic Sciences
** Corresponding author: sm.lis@student.uw.edu.pl*

Abstract: No model dominates existing VaR forecasting comparisons. This problem may be solved by combine forecasts. This study investigates the daily volatility forecasting for commodities (gold, silver, oil, gas, copper) from 2000-2020 and identifies the source of performance improvements between individual GARCH models and combining forecasts methods (mean, the lowest, the highest, CQOM, quantile regression with the elastic net or LASSO regularization, random forests, gradient boosting, neural network) through the MCS. Results indicate that individual models achieve more accurate VaR forecasts for the confidence level of 0.975, but combined forecasts are more precise for 0.99. In most cases simple combining methods (mean or the lowest VaR) are the best. Such evidence demonstrates that combining forecasts is important to get better results from the existing models. The study shows that combining the forecasts allows for more accurate VaR forecasting, although it's difficult to find accurate, complex methods.

Keywords: Combining forecasts, Econometric models, Finance, Financial markets, GARCH models, Neural networks, Regression, Time series, Risk, Value-at-Risk, Machine learning, Model Confidence Set

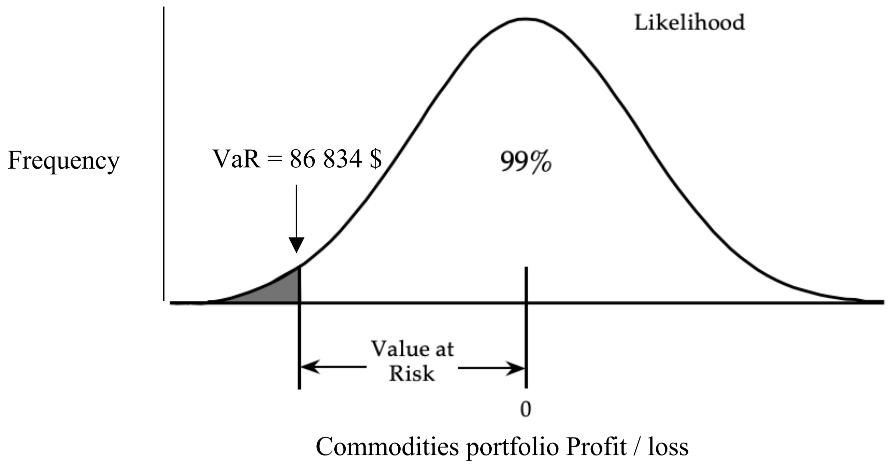
JEL codes: C51, C52, C53, G32, Q01

1. Introduction

At the end of XIX century the word risk hasn't got any technical definition in economics (Haynes 1895). In the 1920's and 1930's the beginnings of the portfolio theory can be found. Hardy (1923), Hicks (1935), and Leavens (1945) wrote papers on non-mathematical discussions of portfolio building, presenting the advantages of diversification. Then, in the following years, much was devoted to the optimization of the portfolio. This time quantitative methods were used (Roy (1952),Markowitz (1952), Sharpe (1964)). However, it was only since the stock market crash of 1987 that modelling and forecasting the volatility of financial markets has received much attention from academics, practitioners, and regulators. It started to play the central role in several financial applications (Liu and Hung 2010). In the following year - 1988 - the set of minimum capital requirements for banks was published by the Basel Committee on Banking Supervision (BCBS). Finally, in 1996 by introducing the market risk amendment new risk measures, generally based on 'Value at Risk' (VaR), were adopted (Goodhart 2011 p. 262). In 2004 The Basel II Capital Accord has established VaR as the official measure of market risk and placed it at the center of the determination of capital charges (Rossignolo, et al. 2012). In 2010, in response to the shortcomings of financial regulation revealed during the 2007–2008 financial crisis, another agreement was approved - Basel III. The BCBS (2010) found that risk modeling practices at the time were developed to manage operations under standard conditions with little financial stress, and performed poorly in times of crisis. Basel III introduced a new additional measure - Stressed VaR (sVaR), which requires the computation of an additional VaR measure based on annual data from a period of significant financial stress related to the portfolio. Since then, VaR is calculated as the sum of both of these values (Rossignolo et al. 2012). In 2017, the Basel Committee agreed to abandon VaR and move to the expected shortfall (ES), because there were a number of weaknesses identified with the use of VaR to determine regulatory capital requirements, including its inability to capture the 'tail of risk'. The committee agreed to use 97.5% ES. The agreement is expected to enter into force in 2023 (BCBS, 2017). However, it is still worth modeling VaR since it will be valid for a few more years, and the accuracy of the calculated ES depends on how correctly the VaR level is calculated. Furthermore, Danielsson (2013) analyzed that 97.5% ES and 99% VaR are the same. His work showed that the 99% VaR risk forecast are generally less volatile than 97.5% ES.

VaR is a statistical measure of possible portfolio losses at the specified time horizon (in Basel requirements 10-days for market risk level estimation, 1-day for backtesting) with a certain probability – confidence level (1%, 2.5% and 5%), resulting from price movements at the market. This means that losses greater than VaR are only incurred with a specified low probability (Linsmeier and Pearson 2000). This is illustrated in Figure 1 for example of 1% VaR (Formula for Value at Risk will be included in section "Method")

Figure 1. Graphical representation of the Value at Risk.



Source: based on Duffie and Pan (1997).

In addition, the mathematical equation that defines VaR can be represented as follows (Jorion 2010):

$$P(r_t < VaR_\alpha(t)|\Omega_{t-1} = \alpha \tag{1}$$

where:

r_t - financial return,

α - confidence level,

t - time interval,

Ω_{t-1} - information set available at time t-1.

VaR estimation methods can be divided into three groups:

1. Non-parametric methods, such as historical simulation or monte carlo simulation (Cabedo and Moya 2003).
2. Fully parametric methods, usually econometric models for volatility modelling, e.g. GARCH or GJR-GARCH (Bollerslev (1986); Glosten et al. (1993)).
3. Semi-parametric methods, where part of the model is estimated using parametric model, but the other part is non-parametric. The examples are: the Extreme Value Theory Peak-

over-Threshold model that uses a Generalized Pareto Distribution for a left tail of the return distribution starting from arbitrary chosen threshold (Marimoutou et al. 2009) or QML-GARCH, where GARCH model is used to model conditional volatility and then empirical distribution of standardised residuals are taken as a part of VaR calculation (Engle and Manganelli 1999).

A performance and reliability of models in accurately predicting VaR depend largely on data. While the parsimonious model may perform well in economically stable periods, it can fail in times of greater volatility in the markets (Angabini and Wasiuzzaman 2011). Likewise, highly parameterized models may be appropriate in times of high volatility, but not necessarily in times of "calm" (Laurent et al. (2012), Abad and Benito (2013)). So far, no unique model or approach dominates existing VaR forecasting comparisons (Kuester et al. (2006), Abad and Benito (2013), Bernardi and Catania (2016), Bayer (2018), Buczyński, and Chlebus (2018)). The solution to this problem may be to create more complex models that will better match current economic conditions or combining forecasts. However, as indicated above, despite the continuous development of models, there are no clear results which the model predicts better. However, there is a lot of evidence that predictions that combine models often outperform individual models (among others see Chiu et al. (2010), Jeon and Taylor (2013), Bayer (2018); Taylor (2020)). Some of the studies show that regardless of the asset, the models used, and the assessment period, the combination of forecasts generates more accurate forecast forecasts, which are located in the "green zone", in accordance with Basel II regulations (Halbleib and Pohlmeier 2012). On the other hand, older works on time series forecasting question the predictive increase in accuracy by combining forecasts (Makridakis and Winkler (1983), Armstrong (1989), Terui and Van Dijk (2002)). The last one advantage of combining forecasts is that combining forecasts instead of choosing an individual prognostic model reduces the modeling risk, i.e. the difference in Symmetric mean absolute percentage error (sMAPE) compared to the best possible model is smaller for combining forecasts rather than selecting an individual forecasting method (Hibon and Evgeniou 2005). For these reasons, I found that combining forecasts gives more promising results than creating more complex models.

What are the reasons to choose such an approach? In a review of a combination of forecasts, Timmermann (2006) presents three arguments for combining forecasts to stabilize and improve predictive results from standalone models. First, there are benefits from different model assumptions, specifications, or information sets. Second, the combined forecasts seem to be immune to structural breakdowns. Third, the impact of potential misspecification of individual models is reduced by combining a set of predictions from several models.

There are many methods for combining forecasts. Some of them have already been used in relation to VaR (Giacomini and Komunjer (2005), Huang and Lee (2013), McAleer et al. (2013)). However, usually one method of combining forecasts is compared with the mean, median and/or the results of individual methods. In the literature, there is no thorough comparison of the accuracy of forecasts made with the use of different combination methods for Value at Risk. In this paper, VaR forecasts for commodities prices (gold, silver, copper, oil and gas) using standalone methods (GARCH, GARCH-t, GARCH-st, QML-GARCH, and Indirect GARCH (CaViaR) and combined forecasts (average of forecasts, most conservative forecast, most liberal forecast, quantile regression with various loss functions, quantile random forests, generalized boosted regression model and quantile regression using neural network) was compared. Subsequently, their accuracy over a whole period (from mid 2004 – 2020), a period of calm (July 2004 – 2006, 2009 - 2013, and 2016 - 2019), a period of crisis (2007 – 2008, 2014 – 2015, March 2020 – December 2020), and a coronavirus pandemic period (March 2020 – December 2020).

According to the aim of the study following hypotheses were put forward in the study:

Hypothesis 1. "*Over the entire period, forecast combining methods will be more accurate than individual methods*." This is because there are three periods of crisis throughout the whole period in which individual methods often fail.

Hypothesis 2. "*In the period of calm, forecast combining methods will prove to be more accurate than the individual methods*". This is because these methods are still able to use the best features of each model by weighing them appropriately, and therefore combining predictions will produce more accurate results.

Hypothesis 3. "*In times of crisis, the methods of combining forecasts will turn out to be more accurate than individual methods*." The justification is the same as in the case of the hypothesis 3.

Hypothesis 4. "*During the period of data available for the current coronavirus pandemic, forecast combining methods will be more accurate than standalone VaR*" The rationale is the same as for the last two hypotheses.

In all hypotheses, when the greater accuracy of the forecast combining methods was meant, it was understood as supremacy of at least one of these methods, not all of them. As most of the methods are going to be used for the first time for combining Value at Risk forecasts, hypotheses regarding the primacy of one of the combination methods were not stated.

The remainder of this paper is organized as follows. Section 2 introduces the methodology and provides details on individual methods, ways of combining forecast and exploratory data analysis. Section 3 introduces the results of the empirical application. Section 4 consists of a conclusion and an outlook on potential future research areas.

2. Methodology

The methodology describes the scope of data, individual models, methods of combining forecasts, and methods of backtesting and comparing models.

2.1 Data

Data on futures prices for gold, silver, copper, oil and gas were collected for the period from 01/09/2000 to 01/12/2020 (6,215 records) from Yahoo Finance (COMEX Gold futures. (GC). (2020, December 7). YahooFinance. Retrieved December 7, 2020, from http://finance.yahoo.com/q?s=GC=F; COMEX Silver futures. (SI). (2020, December 7). Yahoo!Finance. Retrieved December 7, 2020, from http://finance.yahoo.com/q?s=SI=F; COMEX Copper futures. (HG). (2020, December 7). Yahoo!Finance. Retrieved December 7, 2020, from http://finance.yahoo.com/q?s=HG=F; NYMEX WTI Crude Oil futures. (CL). (2020, December 7). Yahoo!Finance. Retrieved December 7, 2020, from http://finance.yahoo.com/q?s=CL=F; NYMEX Gas futures. (NG). (2020, December 7). Yahoo!Finance. Retrieved December 7, 2020, from http://finance.yahoo.com/q?s=NG=F). The first three instruments are traded on the COMEX exchange, and the remaining two on the NYMEX exchange.

For each financial instrument, the number of non-available data is different - it ranges from 1079 to 1138. Definitely, most of them come from values on Sundays where there is no listing on the stock exchange. Thus, the observations for Sundays have been deleted. The rest of the NA value has been filled in with the last available value. Broad research has shown that this approach is as robust for time series data as other known data gap-filling methods (Caillault et al. 2017).

The log returns were calculated based on the adjusted price according to below formula:[1]

[1] April 20, 2020, was the first day in history for oil to record negative prices. Therefore, the log-return formula could not be used here, so that day was assigned with the minimum log-return from the rest of the period, and the following day with the maximum log-return from the rest of the period.

$$R_{log,t} = \ln(\frac{P_t}{P_{t-1}}) \tag{2}$$

where:

P_t – price of the asset in period t,

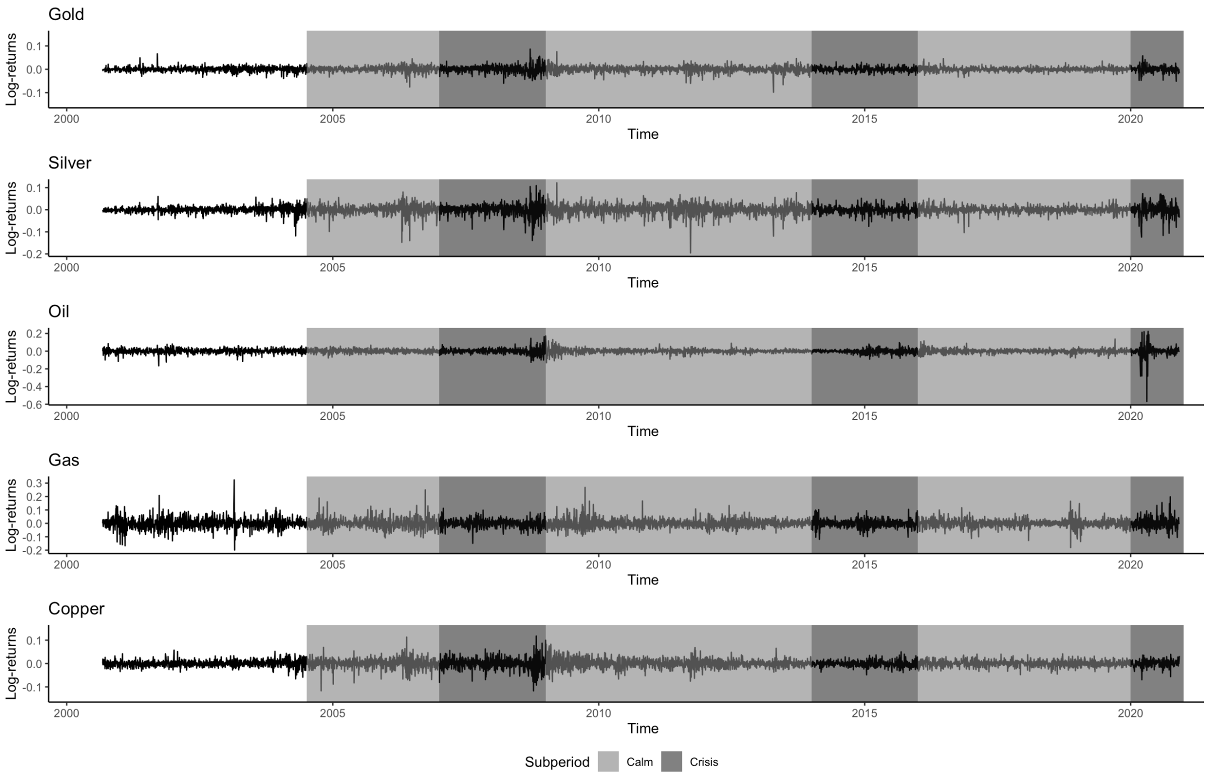
P_{t-1} - price of the asset in period t-1.

The VaR assessment horizon will cover the period from July 5, 2004, to December 1, 2020 (4157 days). One-day ahead VaR with 99^{th} and 97.5^{th} alpha level will be used. Estimation will be performed on the basis of data from the last 1000 observations and the model parameters will be updated with each observation (rolling window approach). It is a commonly used rolling window (among others Danielsson and Morimoto (2000), Bayer (2018)). Moreover, the research showed that for the window sizes of 500 and 2000 observations, the predicted VaR does not differ significantly from the results for window size of 1000 observations (Gençay et al. 2003). Forecasts will be divided into two sub-periods – the calm period and the crisis period. The latter will be much more volatile compared to the former[2]. The purpose of this apportionment is to evaluate models for situations with different volatility. The division of the assessment period is shown in Figure 2, which shows the logarithmic returns for each of the commodities. The dark grey areas represent the three periods of crisis. First from January 1, 2007, to December 31, 2008. This period covers the 2007-2009 financial crisis. After property prices in the US dropped drastically, and banks faced many problems and people's confidence in them declined, investors began to invest more in commodities. This resulted in a large increase in commodities prices in 2007-2008 (Phillips and Yu 2011). The second period of the crisis, from 1 January 2014 to 31 December 2015, for most prices (except for copper) meant much more frequent drops than in the previous period. They were not always significant, although this period is interesting due to the fact that declines can be recorded much more clearly in this period. The main reason for the decline in commodity prices is the surplus of supply in relation to demand, which originated in the commodity boom at the end of the previous decade, the slowdown in the development of the world economy, including China and several other large developing countries, as well as the boom on the stock market (Dudziński 2016). The last, third period of the crisis marks the beginning of the coronavirus pandemic, i.e. the period from January 1, 2020, to December 1, 2020. Global uncertainty related to the

[2] High volatility also occurs during calm periods. The division results not only from the graphical analysis of the charts but also from considering the periods distinguished in the literature as crises. In the study, it was decided to distinguish the most significant ones and common for every commodity.

emergence of a new disease, COVID-19, has significantly disrupted the dynamics of prices of all raw materials. With the exception of gold prices, this was a negative effect. It turns out that information about the pandemic around the world also changes investors' decisions. Icheck and Marinc (2018) argue that the Ebola virus in 2014–2016, combined with high media coverage, significantly influenced investors' strategies, including by lowering the share prices of companies operating in Africa (Mensi et al. 2020). Time remaining (light gray areas) means a calm period, i.e. from July 5, 2004, to December 31, 2006, from January 1, 2009, to December 31, 2013, and from January 1, 2016, to December 31, 2019. Both areas together represent the entire horizon of the assessment. Unfortunately, for gas throughout the entire VaR testing period, the variability is consistently high, and for copper, it is stably low, which may result in a smaller benefit from the use of the proposed forecasts combining methods.

Figure 2. Logarithmic returns for each commodity (light grey for clam period, dark grey for crisis period).



Source: Own calculations.

2.2 Standalone models

This section describes individual methods for VaR forecasting. The description consists of the history of the model, the mathematical formula, and the advantages and drawbacks of each approach.

2.2.1 GARCH family

The GARCH process (Generalized Autoregressive Conditional Heteroskedasticity) was proposed by Bollerslev (1986). This is a generalization of the ARCH model created by Engle (1982). In GARCH model the conditional variance is not only the function of lagged random errors, but also of lagged conditional variances. Let Z_t be a sequence of i.i.d. random variables such that $Z_t \sim$ N(0,1) and $\epsilon_t \sim$ N(0, σ_t^2). Then standard GARCH model (p,q) will be written as:

$$r_t = \mu_t + \epsilon_t \ = \mu_t + Z_t * \sigma_t \quad \mathrm{t} \in \mathrm{Z} \tag{3}$$

where σ_t is a nonnegative process such that

$$\sigma_t^2 = \ \omega + \alpha_1 * X_{t-1}^2 + \cdots \ \alpha_q * X_{t-q}^2 + \beta_1 * \sigma_{t-1}^2 + \cdots + \beta_1 * \sigma_{t-p}^2, \mathrm{t} \in \mathrm{Z}\ (2.2) \tag{4}$$

and

$$\omega > 0, \alpha_i \geq 0, \mathrm{i} = 1,\ldots,\mathrm{q}, \beta_i \geq 0 \ i = 1, \ldots, p \tag{5}$$

Normality assumption could not be realistic for many cases in real-life problems (Holthausen and Hughes (1978), Szakmary et al. (2010), Youssef et al. (2015)) Therefore, researchers are interested to construct more flexible distributions as alternative to normal distribution to model both skewness and kurtosis. The GARCH-t model of Bollerslev (1987) assumes that errors follow a standardized Student t-distribution and GARCH-st that they have a skewed t-Student distribution.

2.2.2 QML-GARCH

The QML-GARCH (Quasi-Maximum Likelihood GARCH) model is based on the work of Bollerslev and Woolridge (1992). They proved that in the GARCH model estimators are consistent even if the random errors don't come from the normal distribution. Therefore, the use of the GARCH process to standardize the residuals that are not derived from the normal distribution is still appropriate. Based on this, Engle and Manganelli (1999) proposed the QML-GARCH model. It uses the GARCH model to estimate a conditional variance, and then estimates the VaR value as the empirical distribution quantile of the standardized residuals of this model. This is a combination of the GARCH model with historical simulation for a series of standardized residuals.

2.2.3 CaViar

In 2004 Engle and Manganelli proposed the CAViaR model (Conditional Autoregressive Value-at-Risk), that completely refuse to model of the rates of return distribution and directly models the distribution of the quantile. The concept is based on the stylized fact that there is a high autocorrelation in the variance of financial series. Value at Risk is strongly related to variance, so the autocorrelation in VaR should be also present. The basic specification of the CAViaR model is as follows:

$$VaR_\alpha(t) = \beta_0 + \sum_{i=1}^{q} \beta_i * VaR_\alpha * (t-1) + \sum_{j=1}^{r} \beta_j * l(r_{t-j}) \qquad (6)$$

where

$VaR_\alpha(t)$ - VaR at the α level in the t period,

β_0 - model constant,

$\beta_i, \dots, \beta_q$ - weights of the lagged VaRs,

$\beta_j, \dots, \beta_r$ - weights of lagged rates of return,

$l(r_{t-j})$ - function of a finite number of rates of return

For all models, parameters were selected based on the lowest AIC, which is one of the most commonly used indicators for selecting models (Tsay, 2005). All models met the following conditions:

1) Sum of the parameters was lower than 1,
2) All of the parameters were statistically significant,
3) Ljung-Box test on standardized squared residuals indicated that standardized residuals were white noise (ARCH effect removal),
4) The p-value of the LM ARCH test indicated no ARCH effects among the residuals of the model.

When the autocorrelation in the residuals was observed, then it was eliminated by adding the ARMA process to the GARCH, creating ARMA-GARCH.

2.3 Combined forecasts

This section will describe the methods of forecasting. It includes the history of a given method, its use to combine VaR forecasts or other rare events, a mathematical formula or algorithm in the case of machine learning methods, and the advantages and disadvantages of the proposed approaches.

2.3.1 Mean of forecast

In previous studies, researchers in order to combine forecasts have often used the mean of all the forecasts obtained (Halbleib and Pohlmeier (2012), Huang and Lee (2013)). Due to its simplicity and clarity, the simple average found recognition in non-economic literature and actually performed well (Clemen and Winkler (1986), Timmermann (2006)). However, it also carries some dangers. If the values are correlated, then combining them with the mean is capturing the same information again, which is a mere bias-variance trade-off (Hastie, Tibshirani and Friedman 2011, p. 223). Therefore, the increase in the variance, e.g. by collinearity of covariates increases the expected square error of the prediction. Thus, it's better to look at the correlations between forecast and choose the least correlated ones. Therefore, the model with the best backtesting results in the in-sample period will be combined with the forecast from the other model with which it has the lowest correlation.

2.3.2 The lowest VaR

It is interesting to check the most conservative VaR, so to calculate minimum VaR from standalone models for each time. The advantage of this approach will be to always choose the most conservative forecast. If all models gave good results during the calm period, and during the crisis, only the most conservative gave, the use of this approach would be appropriate, because such a solution automatically chooses the most conservative VaR. On the other hand, a drawback may come from constantly overestimating VaR, even during the period, which is also not desired. The approach is described in the following formula:

$$VaR_{min,t} = \min\left(VaR_t^1, VaR_t^2, \dots, VaR_t^n\right) \quad (7)$$

Where:

VaR_t^n – VaR forecast from model n for the period t.

This method has been applied by McAleer et al. (2010), and Buczyński and Chlebus (2019).

2.3.3 The highest VaR

Another interesting combination is to check the most liberal VaR, so to calculate maximum VaR from individual models for each time. If all measures turned out to be overly conservative regardless of time, then it would be a good solution to use the most liberal measure, i.e. the maximum of all predictions. At the same time, such an approach poses a risk of a constant underestimation of VaR. The approach is described in the following formula:

$$VaR_{max,t} = \max\left(VaR_t^1, VaR_t^2, \dots, VaR_t^n\right) \tag{8}$$

Where:

VaR_t^n – VaR forecast from model n for the period t.

This method has been applied by McAleer et al. (2010), and Buczyński and Chlebus (2019).

2.3.4 Conditional quantile optimization method

This method models a conditional p-order quantile using a linear combination of two known quartiles determined using individual methods:

$$Q_p(r_{T+s}) = \beta_{T,0} + \beta * VaR_{T+s}^1 + \left(1 - \beta_{T,1}\right) * VaR_{T+s}^2 \equiv VaR_{12,T+s} * \beta_T = 1,2,\dots,S \tag{9}$$

Where:

$VaR_{12,T+s} = (1, VaR_{T+s}^1, VaR_{T+s}^2)$ is a vector of VaRs predicted from standalone models (1 is here to capture the intercept),

$\beta_T = (\beta_{T,0}, \beta_{T,1}, \left(1 - \beta_{T,1}\right))$ is a vector of parameters for VaRs predicted from standalone models,

The vector of optimal weights λ_T is determined by solving the following minimization problem:

$$\hat{\lambda}_T = \underset{\lambda_T}{\operatorname{argmin}} \begin{matrix} \{\sum_{r_T \geq VaR_{12,T} * \lambda_T} p * \left|r_T - VaR_{12,T} * \beta_T\right| + \\ \sum_{r_T < VaR_{12,T} * \lambda_T} (1-p) * \left|r_T - VaR_{12,T} * \beta_T\right|\} \end{matrix} \tag{10}$$

Where:

r_T – returns for a particular commodity in time T.

The main advantage of the quantile regression approach is that it does not require explicit distribution assumptions for return data. The same forecasts as for the average were used to calculate the combined forecast by this method.

2.3.5 Penalised quantile regression – LASSO

There are many different types of penalties under regularization introduced to obtain a selection of variables. The least shrinkage and selection operator (LASSO) penalty was applied in proposed by Tibshirani (1996) for the selection of variables. This method retains a lot of advantages of best subset selection: gives a sparse solution; ensures the stability of the model selection; provides objective estimates for large coefficients. These are the desirable qualities of good punishment. (Fan and Li 2001). The objective of the LASSO is to solve:

$$\beta_T = \underset{\beta_T}{\operatorname{argmin}} \begin{matrix} \{\sum_{r_T \geq VaR_{comb,T} * \lambda_T} p * \left| r_T - VaR_{comb,T} * \beta_T \right| + \\ \sum_{r_T < VaR_{comb,T} * \lambda_T} (1-p) * \left| r_T - VaR_{comb,T} * \beta_T \right| + \lambda * ||\beta_T||_1\} \end{matrix} \tag{11}$$

Where:

$VaR_{12,T} = (1, VaR_T^1, VaR_T^2)$ is a vector of VaRs predicted from standalone models (1 is here to capture the intercept);

$\beta_T = (\beta_{T,0}, \beta_{T,1}, (1 - \beta_{T,1}))$ is a vector of parameters for VaRs predicted from standalone models.

2.3.6 Penalised quantile regression – elastic net

Another penalized quantile regression is elastic net, which was applied as a VaR combination technique (Bayer 2017). Results obtained by Bayer suggest that there is no difference between regularization with the elastic net penalty of Zou and Hastie (2005) and the LASSO. The elastic net offers a compromise between ridge and lasso. This method shrinks variables into groups and sets some coefficients to zero. Thus, the elastic net, combines the strengths of both approaches. The objective of the net penalty is to solve below expression:

$$\beta_T = \underset{\beta_T}{\operatorname{argmin}} \begin{matrix} \{\sum_{r_T \geq VaR_{comb,T} * \lambda_T} p * \left| r_{T+s} - VaR_{comb,T} * \beta_T \right| + \\ \sum_{r_T < VaR_{comb,T} * \lambda_T} (1-p) * \left| r_T - VaR_{comb,T} * \beta_T \right| + \\ \lambda * (\delta * ||\beta_T||_1 + (1-\delta) * ||\beta_T||_2^2/2\} \end{matrix} \tag{12}$$

Where:

$\delta \in [0{:}1]$ – the parameter, which balances the ridge and the lasso, which in the case of the study is set to be equal to 0.5.

2.3.7 Quantile Random forest

Breiman (2001) introduced random forests as a machine learning tool. Since then, it has become very popular and powerful in case of regression and multivariate classification. Meinshausen and Ridgeway (2006) proved that random forests provide information about the complete conditional distribution of the response variable, not only about the conditional mean. Conditional quantiles can be modelled by quantile regression forests, generalization of random forests. This method provides a non-parametric and accurate way to estimate conditional quantiles for multivariate predictor variables. Numerical examples suggest that the algorithm is competitive in its predictive power. A quantile loss function was applied without any penalty:

$$\beta_T = \underset{\beta_T}{\operatorname{argmin}} \begin{matrix} \{\sum_{r_T \geq VaR_{comb,T+s} * \lambda_T} p * \left| r_T - VaR_{comb,T} * \beta_T \right| + \\ \sum_{r_T < VaR_{comb,T} * \lambda_T} (1-p) * \left| r_{T+s} - VaR_{comb,T} * \beta_T \right| \} \end{matrix} \quad (13)$$

To estimate the influence of the individual forecasts on the combined forecast, in the study the importance measure is used. The influence is computed from permuting out-of-bag data. A forecast error (mean squared error) is logged for each tree. Then the same thing happens after the permutation of each predictor variable. The difference between them is then averaged over all trees and normalized by the standard deviation of the differences (Grömping 2009).

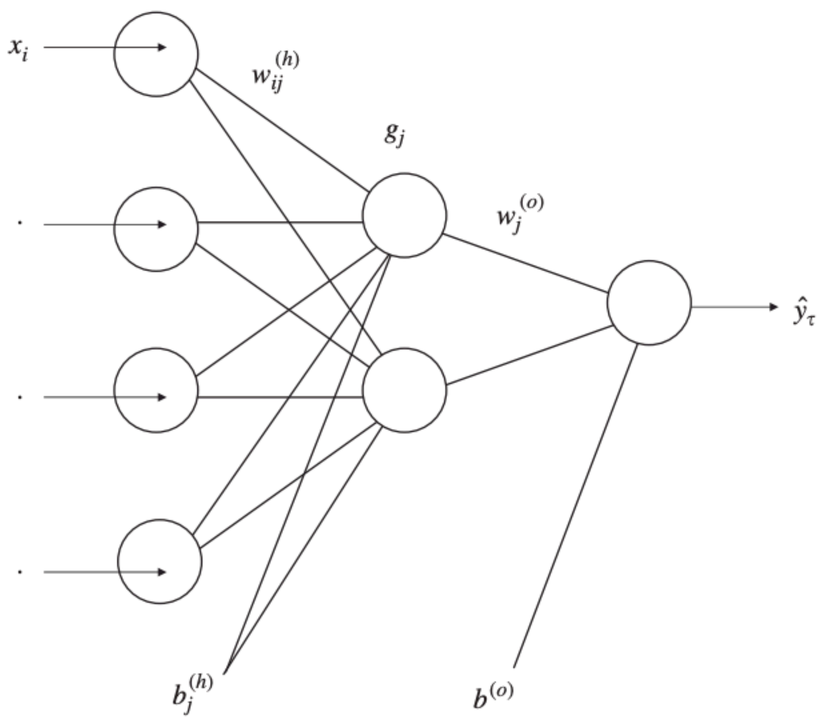
2.3.8 Generalized Boosted Regression Model

Boosting algorithms were originally introduced to solve classification problems. (Freund 1995). The basic approach is to iteratively combine a few simple models, called "weak learners", to obtain a "strong learner" with better prediction accuracy (Friedman et al. 2000). Currently boosting takes different forms with different loss functions, basic models, and different optimization schemes. Friedman (2001) created the groundwork for a new generation of boosting algorithms by finding the link between boosting and optimization. His work proposed a Gradient Boosting Machine (GBM), which can be easily applied to forecast quantile distribution. the GBM algorithm iteratively adds in at each step a new decision tree (i.e. "weak learner") that best limits the loss function. More specifically, in regression, the algorithm starts by initializing the model through the first guess, which is usually a decision tree that reduces the loss function as much as possible, and then at each step the new decision tree is matched with the current remainder and added to the previous model to update the rest. The algorithm continues iterating until the maximum number of iterations specified by the user is reached. One of the greatest practical advantages of using the GBM model is its flexibility and accuracy of forecasts. In addition, the possibility of selecting variables in the GBM model allows the inclusion of non-influencing parameters without reducing the predictability of the model (Touzani et al. 2018). In this method, it is not easy to determine the influence of individual dependent variables on the final result. In order to estimate the influence of inputs, the relative influence presented by Friedman (2001) was used.

2.3.9 Quantile regression neural network

The Quantile Regression Neural Network (QRNN) model is based on the standard multilayer perceptron Artificial Neural Network (ANN) (Gardner and Dorling (1998), Hsieh and Tang, (1998)). Method of calculation is presented in figure 3. First, output from the j-th hidden-layer node $g_j(t)$ is given by applying the hyperbolic tangent, a sigmoidal transfer function, to the inner product between $x_i(t)$ and the hidden-layer weights $w_{ij}^{(h)}$ plus the hidden-layer bias $b_j^{(h)}$. An estimate of the conditional quantile is then given by applying sigmoid transfer function to are the output-layer weights, $w_j^{(o)}$, and is the output-layer bias, $b^{(o)}$.

Figure 3. The diagram of QRNN model with four predictors and two hidden nodes.



Source: Cannon, 2011.

The QRNN model may be a viable alternative to parametric ANN models for modelling extremes (Cannon 2010). However, a drawback of this model is that plots of quantile regression coefficients can be used to gain insight into predictor-prediction relationships and predictive distribution relationships, while performing QRNN analysis is complicated due to the fact that the model is nonlinear, which means that partial derivatives of the model results for the predictors may differ across the predictor space. The evaluation of the impact of individual variables consists of calculating the partial derivatives of the output according to the input variables. This method turned out to be the most useful in Gevrey et al. (2003).

2.4 Bactesting

This section describes the backtesting tests (Excees Ratio, Kupiec test, Christoffersen test, Dynamic Quantile test, Traffic light test) and how models were compared (Model Confidence Set Procedure). It consists of mathematical formulas, informativeness as well as advantages and disadvantages of each approach.

2.4.1 Excess ratio

The excess ration measures the number of cases, when VaR is greater than the observable return, relative to the total number of observations. The excess ratio is described by the formula:

$$ER = \frac{\sum_{t=1}^{n} 1_{r_t < VaR_\alpha(t)}}{n} \tag{14}$$

where:

n - the number of VaR forecasts,

$1_{r_t < VaR_\alpha(t)}$ - the number of cases, for which the VaR forecast was larger than return on the same day.

The excess ratio can then be expressed as the percentage of model failure, which, for the correctly forecasting model, should (in theory) be equal the significance level, at which VaR was calculated.

2.4.2 Kupiec test

The Kupiec Test (1995) is a non-parametric test based on the proportion of exceedances. Due to this, the deviations on both sides of the assumed number of exceedances are considered and the test statistics are built on this difference. It has a chi square distribution with one degree of freedom and looks like this:

$$LR_{uc} = 2 * \ln\left(\frac{1-\hat{\alpha}}{1-\alpha}^{N-X} * \left(\frac{\hat{\alpha}}{\alpha}\right)^{X}\right) \sim \chi^2(1) \tag{15}$$

Where:

α - assumed excess ratio,

$\hat{\alpha}$ - empirical excess ratio,

N - number of VaR forecasts,

X - number of VaR forecasts exceedances.

In the Kupiec test, a null hypothesis H0: $\alpha = \hat{\alpha} = X / N$ is tested, i.e. the assumption that the ratios of the theoretical and empirical excess are equal. This test is used to check both models in terms of underestimation and overestimation.

2.4.3 Christoffersen test

The Kupiec test verifies the hypothesis about the correct quality of exceedances. However, the test does not respond to the presence of clusters in VaR exceedances. It is important to identify the excesses that do not meet the independence condition. Independence is inviolable a feature, as the concentration of VaR exceedances increases the risk and contributes to the accumulation of losses (Jeziorski 2014). The goal of the Christoffersen test (1998) is not to test the significance of the model, but to focus on the fact that the fraction of exceedances is consistent with the assumed one and that the sequence of exceedances is independent. The test statistics is the chi-squared distribution with two degrees of freedom. Its formula is as follows:

$$LR_{cc} = LR_{uc} + LR_{ind} \sim \chi^2(2) \quad (16)$$

Where:

LR_{uc} - statistic of the Kupiec test,

LR_{ind} - statistic of the VaR forecast independence test, coming from the chi-squared distribution with one degree of freedom. It tests the null hypothesis of independence of exceedances against the alternative hypothesis that the exceedances are characterized by a first-order Markov chain.

2.4.4 Dynamic Quantile test

Another test considered in the research was the DQ test proposed by Engle and Manganelli (2004). The purpose of this test is to jointly check the presence of autocorrelation in the VaR forecasts and whether the number of exceedances is as expected. The null hypothesis of the DQ test is that all coefficients in the regression below are zero.

$$I_t = \beta_0 + \sum_{i=1}^{p} \beta_i * I_{t-1} + \sum_{j=1}^{q} \mu_j * X_j + \varepsilon_t \quad (17)$$

Where:

I_t - $\frac{1 - \alpha \; for \; r_t < VaR_\alpha(t)}{\alpha \; for \; r_t \geq VaR_\alpha(t)}$, where α = significance level of VaR forecasts,

X_j - all explanatory variables used as the information set for forecasting,

p - number of lags of the dependent variable,

q - number of lags of independent variables.

The most commonly used explanatory variable in the DQ test is lagged VaR forecasts. The DQ test statistic is derived from the chi-square distribution with $p + q$ degrees of freedom and is as follows:

$$DQ = \frac{I_t' X(X'X)^{-1} X' I_t}{\alpha * (1-\alpha)} \sim \chi^2(p+q) \tag{18}$$

Where:

I_t - vector of exceedances,

X - matrix, that the columns are p lags of the exceedances and q lags of explanatory variables.

2.4.5 Traffic light test

The Basel traffic light test is based on the exceedance index value. The assessment of the quality of the VaR forecast is made on the basis of assigning, respectively: green (no problems with forecast quality, model recognized as valid), yellow (recommended model supervision, warning zone) and red (the model almost certainly generates VaR forecasts of poor quality) zones. Lights are assigned on the basis of exceeding successive thresholds of the exceedance indicator. The green zone is where the cumulative binomial distribution (with: α probability of success, where in this study α is equal to 1% or 2.5%, and N trials, where N is equal to the number of VaR forecasts) is lower than 0.95. The yellow zone begins at the point where the distribution is greater than or equal to 0.95. Likewise, the red zone begins at the point where the value of the same distribution is greater than or equal to 0.9999 (BCBS 1996).

2.4.6 Model Confidence Set procedure

The Hansen et al. (2011) procedure consists of a sequence of statistical tests that allow the construction of a set of "superior" models in which the null hypothesis of equal predictive power (EPA) is not rejected with a certain probability of the α level. This is done by sequentially removing the worst model from the set. To calculate EPA statistics, the asymmetric VaR loss function by González-River Lee, and Mishra (2004) should be defined:

$$l(r_t, VaR_t) = \big(\tau - 1(r_t < VaR_t)\big) * (r_t - VaR_t) \tag{19}$$

The asymmetric VaR loss function is a natural candidate for testing historical quantile risk measures as it is more affected by observations below the τ quantile level.

The procedure starts with an initial set of models (M_0) of dimension m, including all alternative model specifications, and provides a smaller and better model set (M_1) with dimension lower than m, which includes all models with higher predictive ability according to the selected loss function. Let $d_{ij,t}$ denote the difference in losses between the models i and j over time t:

$$d_{ij,t} = l_{i,t} - l_{j,t}, \; i,j = 1,\dots,m, \; t = 1,\dots,n \tag{20}$$

The EPA hypothesis for a given set of M models can be formulated as follows:

$$H_{0,M} = E\left(d_{ij}\right) = 0, \qquad i,j = 1,\dots,m$$

$$H_{A,M} = E\left(d_{ij}\right) \neq 0, \; i,j = 1,\dots,m \tag{21}$$

This hypothesis can be tested by constructing the following statistic:

$$t_{ij} = \frac{n^{-1} * \sum_{t=1}^{n} d_{ij,t_{,t}}}{\sqrt{\widehat{VaR}\left(n^{-1} * \sum_{t=1}^{n} d_{ij,t_{,t}}\right)}} \tag{22}$$

The variance $\widehat{VaR}$ is bootstrap estimates of $VaR(n^{-1} * \sum_{t=1}^{n} d_{ij,t_{,t}})$.

3. Results

This section presents data analysis, compares the accuracy of individual models and methods for combining forecasts, and presents weights assigned to individual models for combining forecasts.

3.1 Data analysis

In order to check the characteristics of the returns, the basic statistics were counted. Table 1 shows the statistics for the daily logarithmic rate of return (minimum, maximum, skewness, kurtosis and quantiles), the Jarque-Bera's test value, and its p-value (in parentheses).

Table 1. Statistics of prices' log-returns

Commodity	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	J-B test	Skewness	Ex. Kurtosis
Gold	-0.0982	-0.0049	0.0005	0.0004	0.006	0.0864	6398 (<0.001)	-0.2658	8.7160
Silver	-0.1955	-0.0080	0.0011	0.0003	0.0090	0.1220	13942 (<0.001)	-0.9263	10.8079
Oil	-0.2799	-0.0128	0.0008	0.0001	0.0130	0.3196	52559 (<0.001)	-1.9164	52.6291
Gas	-0.1990	-0.1911	-0.0007	-0.0001	0.0173	0.3238	6833 (<0.001)	0.5643	8.7537
Copper	-0.1169	-0.0082	0.0002	0.0003	0.0089	0.1177	4279 (<0.001)	-0.1731	7.6239

Source: Own calculation.

No commodity has a distribution corresponding to the normal distribution. All distributions are leptokurtic (Excess Kurtosis far above 0). Moreover, the distributions are left-skewed for oil, gold, silver and copper and right-skewed for gas. The last one is not surprising because previous studies (Tse 2016) have shown the same characteristic. The highest values of kurtosis and skewness are for oil. This fact can be explained by the steady rises in oil prices over the past 20 years, with the exception of sharp slumps in bad economic times.

This was the case for gold and silver. Following the above procedure, we obtained the following models:

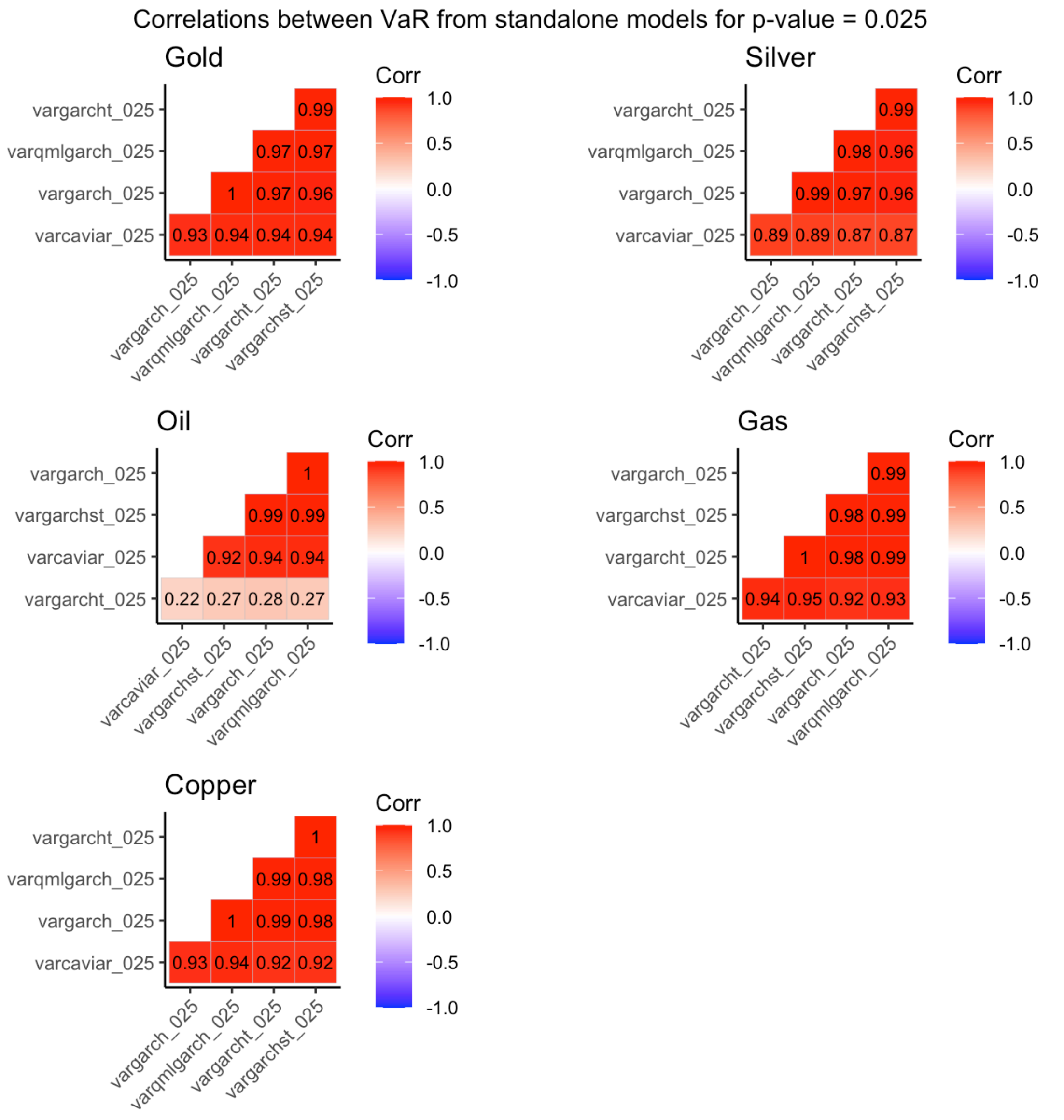
- Gold – GARCH(1,1), AR(1)-GARCH-t(1,1), AR(1)-GARCH-st(1,1), QML-GARCH(1,1), Indirect GARCH(1,1),
- Silver – GARCH(1,2), AR(1)-GARCH-t(1,1), AR(1)-GARCH-st(1,1), AR(1)-QML-GARCH(1,1), Indirect GARCH(1,1),
- Oil – GARCH(1,1), GARCH-t(1,1), GARCH-st(1,1), QML-GARCH(1,1), Indirect GARCH(1,1),
- Gas – GARCH(1,1), GARCH-t(1,1), GARCH-st(1,1), QML-GARCH(1,1), Indirect GARCH(1,1),
- Copper – GARCH(1,1), GARCH-t(1,2), GARCH-st(1,1), QM-GARCH(1,1), Indirect GARCH(1,1).

Figure 4 and Figure 5 show the correlation between the forecasts for individual models for 0.025 and 0.01 VaR' alpha level. It is the best to combine the least correlated forecast, but choosing based only on correlations suboptimal solution may be chosen. That's why mean of the best performing model (based on MCS procedure in the in-sample period – 1 September, 2000 to 2 July, 2004) and the least correlated with that model VaR was calculated. it has been concluded that the best averages for commodities are as follows:

- gold – GARCH + CaViaR for both p-value (correlation 0.93 for p-value = 0.025 and 0.81 for p-value = 0.01);
- silver – GARCH-st + CaViaR (correlation 0.87 for p-value = 0.025 and correlation 0.8 for p-value = 0.01);
- gas - GARCH-t + CaViaR for both p-value (correlation 0.94 for p-value = 0.025 and 0.89 for p-value 0.01);
- oil - GARCH-t + CaViaR for both p-value (correlation 0.22 for p-value = 0.025 and 0.27 for p-value 0.01);

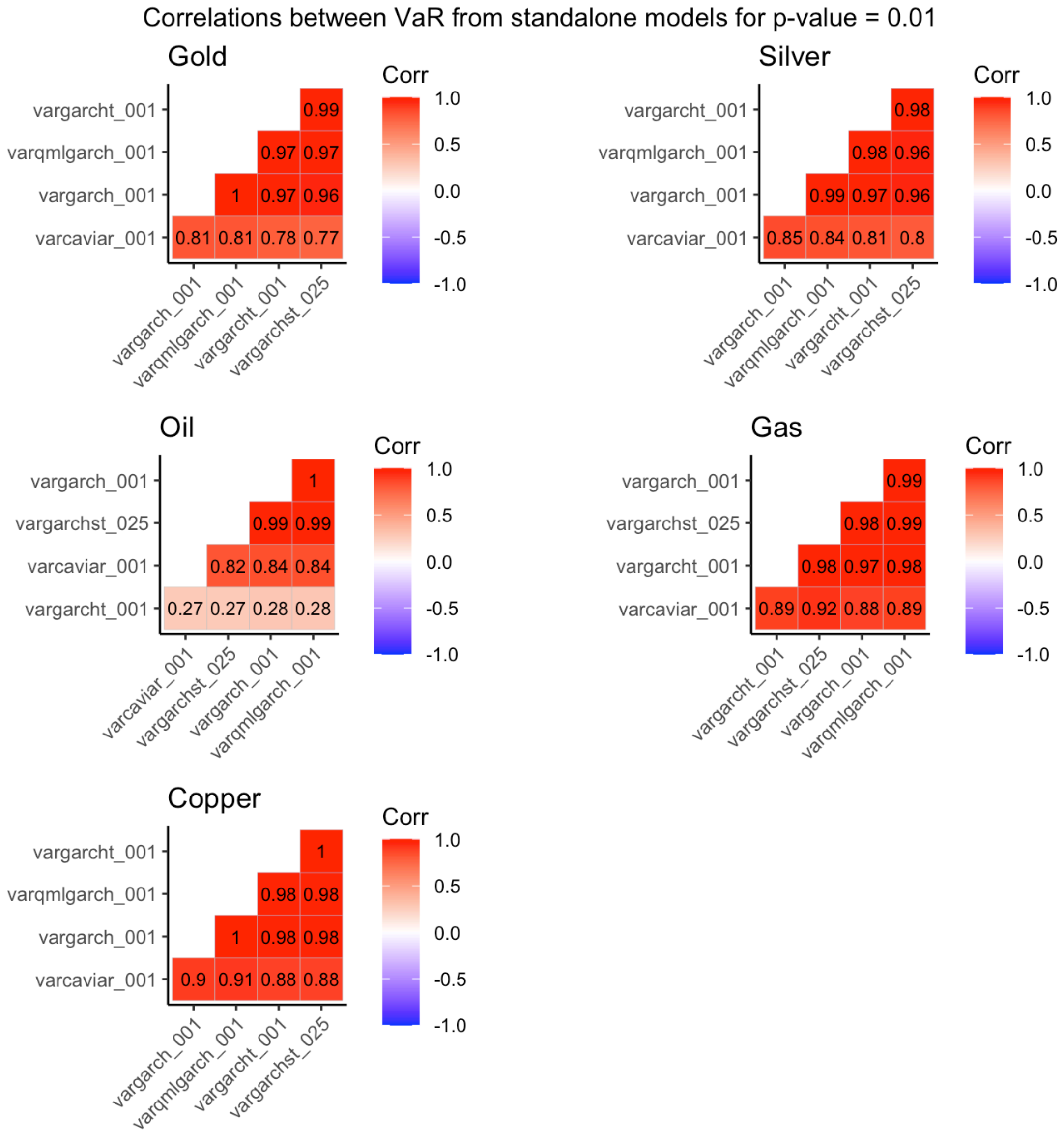
- copper - GARCH-st + CaViaR for both p-value (correlation 0.93 for p-value = 0.025 and 0.9 for p-value 0.01).

Figure 4. Correlations between VaR forecasts for p-value = 0.025



Source: Own calculations.

Figure 5. Correlations between VaR forecasts for p-value = 0.01

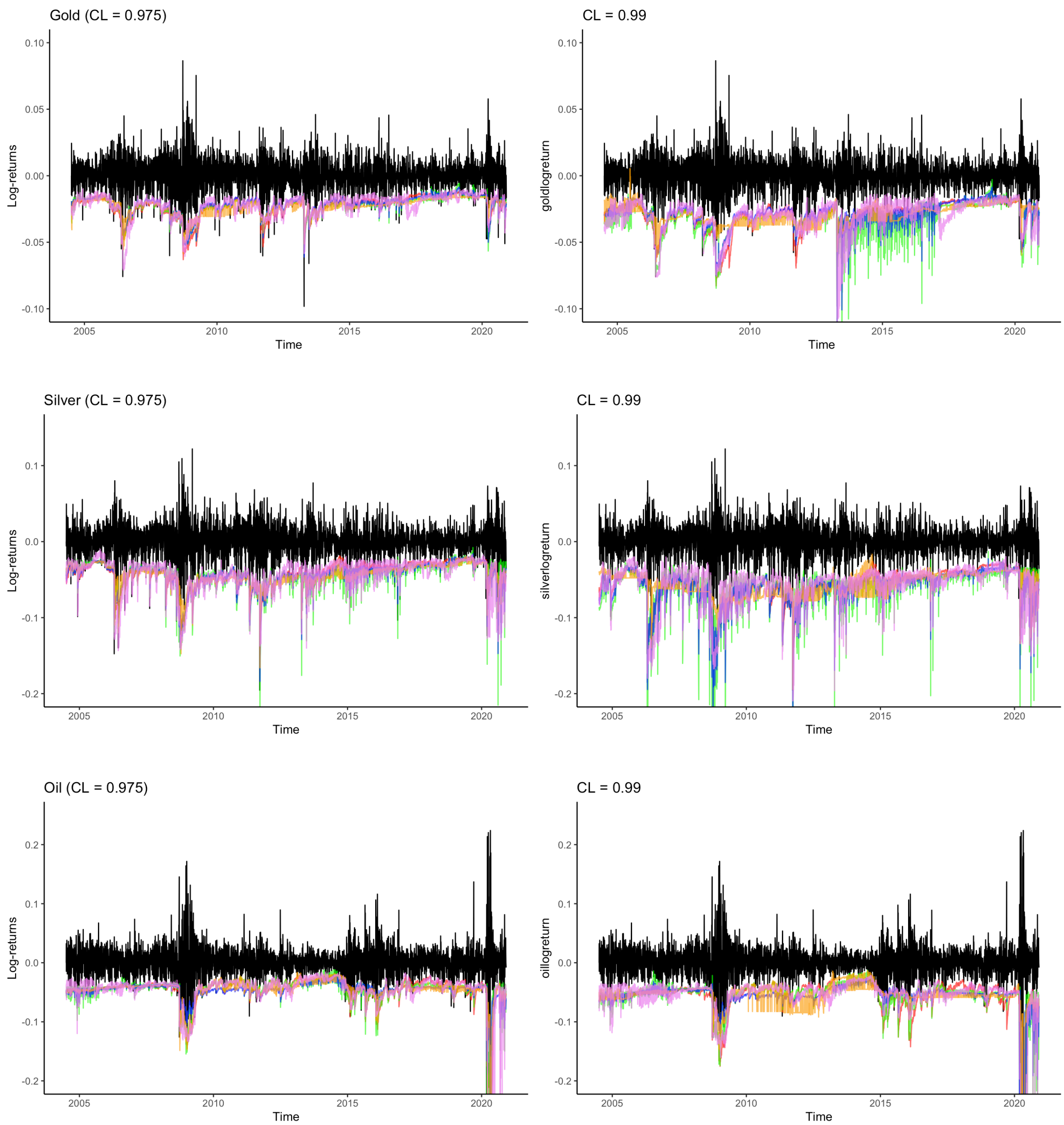


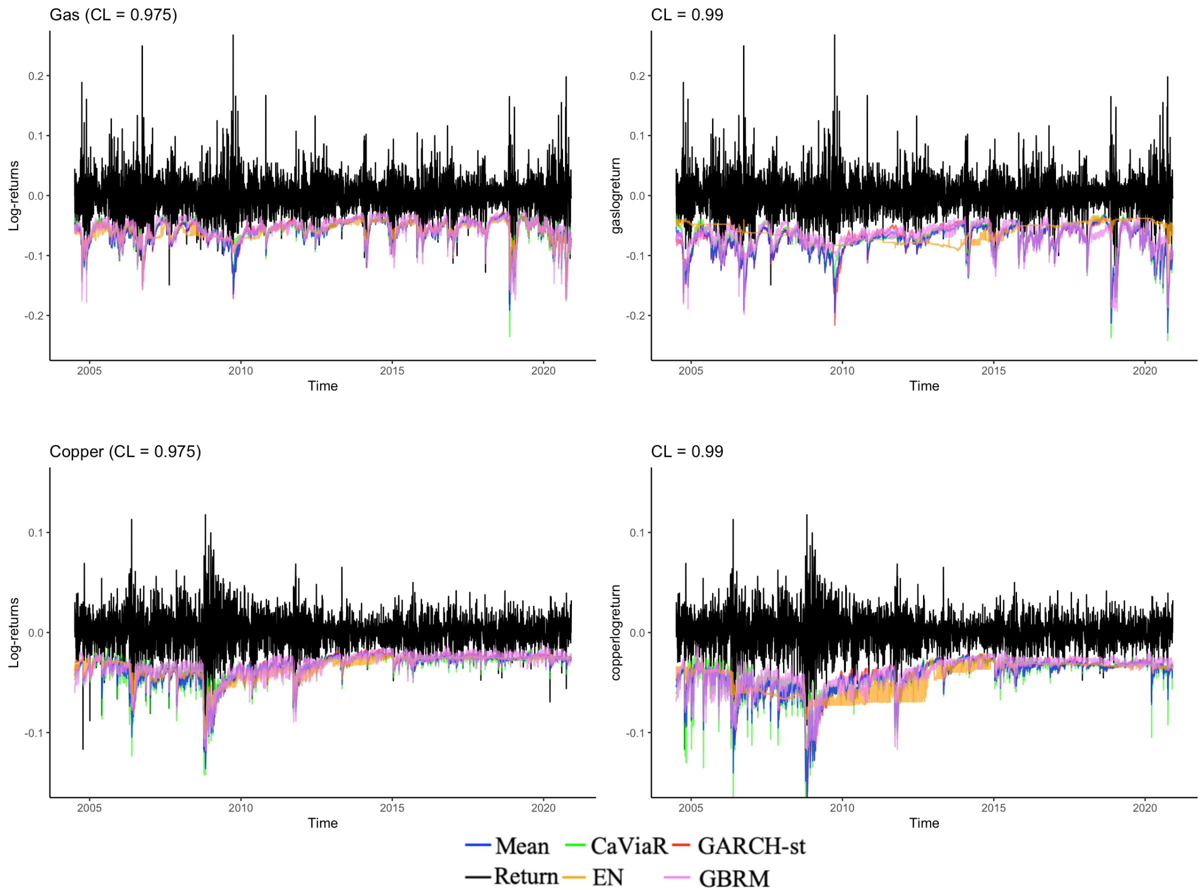
Source: Own calculations.

3.2 Empirical results for individual and combined methods

The analysis began with an analysis of the graphical accuracy of the models. Figure 6 shows the log-returns and out-of-sample VaR sequences for selected models. For all assets, the CaViaR model turned out to be the most conservative measure. Often when other models took a relatively liberal VaR value, CaViaR estimated it much lower (see, for example, Gold or Silver for a confidence level of 0.99). Probably it is the consequence that for all of them the mean contained CaViaR, so its value was between CaViaR and some other model. Forecasts achieved by quantile regression using elastic net regularization have proven to flatten the VaR along the entire time horizon. Probably using this method will record many exceedances. The most liberal method (except the highest VaR), occurs to be combining forecast using gradient boosting quantile regression. The conditional quantile optimization method turned out to be in the middle, however, sometimes, as for example for oil and copper for a confidence level of 0.99, it tended to stay at a low level for quite a long time.

Figure 6. Log-returns and VaR forecast for all assets for both confidence level (CL) – 0.975 (on left) and 0.99 (on right).





Source: Own calculations.

Tables 2 to 11 presents the results of backtesting for each of the assets, and each method for two different confidence levels - 0.975 and 0.99. The tables contain excess ratio, merchant test, charvre, dynamic quantile, and traffic light results. The results are presented separately for four periods. The entire assessment period, the period beyond crises, the period of crises, and the period of the current coronavirus crisis.

For all commodities throughout the all assessment horizon, at least one individual method works well, i.e., at least one of the tests (UC, CC or DQ) has a p-value greater than 0.05 and the traffic light test result is green. GARCH-st is the only model that has always met the conditions for each asset during whole assessment period. CaViaR failed to meet them in only one case, i.e. for oil at a confidence level of 0.99. GARCH-t works well for gold, gas and copper. GARCH only gives accurate results for Oil Confidence Level of 0.975 and gas for both Confidence Levels, while QML-GARCH only works for Gas Confidence Levels. The results for the combined models are not so clear. For gold at a confidence level of 0.975, almost all forecast combining methods give promising results. Only the highest VaR, random forests and neural networks fail here. On the other hand, for confidence level 0.99, only the lowest VaR meets the criteria listed in the previous paragraph. For silver at both levels, the simple mean met all conditions. Additionally, for the 0.975 level, also the lowest VaR, CQOM, and two regression methods (elastic net and lasso) give good results. For oil for the confidence level of 0.975, the mean gives accurate forecasts, and for the 0.99 level no method of combining forecasts gives promising results. For gas at the confidence level of 0.975, the exact results are given by the mean, the highest VaR, the lowest VaR, elastic net and LASSO, while for the level 0.99 only mean and lowest VaR give accurate forecasts. For both levels of confidence, good results are given by the average, lowest VaR and elastic net. In addition, for confidence level 0.975 also LASSO. Summarizing the analysis of the entire assessment period, the GARCH-st or CaViaR model seems to be the most appropriate among the individual models, while the mean and regression based combined models appear most often as the best from forecast combining models, although this differs depending on the commodity.

For a calm period, i.e. for all periods outside of crisis periods, many models, both individual and combined, seem to give accurate results. Of the individual models, only GARCH fails for silver for both confidence levels and for gold for the confidence level of 0.99, and GARCH-t fails for silver for the confidence level of 0.975. Among the forecast combining models, random forests and neural networks do not provide accurate results for all commodities, regardless of the confidence level. The highest VaR and CQOM also don't perform well. The best results are obtained from the mean, lowest VaR and two regression methods. These give good results for any confidence level and for any raw material. Once again mean and regression based combining models gives promising results.

Standalone methods do not work well in times of crisis. For gold, for the confidence level of 0.975, exact results are given by GARCH-st and CaViaR, and for the level of 0.99, also GARCH-t. In the case of silver, none of the individual models gave good results. Excess ratio for all of them is above 3.3% for confidence level of 0.975 and 1.53% for 0.99. For oil only at the confidence level of 0.99 GARCH-st gave accurate forecasts. For gas at the confidence level of 0.975 GARCH, GARCH-t and QML-GARCH forecasted well, and at the level of 0.99 only GARCH-t. For copper at a confidence level of 0.975, neither model gives good results, and at the level of 0.99 CaViaR gives the exact VaR value.

During the coronavirus pandemic crisis for gold, silver and oil, no individual model gives a good VaR forecast, regardless of the confidence level. For these assets the excess ratio is on average around 5% for a confidence level of 0.975 and 2.5% for a confidence level of 0.99. For the gas at the confidence level of 0.975 all individual methods give good predictions, and for the confidence level of 0.99 GARCH-t, GARCH-st and CaViaR. For copper the confidence level of 0.975, GARCH-t gives good results, and for the level of 0.99 no individual model gives good results (its excess ratios are above 2.16%). When it comes to combining forecasts, the lowest VaR proved particularly successful in this period. The exception to this rule is gold for the confidence level of 0.99, silver for both confidence levels and copper for the confidence level of 0.99. In addition, the CQOM also performed well here in two cases, i.e. gold and gas for the confidence level of 0.975. Other methods appeared only for forecasting VaR for gas. For the confidence level of 0.975, they were the mean, elastic net and lasso, and for the confidence level of 0.99 it was the mean.

Table 2. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for gold for confidence level equal to 0.975.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	3.10%	5.77	6.02	21.07	Y	2.85%	1.38	1.55	9.22	G	3.70%	6.45	6.50	16.92	Y	5.63%	6.88	6.98	26.48	Y
GARCH-t	2.81%	1.62	1.77	18.08	G	2.57%	0.06	0.07	2.48	G	3.38%	3.57	3.80	28.92	Y	4.76%	3.85	4.23	21.16	Y
GARCH-st	2.38%	0.24	0.41	16.80	G	2.20%	1.16	1.29	5.88	G	2.82%	0.50	1.36	21.53	G	4.33%	2.61	3.22	22.78	Y
QML-GARCH	3.17%	5.91	6.02	22.59	Y	2.88%	1.66	1.80	9.14	G	3.62%	5.66	5.74	19.26	Y	5.59%	6.80	6.98	26.36	Y
CaViaR	2.65%	0.36	0.74	20.85	G	2.47%	0.01	0.04	15.69	G	3.06%	1.49	2.02	14.36	G	5.19%	5.28	7.53	31.81	Y
Mean	2.81%	1.62	1.77	14.83	G	2.64%	0.24	0.24	9.74	G	3.22%	2.43	2.79	11.04	Y	4.76%	3.85	4.23	15.45	Y
Highest VaR	3.46%	14.18	14.38	41.56	R	3.19%	5.25	5.25	15.70	Y	4.11%	11.05	11.42	33.81	Y	6.49%	10.57	11.54	41.72	Y
Lowest VaR	2.07%	3.37	3.39	10.49	G	1.92%	4.35	4.36	8.59	G	2.42%	0.04	0.14	7.87	G	3.46%	0.79	2.04	7.83	G
CQOM	2.69%	0.63	9.01	56.89	G	2.47%	0.01	4.21	39.29	G	3.25%	2.48	6.46	44.63	Y	3.90%	1.58	2.48	11.99	G
Elastic Net	2.50%	0.00	1.85	24.77	G	2.20%	1.16	1.39	7.99	G	3.27%	2.49	4.23	30.53	Y	6.49%	10.57	11.54	81.74	Y
LASSO	2.62%	0.25	6.95	30.04	G	2.26%	0.69	6.14	14.90	G	3.46%	4.22	5.51	44.51	Y	7.36%	14.83	15.27	91.46	R
QRF	5.15%	92.01	92.12	246.98	R	4.87%	52.90	52.90	158.42	R	5.80%	40.61	41.02	93.76	R	8.66%	22.16	22.59	57.90	R
GBRM	2.89%	2.43	2.51	16.97	G	2.57%	0.06	0.07	5.19	G	3.62%	5.66	5.74	19.85	Y	4.76%	3.85	4.23	17.65	Y
QRNN	4.71%	66.66	70.04	223.39	R	4.49%	38.60	39.35	105.05	R	5.23%	29.10	32.44	212.53	R	9.09%	24.82	29.35	176.38	R

Source: based on own calculations.

Note: Gray fields indicate p-values greater than 5%. GARCH stands for GARCH(1,1), GARCH-t - AR-GARCH-t(1,1), GARCH-st - AR-GARCH(1,1), QML-GARCH - QML-GARCH(1,1), CaViaR - Indirect GARCH(1,1), Mean stands for simple average from GARCH and CaViaR, Highest VaR means the maximum from GARCH, GARCH-t, GARCH-st, QML-GARCH, and CaViaR, Lowest VaR stands for the minimum from individual models, CQOM stands for Conditional Quantile Optimization Method applied for GARCH and CaViaR (described in section 2.3.4), Elastic Net stands for forecast combined using quantile regression with elastic net regularization (described in section 2.3.6), LASSO stands for forecast combined using quantile regression with LASSO regularization (described in section 2.3.5), QRF stands for forecast combined using Quantile Regression Forests (described in section 2.3.7), GBRM stands for forecast combined using Gradient Boosting Regression Model (described in section 2.3.8), QRNN stands for forecast combined using Quantile Regression Neural Network (described in section 2.3.9). In TL: 1) G stands for green, 2) Y stands for yellow, 3) R stands for red.

Table 3. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for gold for confidence level equal to 0.99.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	1.85%	24.38	24.59	47.78	R	1.65%	10.30	10.36	20.23	Y	2.33%	16.25	16.39	40.37	R	3.46%	8.64	9.89	31.80	Y
GARCH-t	1.13%	0.69	6.40	18.19	G	1.06%	0.12	1.02	3.17	G	1.29%	0.96	6.88	34.25	G	2.19%	2.39	5.38	30.03	Y
GARCH-st	0.91%	0.32	4.16	9.17	G	0.89%	0.36	1.79	3.81	G	0.97%	0.01	2.69	9.69	G	2.16%	2.37	5.38	30.07	Y
QML-GARCH	1.90%	24.66	24.59	47.38	R	1.67%	10.35	10.36	20.02	Y	2.37%	16.43	16.39	40.04	R	3.55%	8.87	9.89	31.55	Y
CaViaR	1.03%	0.05	3.07	15.54	G	1.10%	0.27	1.09	6.80	G	0.89%	0.17	3.18	13.09	G	2.60%	4.13	6.42	26.36	Y
Mean	1.30%	3.43	5.09	18.31	Y	1.30%	2.48	2.89	8.34	G	1.29%	0.96	2.61	18.09	G	3.46%	8.64	9.89	31.23	Y
Highest VaR	1.88%	25.64	26.93	52.17	R	1.65%	10.30	10.36	19.95	Y	2.42%	18.01	19.63	49.68	R	3.46%	8.64	9.89	31.27	Y
Lowest VaR	0.84%	1.11	5.53	12.07	G	0.86%	0.63	2.19	3.78	G	0.81%	0.51	3.89	13.52	G	2.16%	2.37	5.38	30.17	Y
CQOM	1.54%	10.50	13.28	51.94	Y	1.44%	5.04	7.12	34.06	Y	1.77%	6.07	6.77	35.15	Y	2.60%	4.13	6.42	52.00	Y
Elastic Net	1.32%	3.98	11.47	62.74	Y	1.10%	0.27	8.30	39.08	G	1.85%	7.28	7.86	40.96	Y	3.46%	8.64	9.89	43.82	Y
LASSO	1.27%	2.92	4.68	34.99	Y	1.06%	0.12	1.02	7.58	G	1.77%	6.07	6.77	42.25	Y	3.90%	11.30	12.19	48.62	Y
QRF	3.61%	171.00	171.03	448.16	R	3.64%	122.05	122.39	323.10	R	3.54%	48.97	52.20	148.36	R	4.76%	17.29	18.39	73.79	R
GBRM	1.35%	4.56	11.81	80.37	Y	0.96%	0.05	4.83	17.66	G	2.25%	14.56	16.57	99.99	R	2.60%	4.13	6.42	47.14	Y
QRNN	3.42%	150.49	160.82	541.10	R	3.29%	96.70	102.40	349.83	R	3.70%	54.22	58.90	206.00	R	6.06%	27.68	27.71	80.57	R

Source: based on own calculations.

Note: The same as for the pervious table.

Table 4. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for silver for confidence level equal to 0.975.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	3.58%	17.72	21.25	41.94	R	3.46%	9.96	11.56	22.90	Y	3.86%	8.16	10.22	22.05	Y	5.63%	3.85	6.69	28.62	Y
GARCH-t	3.42%	12.86	21.07	38.19	Y	3.16%	4.76	9.27	19.21	Y	4.03%	10.04	13.55	24.26	Y	4.76%	2.85	6.14	22.35	Y
GARCH-st	2.91%	2.74	13.82	34.46	Y	2.68%	0.36	8.13	24.97	G	3.46%	4.22	7.42	16.69	Y	4.33%	2.74	6.14	23.15	Y
QML-GARCH	3.44%	13.51	21.50	43.07	R	3.40%	8.65	11.93	24.52	Y	3.54%	4.91	10.27	22.85	Y	5.69%	2.90	6.14	23.09	Y
CaViaR	2.84%	1.87	3.74	25.47	G	2.64%	0.24	0.24	9.71	G	3.30%	2.98	6.86	29.91	Y	5.19%	10.57	11.72	38.85	Y
Mean	2.77%	1.17	4.91	16.11	G	2.47%	0.01	0.75	5.15	G	3.46%	4.22	7.55	18.32	Y	4.76%	6.88	8.88	31.12	Y
Highest VaR	4.14%	38.32	44.18	74.44	R	3.81%	17.68	20.61	38.17	R	4.91%	23.23	25.99	42.50	R	6.49%	12.63	13.45	34.59	R
Lowest VaR	2.21%	1.46	6.86	19.48	G	2.09%	2.10	2.47	10.64	G	2.50%	0.00	7.41	22.49	G	3.46%	2.61	6.14	30.82	Y
CQOM	2.38%	0.24	18.45	51.46	G	2.16%	1.44	14.15	34.88	G	2.90%	0.77	6.12	20.39	G	3.90%	2.69	6.14	27.15	Y
Elastic Net	2.48%	0.01	8.00	30.60	G	2.30%	0.50	3.40	15.17	G	2.91%	0.78	6.12	22.45	G	6.49%	3.85	6.69	34.39	Y
LASSO	2.72%	0.79	17.15	72.79	G	2.26%	0.69	12.27	39.12	G	3.78%	7.28	11.65	55.77	Y	7.36%	17.15	17.42	73.08	R
QRF	4.81%	72.00	73.95	217.41	R	4.39%	35.03	35.37	107.96	R	5.80%	40.61	42.44	126.89	R	8.66%	10.57	10.57	55.56	Y
GBRM	3.32%	10.41	19.53	59.65	Y	3.05%	3.42	6.62	22.87	Y	3.95%	9.08	15.27	47.33	Y	4.76%	6.88	8.63	29.33	Y
QRNN	5.00%	83.19	102.20	395.17	R	4.70%	46.17	55.58	253.17	R	5.72%	38.88	48.45	157.84	R	9.09%	39.66	40.11	147.69	R

Source: based on own calculations.

Note: The same as for the table 2, but here GARCH stands for GARCH(1,2), QML-GARCH stands for AR-QML-GARCH(1,1), Mean stands for simple average from GARCH-st and CaViaR.

Table 5. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for silver for confidence level equal to 0.99.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	2.36%	56.00	62.62	114.64	R	2.30%	36.32	39.22	67.05	R	2.50%	19.83	23.85	56.09	R	3.46%	12.30	15.63	76.45	Y
GARCH-t	1.30%	3.43	19.68	58.92	Y	1.06%	0.12	8.50	25.70	G	1.85%	7.28	14.45	46.95	Y	2.23%	11.52	15.63	73.12	Y
GARCH-st	1.03%	0.05	21.68	77.52	G	0.75%	1.94	14.40	42.41	G	1.69%	4.96	13.18	47.30	Y	2.16%	11.30	15.63	75.03	Y
QML-GARCH	2.31%	52.56	62.26	120.78	R	2.20%	31.39	34.78	61.32	R	2.58%	21.72	28.68	71.24	R	3.51%	12.73	15.63	74.50	Y
CaViaR	1.11%	0.46	3.05	19.34	G	0.93%	0.16	1.48	4.38	G	1.53%	3.03	4.14	39.88	Y	2.60%	20.58	20.79	88.35	R
Mean	1.01%	0.00	11.44	36.79	G	0.82%	0.98	12.38	38.30	G	1.45%	2.22	3.49	11.08	Y	3.46%	11.90	12.19	48.59	Y
Highest VaR	2.65%	78.37	84.86	157.60	R	2.44%	43.32	45.64	76.26	R	3.14%	36.67	41.01	101.95	R	3.46%	31.46	32.44	130.63	R
Lowest VaR	0.70%	4.29	10.09	20.68	G	0.62%	4.99	7.69	9.93	G	0.89%	0.17	3.18	21.35	G	2.16%	4.13	6.42	55.81	Y
CQOM	1.64%	14.24	29.29	108.75	R	1.51%	6.61	19.03	76.54	Y	1.93%	8.57	11.55	52.32	Y	2.60%	11.60	12.19	54.94	Y
Elastic Net	1.30%	3.43	19.68	87.01	Y	1.10%	0.27	8.30	44.48	G	1.77%	6.07	13.76	55.34	Y	3.46%	14.19	17.72	82.18	R
LASSO	1.20%	1.62	24.93	104.71	G	1.06%	0.12	13.66	64.24	G	1.53%	3.03	12.44	50.30	Y	3.90%	11.93	15.63	65.06	Y
QRF	2.89%	99.07	99.71	290.41	R	2.64%	54.69	56.27	136.88	R	3.46%	46.41	46.60	187.82	R	4.76%	13.24	12.03	98.63	Y
GBRM	1.56%	11.38	23.52	68.29	Y	1.41%	4.32	13.57	42.35	Y	1.93%	8.57	11.55	40.85	Y	2.60%	2.37	11.96	92.74	Y
QRNN	3.61%	171.00	177.55	776.88	R	3.46%	109.12	115.76	459.11	R	3.95%	62.45	63.00	339.63	R	6.06%	47.94	48.22	305.35	R

Source: based on own calculations.

Note: The same as for the previous table.

Table 6. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for oil for confidence level equal to 0.975.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	2.96%	3.39	3.43	10.66	Y	2.37%	0.22	3.56	7.97	G	4.35%	14.30	15.37	19.82	R	5.63%	6.88	8.63	20.99	Y
GARCH-t	3.54%	16.26	37.79	313.36	R	1.75%	7.51	8.57	55.08	G	7.73%	90.37	99.57	311.51	R	4.76%	22.16	24.91	157.82	R
GARCH-st	2.43%	0.09	0.95	5.04	G	1.96%	3.83	6.10	8.43	G	3.54%	4.91	7.86	9.45	Y	4.33%	2.61	6.14	21.69	Y
QML-GARCH	3.04%	3.53	3.43	10.71	Y	2.39%	0.24	3.56	8.09	G	4.35%	14.30	15.37	19.84	R	5.77%	6.97	8.63	20.99	Y
CaViaR	2.57%	0.09	1.66	14.89	G	1.96%	3.83	3.84	7.94	G	4.03%	10.04	11.73	24.06	Y	5.19%	3.85	6.69	21.18	Y
Mean	2.50%	0.00	0.68	18.03	G	1.37%	18.14	19.25	26.72	G	5.15%	27.59	27.74	37.72	R	4.76%	5.28	5.49	11.16	Y
Highest VaR	4.86%	74.74	83.96	264.82	R	3.16%	4.76	4.76	29.49	Y	8.86%	125.64	131.16	328.47	R	6.49%	39.66	41.25	172.61	R
Lowest VaR	1.59%	16.27	18.80	24.20	G	0.99%	34.98	35.56	29.30	G	2.98%	1.10	3.52	7.61	G	3.46%	0.25	1.96	6.54	G
CQOM	3.01%	4.12	22.26	157.69	Y	1.78%	6.80	6.81	45.91	G	5.88%	42.38	56.14	181.95	R	3.90%	27.60	34.10	133.16	R
Elastic Net	2.98%	3.75	5.04	22.36	Y	1.96%	3.83	3.84	9.29	G	5.39%	32.24	32.77	59.60	R	6.49%	14.83	15.27	63.54	R
LASSO	3.03%	4.51	10.03	69.61	Y	1.78%	6.80	7.78	12.84	G	5.96%	44.17	45.66	108.55	R	7.36%	17.15	18.83	85.82	R
QRF	5.82%	137.72	139.48	253.06	R	5.28%	70.54	70.63	217.80	R	7.09%	72.17	74.44	124.92	R	8.66%	19.60	25.83	74.65	R
GBRM	3.39%	12.23	12.24	24.87	Y	2.81%	1.13	2.11	15.93	G	4.75%	20.50	21.00	32.50	R	4.76%	3.85	4.23	13.35	Y
QRNN	5.65%	125.62	126.08	208.16	R	4.73%	47.48	50.07	297.87	R	7.81%	92.75	92.77	156.26	R	9.09%	53.11	53.12	134.19	R

Source: based on own calculations.

Note: The same as for the table 2, but here GARCH-t stands for GARCH-t(1,1), GARCH-st stands for GARCH-st(1,1), Mean stands for simple average from GARCH-t and CaViaR.

Table 7. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for oil for confidence level equal to 0.99.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	1.54%	10.50	13.28	25.45	Y	1.27%	1.97	2.92	11.38	G	2.17%	12.95	18.36	24.49	R	3.43%	8.41	13.93	72.02	Y
GARCH-t	2.41%	59.53	80.88	490.49	R	0.96%	0.05	1.25	106.88	G	5.80%	136.83	145.95	536.93	R	2.16%	39.45	44.28	346.76	R
GARCH-st	1.01%	0.00	0.58	3.13	G	0.89%	0.36	0.82	10.06	G	1.29%	0.96	2.61	6.28	G	2.16%	4.13	6.42	24.85	Y
QML-GARCH	1.52%	9.64	12.55	24.38	Y	1.29%	2.01	2.92	11.47	G	2.09%	11.41	17.23	23.12	Y	3.45%	8.55	13.93	72.03	Y
CaViaR	1.49%	8.81	8.82	14.83	Y	1.10%	0.27	0.98	5.46	G	2.42%	18.01	18.10	27.26	R	2.60%	6.24	7.95	36.61	Y
Mean	1.44%	7.26	7.28	20.93	Y	0.96%	0.05	0.59	21.48	G	2.58%	21.72	21.76	30.45	R	3.41%	8.24	7.95	37.61	Y
Highest VaR	3.49%	158.07	169.88	460.74	R	1.99%	22.40	22.42	79.80	R	7.00%	194.17	201.30	574.54	R	3.46%	56.90	61.42	306.55	R
Lowest VaR	0.46%	15.51	18.66	17.14	G	0.27%	21.77	21.81	15.72	G	0.89%	0.17	3.18	8.15	G	2.16%	1.02	4.98	40.55	G
CQOM	2.36%	56.00	94.71	707.99	R	1.51%	6.61	11.24	151.64	Y	4.35%	76.99	104.99	671.01	R	2.60%	39.45	44.28	297.04	R
Elastic Net	1.64%	14.24	14.83	27.95	R	0.99%	0.00	0.58	15.67	G	3.14%	36.67	37.11	69.57	R	3.46%	14.19	14.79	46.58	R
LASSO	1.59%	12.31	14.83	120.13	Y	0.79%	1.41	1.78	41.45	G	3.46%	46.41	47.70	137.00	R	3.90%	17.29	17.67	112.70	R
QRF	3.68%	178.14	179.09	390.91	R	3.26%	94.28	94.76	279.50	R	4.67%	89.32	92.69	240.24	R	4.76%	20.58	26.33	125.31	R
GBRM	1.80%	21.93	22.21	34.37	R	1.48%	5.80	7.09	38.64	Y	2.58%	21.72	23.01	40.31	R	2.60%	6.24	7.95	25.84	Y
QRNN	4.43%	267.55	269.31	1197.81	R	3.40%	104.09	104.84	533.39	R	6.84%	186.18	186.44	816.55	R	6.06%	66.27	69.41	542.78	R

Source: based on own calculations.

Note: The same as for the previous table.

Table 8. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for gas for confidence level equal to 0.975.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	2.19%	1.72	1.72	5.62	G	1.78%	6.80	6.81	8.64	G	3.14%	1.93	1.98	7.75	G	5.63%	0.79	1.36	10.87	G
GARCH-t	2.09%	2.99	3.01	6.79	G	1.75%	7.51	7.52	8.66	G	2.90%	0.77	0.77	5.75	G	4.76%	1.58	2.32	10.97	G
GARCH-st	2.45%	0.04	0.13	2.93	G	2.09%	2.10	2.47	2.96	G	3.30%	2.98	3.08	10.15	Y	4.33%	1.49	2.32	11.11	G
QML-GARCH	2.24%	1.22	1.22	4.27	G	1.85%	5.50	5.50	6.19	G	3.14%	1.93	1.98	7.59	G	5.63%	0.79	1.36	11.01	G
CaViaR	2.53%	0.01	0.06	12.09	G	1.99%	3.34	3.88	9.26	G	3.78%	7.28	7.71	20.93	Y	5.19%	1.57	2.32	13.96	G
Mean	2.24%	1.22	1.60	9.07	G	1.82%	6.13	7.02	11.42	G	3.22%	2.43	2.50	10.27	Y	4.76%	0.79	1.36	10.88	G
Highest VaR	2.86%	2.14	2.20	8.51	G	2.37%	0.22	0.30	1.96	G	4.03%	10.04	10.71	24.08	Y	6.49%	2.61	3.52	14.67	Y
Lowest VaR	1.83%	8.48	8.72	12.70	G	1.41%	16.94	17.20	17.60	G	2.82%	0.50	0.50	5.85	G	3.46%	0.79	1.36	10.63	G
CQOM	4.52%	56.49	67.33	232.88	R	4.22%	29.41	43.33	169.76	R	5.23%	29.10	29.21	76.81	R	3.90%	0.63	0.77	10.55	G
Elastic Net	2.45%	0.04	0.84	8.38	G	2.02%	2.90	7.23	11.76	G	3.46%	4.22	7.31	17.85	Y	6.49%	1.79	2.32	10.66	G
LASSO	2.45%	0.04	2.09	13.69	G	2.09%	2.10	6.04	14.24	G	3.30%	2.98	3.08	17.62	Y	7.36%	1.86	2.32	9.07	G
QRF	5.53%	117.24	117.29	357.06	R	5.25%	68.99	69.14	221.09	R	6.20%	49.73	49.74	141.98	R	8.66%	6.88	8.44	18.56	Y
GBRM	3.30%	9.83	12.20	37.10	Y	2.85%	1.38	2.39	12.48	G	4.35%	14.30	15.37	39.06	R	4.76%	2.61	3.22	12.55	Y
QRNN	5.27%	99.63	105.52	453.83	R	5.21%	67.45	70.62	349.00	R	5.39%	32.24	35.09	119.62	R	9.09%	6.88	6.98	64.95	Y

Source: based on own calculations.

Note: The same as for the table 2, but here GARCH-t stands for GARCH-t(1,1), GARCH-st stands for GARCH-st(1,1), Mean stands for simple average from GARCH-t and CaViaR.

Table 9. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for gas for confidence level equal to 0.99.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	1.03%	0.05	0.95	3.12	G	0.79%	1.41	1.78	5.43	G	1.61%	3.94	4.60	9.29	Y	3.46%	2.37	2.60	26.03	Y
GARCH-t	0.67%	5.05	5.43	8.05	G	0.48%	9.84	9.98	10.05	G	1.13%	0.20	0.51	2.29	G	2.16%	0.19	0.27	2.44	G
GARCH-st	0.99%	0.01	0.82	9.81	G	0.75%	1.94	2.27	6.23	G	1.53%	3.03	3.62	13.25	Y	2.09%	1.02	1.17	27.33	G
QML-GARCH	1.08%	0.28	1.26	3.39	G	0.81%	1.48	1.78	5.38	G	1.77%	6.07	6.87	12.93	Y	3.54%	2.45	2.60	26.19	Y
CaViaR	1.05%	0.07	0.95	5.60	G	0.82%	0.98	1.38	5.06	G	1.53%	3.03	3.62	10.03	Y	2.60%	1.02	1.17	9.73	G
Mean	0.84%	1.11	1.70	4.82	G	0.65%	4.07	4.32	8.78	G	1.29%	0.96	1.37	4.27	G	3.46%	0.29	0.27	2.80	G
Highest VaR	1.25%	2.45	3.77	10.26	Y	0.99%	0.00	0.58	2.86	G	1.85%	7.28	8.14	19.35	Y	3.46%	2.37	2.60	26.07	Y
Lowest VaR	0.63%	6.80	7.12	9.37	G	0.41%	13.10	13.20	12.22	G	1.13%	0.20	0.51	2.41	G	2.16%	0.19	0.27	2.70	G
CQOM	4.40%	264.51	268.51	748.76	R	4.67%	209.24	212.38	615.08	R	3.78%	56.92	57.68	156.26	R	2.60%	11.30	12.03	29.88	Y
Elastic Net	3.49%	158.07	167.67	609.97	R	3.43%	106.60	113.48	430.37	R	3.62%	51.57	54.28	217.59	R	3.46%	61.53	61.96	212.30	R
LASSO	1.35%	4.56	11.81	67.51	Y	1.10%	0.27	8.30	44.77	G	1.93%	8.57	9.06	39.17	Y	3.90%	6.24	6.68	26.86	Y
QRF	3.32%	140.58	140.62	422.57	R	2.78%	62.80	62.83	203.90	R	4.59%	86.18	86.24	262.53	R	4.76%	17.29	18.39	63.33	R
GBRM	1.54%	10.50	10.50	39.50	Y	1.23%	1.51	2.04	16.13	G	2.25%	14.56	15.85	44.46	R	2.60%	2.37	2.60	25.28	Y
QRNN	3.78%	Inf	Inf	1181.52	R	4.12%	160.80	160.80	1024.17	R	2.98%	32.11	34.39	194.18	R	6.06%	4.13	4.46	71.51	Y

Source: based on own calculations.

Note: The same as for the previous table.

Table 10. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for copper for confidence level equal to 0.975.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	2.96%	3.39	4.77	19.79	Y	2.61%	0.14	0.14	4.71	G	3.78%	7.28	9.54	30.64	Y	5.63%	3.85	6.69	29.71	Y
GARCH-t	2.67%	0.48	3.05	18.85	G	2.33%	0.34	0.45	2.93	G	3.46%	4.22	7.42	36.25	Y	4.76%	1.58	5.92	35.76	G
GARCH-st	2.65%	0.36	3.04	20.04	G	2.37%	0.22	0.30	3.05	G	3.30%	2.98	6.71	39.01	Y	4.33%	2.61	6.14	31.14	Y
QML-GARCH	2.99%	3.44	4.77	19.46	Y	2.65%	0.16	0.14	4.21	G	3.78%	7.28	9.54	30.72	Y	5.63%	3.85	6.69	42.44	Y
CaViaR	2.57%	0.09	0.62	10.85	G	2.26%	0.69	3.75	9.95	G	3.30%	2.98	6.71	16.07	Y	5.19%	2.71	6.14	51.69	Y
Mean	2.43%	0.09	0.95	10.86	G	2.23%	0.91	3.87	4.65	G	2.90%	0.77	6.12	22.07	G	4.65%	2.55	6.14	40.23	Y
Highest VaR	3.27%	9.27	10.61	29.07	Y	2.88%	1.66	1.74	12.79	G	4.19%	12.09	15.09	45.05	Y	6.49%	3.85	6.69	30.17	Y
Lowest VaR	2.12%	2.64	4.58	13.18	G	1.92%	4.35	6.54	7.42	G	2.58%	0.03	6.99	23.06	G	3.46%	1.58	5.92	40.33	G
CQOM	4.43%	51.67	Inf	960.75	R	2.98%	2.65	13.39	230.53	G	7.81%	92.75	121.91	806.55	R	3.90%	218.89	219.00	703.95	R
Elastic Net	2.62%	0.25	1.65	17.45	G	2.37%	0.22	0.51	9.27	G	3.22%	2.43	6.46	38.25	Y	6.49%	3.85	6.69	25.60	Y
LASSO	2.62%	0.25	1.65	43.06	G	2.37%	0.22	0.51	36.09	G	3.22%	2.43	6.46	30.80	Y	7.36%	5.28	7.53	31.78	Y
QRF	4.91%	77.52	78.43	231.96	R	4.39%	35.03	35.06	121.60	R	6.12%	47.85	49.03	128.29	R	8.66%	10.57	13.88	49.29	Y
GBRM	3.08%	5.33	7.41	33.89	Y	2.95%	2.29	3.56	21.25	G	3.38%	3.57	12.79	40.09	Y	4.76%	3.85	6.69	23.83	Y
QRNN	4.52%	56.49	61.01	421.32	R	4.49%	38.60	39.35	325.86	R	4.59%	17.91	23.66	136.29	R	9.09%	10.57	11.54	78.96	Y

Source: based on own calculations.

Note: The same as for the table 2, but here GARCH-t stands for GARCH-t(1,2), GARCH-st stands for GARCH-st(1,1), Mean stands for simple average from GARCH-st and CaViaR.

Table 11. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for copper for confidence level equal to 0.99.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	1.71%	17.36	21.56	53.94	R	1.44%	5.04	6.26	9.79	G	2.33%	16.25	24.63	98.67	R	3.46%	11.30	15.63	110.23	Y
GARCH-t	1.18%	1.27	10.40	38.94	G	0.93%	0.16	0.67	8.74	G	1.77%	6.07	18.76	79.58	Y	2.16%	4.13	11.96	99.07	Y
GARCH-st	1.08%	0.28	10.67	40.15	G	0.89%	0.36	0.82	8.16	G	1.53%	3.03	18.17	92.27	Y	2.12%	4.11	11.96	99.36	Y
QML-GARCH	1.78%	20.75	24.47	55.48	R	1.54%	7.47	8.88	12.95	G	2.33%	16.25	24.63	98.80	R	3.49%	11.41	15.63	131.62	Y
CaViaR	1.20%	1.62	3.71	27.18	G	1.13%	0.49	1.25	16.74	G	1.37%	1.53	6.99	30.44	G	2.60%	2.37	5.38	82.34	Y
Mean	1.01%	0.00	11.44	45.52	G	0.89%	0.36	0.82	6.08	G	1.29%	0.96	19.09	118.12	G	3.46%	4.13	11.96	107.81	Y
Highest VaR	1.95%	29.58	32.33	73.24	R	1.72%	12.41	14.15	27.03	G	2.50%	19.83	27.24	95.92	R	3.46%	11.30	15.63	110.69	Y
Lowest VaR	0.87%	0.79	5.01	21.97	G	0.75%	1.94	2.27	8.58	G	1.13%	0.20	7.18	38.09	G	2.16%	2.37	5.38	59.22	Y
CQOM	1.85%	24.38	33.04	376.94	R	1.34%	3.04	3.40	84.60	G	3.06%	34.37	42.11	389.48	R	2.60%	31.46	32.44	159.44	R
Elastic Net	1.27%	2.92	7.43	78.15	Y	1.13%	0.49	1.25	53.05	G	1.61%	3.94	12.74	66.78	Y	3.46%	2.37	5.38	26.06	Y
LASSO	1.23%	2.02	2.21	136.46	G	1.06%	0.12	0.78	55.47	G	1.61%	3.94	4.90	112.99	Y	3.90%	2.42	5.38	27.19	Y
QRF	3.63%	173.63	178.33	542.73	R	3.36%	101.61	101.76	317.23	Y	4.27%	74.00	81.26	275.12	R	4.76%	13.24	12.19	81.67	Y
GBRM	1.73%	18.46	20.31	64.03	R	1.48%	5.80	7.09	29.37	G	2.33%	16.25	20.93	56.39	R	2.60%	9.54	12.19	49.56	Y
QRNN	3.56%	165.79	168.16	374.11	R	4.01%	152.20	153.27	351.61	R	2.50%	19.83	21.28	76.59	R	6.06%	17.29	17.67	104.84	R

Source: based on own calculations.

Note: The same as for the previous table.

Due to the lack of a clearly dominant model, both among individual and combined models, it was decided to carry out the Model Confidence Set procedure. Its results are presented in Table 12. For the confidence level of 0.95, individual models dominate Although for example for silver it has been shown that for all crises, as well as for the coronavirus crisis, the best method of forecasting VaR is to use the mean. However, it should be remembered that for this particular asset no model passed the regulatory traffic light test positively, and the forecasts for the average did not pass any of the tests. Also, for copper, the results show the superiority of the forecast combining methods. For the entire assessment period, the mean is the best model, and for all crises, as well as for the current coronavirus crisis, it is best to combine quantile regression forecasts with the use of elastic net regularization. For the confidence level equal to 0.99, forecast combining methods dominate. The lowest VaR turned out to be the best model for all assets for the entire assessment period. For the calm period for gold, silver and copper, the lowest VaR is the best, for oil it is the mean, and for gas it is GARCH-t. For both periods of crises, the lowest VaR turned out to be the best. The first exception is gold, which for crises, including the coronavirus crisis, forecasts turned out to be a better model for GARCH-st. The second exception is gas, for which CQOM is the best method for combining forecasts during a pandemic crisis. The third exception is copper, for which GARCH-st is the best method of forecasting VaR during a pandemic. This analysis showed a significant dominance of the forecast connection methods for the confidence level of 0.99. Interestingly, the simple methods turned out to be the best.

Table 12. The best model for each commodity (rows) and for all periods (columns) for confidence level of 0.975 (upper part of table), and 0.99 (lower part of table) achieved using MCS procedure.

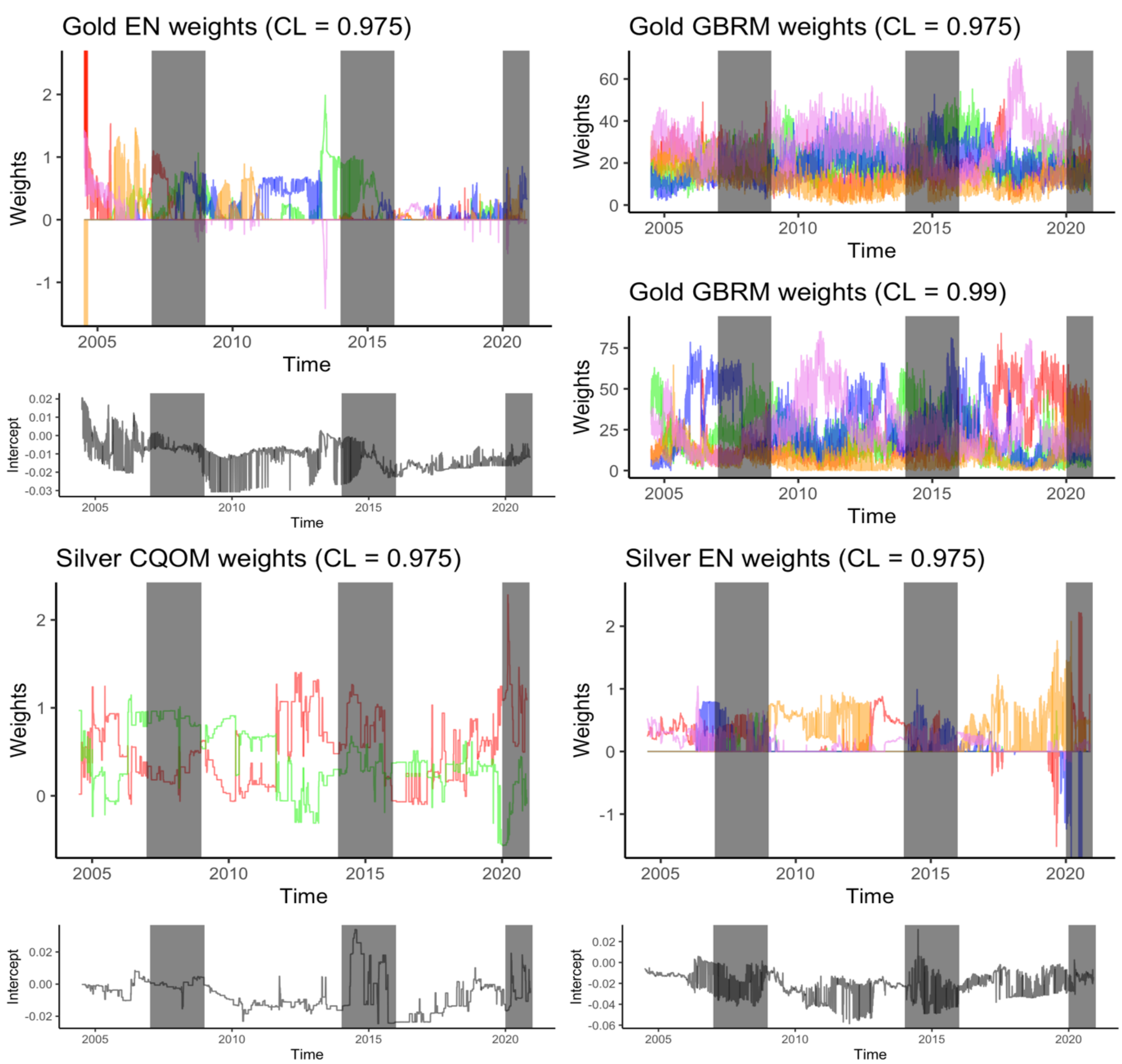
Model	Period I (Whole period)	Period II (All calm periods)	Period III (All crisis periods)	Period IV (COVID period)
Confidence level = 0.025				
Gold	GARCH-t	GARCH-t	GARCH-t	CQOM
Silver	Mean	GARCH-st	Mean	Mean
Oil	GARCH	GARCH	GARCH-st	GARCH-st
Gas	GARCH-st	Highest VaR	GARCH-st	LASSO
Copper	Mean	GARCH	Elastic Net	Elastic Net
Confidence level = 0.01				
Gold	Lowest VaR	Lowest VaR	GARCH-st	GARCH-st
Silver	Lowest VaR	Lowest VaR	Lowest VaR	Lowest VaR

Oil	Lowest VaR	Mean	Lowest VaR	Lowest VaR
Gas	Lowest VaR	GARCH-t	Lowest VaR	CQOM
Copper	Lowest VaR	Lowest VaR	Lowest VaR	GARCH-st

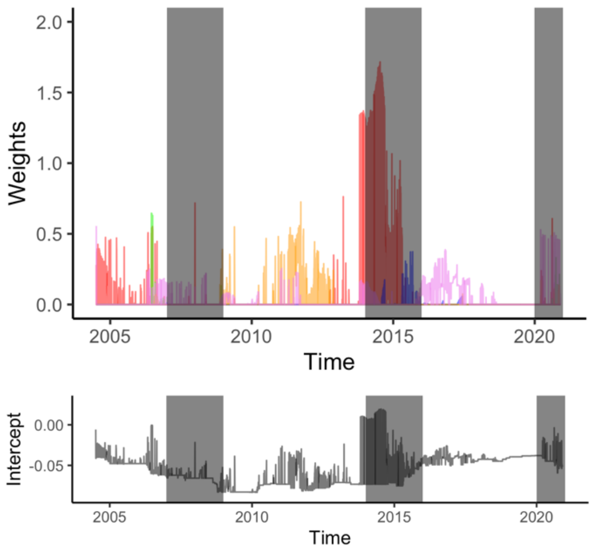
Source: based on own calculations.

To understand where the predominance of given forecast combining methods comes from, one should look at the weights that are assigned to each individual model to create a specific combined forecast. Figure 7 presents charts on which the weights assigned to individual models for selected methods of combining forecasts are depicted. At first glance, it can be seen that in some periods (especially from 2007 to 2009) for both regression methods and CQOM the models assumed weights close to zero or equal to zero, and the forecast was an intercept. Concerning the regularization of elastic net and LASSO for gold, it is difficult to make an unambiguous conclusion. It is only visible that in the periods of crisis, the model assigned greater importance to GARCH-st both for the confidence level of 0.975 and 0.99. For silver CQOM for the confidence level of 0.975 for two crises (from 2014 to 2016 and in 2020) assigned greater weight to the GARCH model, and for the subprime crisis (from 2007 to 2009) it assigned greater importance to the GARCH-t model. For the elastic net and LASSO, during the crisis periods, the model assigned more importance to the GARCH model, and in the period of calm to the QML-GARCH model. This phenomenon was observed regardless of the confidence level. For oil, both for the elastic net and LASSO, it was observed that the CaViaR model prevails in periods of calm. Interestingly for the crisis since 2007, the model did not assign any weights to forecasts, but only regulated the intercept, which may indicate that this crisis was extremely difficult to model for individual models. For gas at the confidence level equal to 0.975, both for the elastic net and LASSO methods, the GARCH-st model was the most preferred model for the calm period. The exception to this dependence is the period from 2017 to 2020. Such a relationship was not observed for the elastic net at the confidence level of 0.99. The results for copper say that there is no single good model for this asset that can be adopted even in times of calm. This method for the period from 2004 to 2007 favours the GARCH model, from 2009 to 2014 it assigns more weight to the CaViar model, and after 2016 to the GARCH-t model. However, the latter dependence is not observed at the confidence level of 0.99. Looking at the graphs for the Quantile Boosting Regression Model, it can be seen that for all commodities in the period of calm, the CaViaR model dominates, except for oil, where different models dominate depending on the period of calm. From 2004 to 2008 it was CaViaR, from 2009 to 2014 it was GARCH, and from 2016 to 2020 it was GARCH-t. The results for the crises are not so clear. The CaViaR model has dominated the crisis since 2007 for oil and gas. For gold, the CaViaR model dominates only during the crisis caused by the coronavirus (from 2020) for a confidence level of 0.975, in other cases, no model seems to be in the lead. For oil, for the confidence level of 0.975, for the crisis from 2014 to 2016, GARCH-t is of the greatest importance, and for the crisis from 2020, CaViaR takes over the leadership role. For gas, for the confidence level of 0.99, for the crisis from 2014 to 2020, CaViaR prevails, while for the crisis from 2020, GARCH prevails. Such an analysis was not performed for silver and copper as this method did not give promising results there. The results indicate that although it is possible to distinguish models which in the given methods of combining forecasts indicate the dominance of one of the individual models over others, in most situations the models are weighted to a comparable extent.

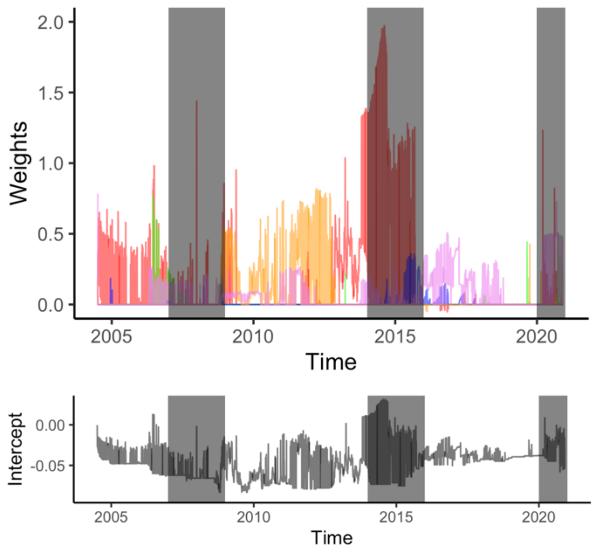
Figure 7. Combing weight for the most promising methods of combining VaR forecasts for all assets for both confidence level (CL) – 0.975 and 0.99.



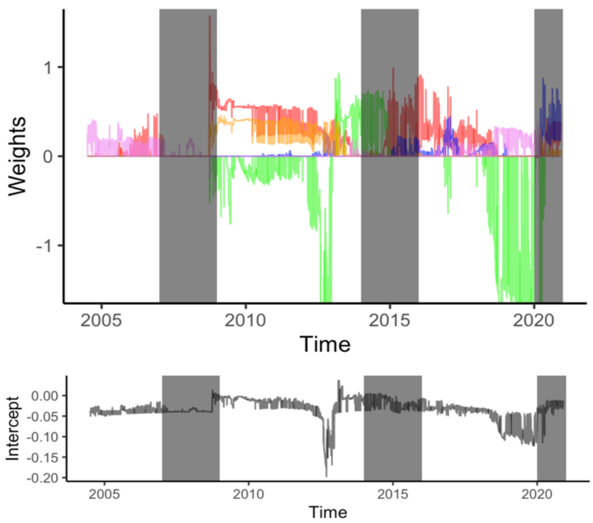
Silver LASSO weights (CL = 0.99)



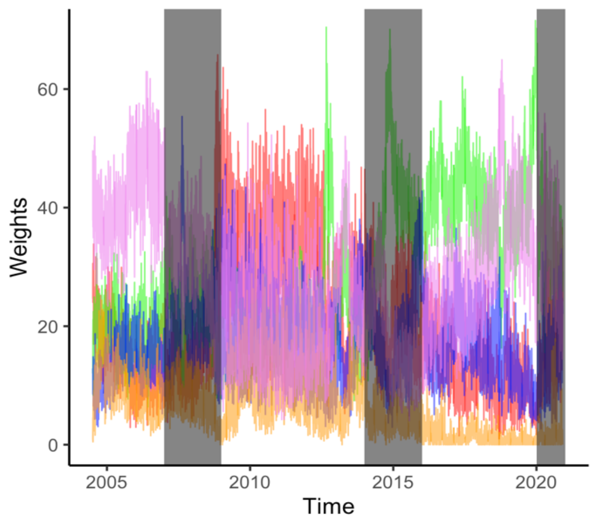
Silver EN weights (CL = 0.99)



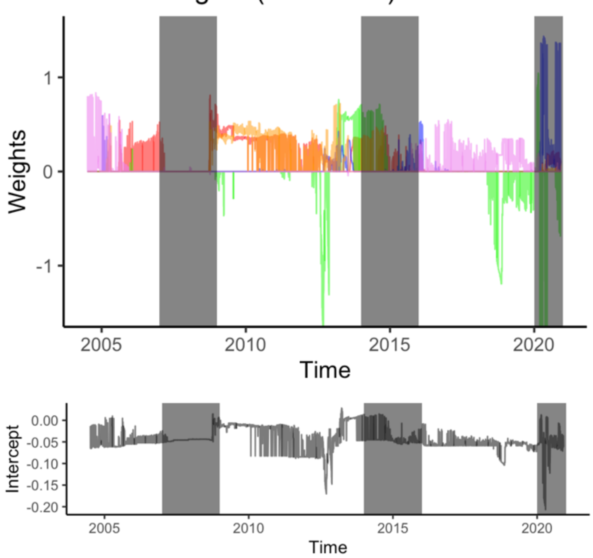
Oil EN weights (CL = 0.975)



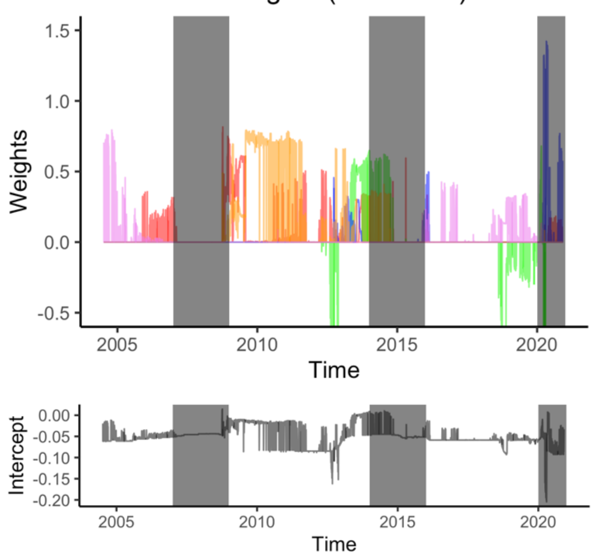
Oil GBRM weights (CL = 0.975)

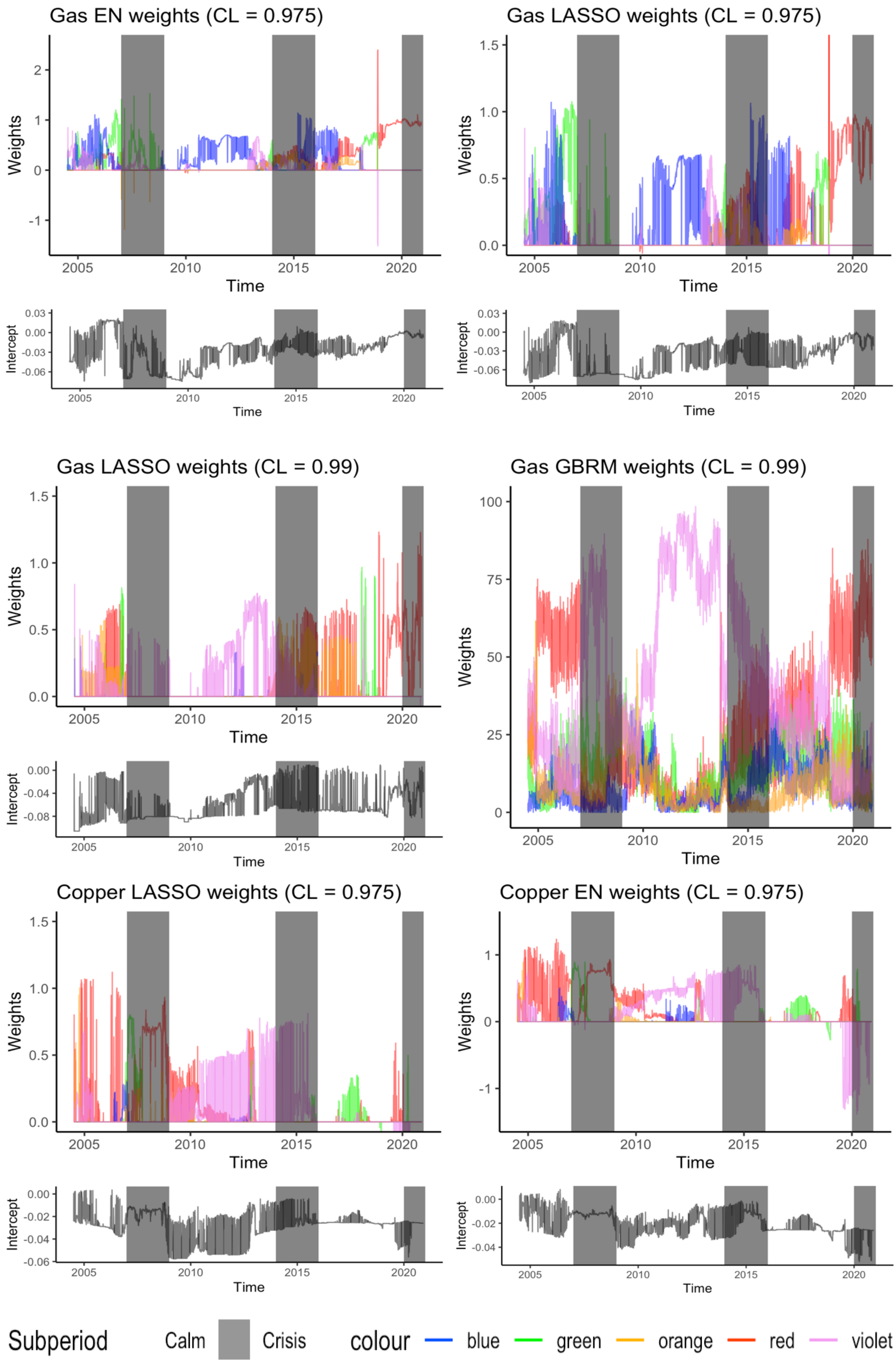


Oil EN weights (CL = 0.99)



Oil LASSO weights (CL = 0.99)





Source: based on own calculations.

Note: Figures above present the results of weights assigned by combining methods: CQOM stands for Conditional Quantile Optimization Method applied for GARCH and CaViaR (described in section 2.3.4), Elastic Net stands for forecast combined using quantile regression with elastic net regularization (described in section 2.3.6), LASSO stands for forecast combined using quantile regression with LASSO regularization (described in section 2.3.5), and GBRM stands for forecast combined using Gradient Boosting Regression Model (described in section 2.3.8). The models shown in the legend are described for each asset in the section 3.1

The first hypothesis about greater accuracy of forecast combining methods in the entire assessment period was fully confirmed for confidence level 0.99, and partially for confidence level 0.975, i.e. for silver and copper. The second hypothesis about greater accuracy of forecast merging models for the calm period was confirmed only on one asset - gas - for the confidence level of 0.975, and almost fully confirmed for the confidence level of 0.99, i.e. only for gas the individual was better model. The third hypothesis regarding the greater accuracy of the forecast connection methods for the crisis periods was only partially confirmed, i.e. for the level of 0.975 for silver and copper, and for the level of 0.99 for all assets except for gold. The fourth hypothesis about the superiority of forecast combining methods over individual modelCaViaR confirmed almost entirely for the confidence level of 0.975 - the exception was an oil, and partially confirmed for 0.99 - the exceptions were gold and copper. The most common forecast combining models, which occurs to be the best, were the lowest VaR and the mean.

4 Conclusions

The study managed to compare the accuracy of individual VaR forecasting models with forecast combining methods for two different confidence levels (0.975 and 0.99), as well as considering the influence of individual forecasts to the combined forecast. The results of the study for the confidence level of 0.975 were, in most cases, only partially in line with the expectations, and for the confidence level of 0.99 they were almost entirely as expected. The results are consistent with the results achieved by other researchers (e.g. Bayer, 2018), as well as predictions and research results for other fields, where the average dominated over other methods of combining forecasts (Timmermann, 2006). Despite the fact that the study also used complicated methods of combining forecasts, in most cases the simplest ones, i.e. the lowest VaR and the average, turned out to be the best. However, regression methods have often shown promising results. It may be an area that should be thoroughly verified by checking whether these methods cannot be improved by changing the parameter estimation window, using different tuning parameters or changing regularization. An interesting fact was the uniqueness of gold, which was the only one that for crises, including the one caused by the coronavirus, completely "resisted" the method of combining forecasts and for it individual models gave the best results. This may be due to its special role in times of crises, where individuals, as well as institutions and countries during crises try to protect their savings there, but this require further research. The main conclusion from the study is that it is worth combining forecasts, however, one should focus on simplicity. The limitation of the study was that it confirmed the results only for commodities. Another disadvantage was backtesting the results for the coronavirus crisis on incomplete data from that period, so the results should not be generalised to application for that period. Another limitation was the adoption of default modelling options using machine learning models. Subsequent research could include testing hypotheses in other markets, using different loss function, combining forecasts from other individual models testing the hypothesis for a pandemic crisis on complete data, and trying to improve both regression and machine learning methods.

Bibliography

Angabini, A., Wasiuzzaman, S., 2011. GARCH Models and the Financial Crisis-A Study of the Malaysian. *The International Journal of Applied Economics and Fin-ance Vol. 5 No*, *3*, 226-236.

Armstrong, J. S., 1989. Combining forecasts: The end of the beginning or the beginning of the end? *International Journal of Forecasting*, *5*(4), 585-588.

Bayer, S., 2018. Combining value-at-risk forecasts using penalized quantile regressions. *Econometrics and statistics*, *8*, 56-77.

Basle Committee on Banking Supervision, 1996. "Supervisory Framework for the Use of 'Backtesting' in Conjunction with the Internal Models Approach to Market Risk Capital Requirements".

Basle Committee on Banking Supervision, 2010. The Basel III Capital Framework: A Decisive Breakthrough. Speech by Hervé Hannoun at BoJ-BIS High Level Seminar on Financial Regulatory Reform: Implications for Asia and the Pacific, Hong Kong SAR.

BCBS (2017), Basel III: Finalising Post-crisis Reforms, December 2017,

Bernardi, M., Catania, L., 2016. Comparison of Value-at-Risk models using the MCS approach. *Computational Statistics*, *31*(2), 579-608.

Bollerslev T., 1986. Generalized Autoregressive Conditional Heteroscedasticity. *Journal of Econometrics. 31(3)*, 307-327.

Bollerslev T., 1987. Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *The Review of Economics and Statistics. 69(3),* 542-547.

Bollerslev, T., Woolridge, J. M., 1992. Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances *Econometric Reviews 11.*

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Buczyński, M., Chlebus, M., 2018. Comparison of semi-parametric and benchmark value-at-risk models in several time periods with different volatility levels. *e-Finanse: Financial Internet Quarterly*, *14*(2), 67-82.

Buczyński, M., Chlebus, M., 2019. Old-fashioned parametric models are still the best. A comparison of Value-at-Risk approaches in several volatility states. *Faculty of Economic Sciences, University of Warsaw Working Papers*, *12*, 297.

Cabedo, J. D., Moya, I., 2003. Estimating oil price 'Value at Risk'using the historical simulation approach. *Energy economics*, *25*(3), 239-253.

Caillault, É. P., Lefebvre, A., and Bigand, A., 2017. Dynamic time warping-based imputation for univariate time series data. *Pattern Recognition Letters*.

Cannon, A. J., 2010. A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes: An International Journal*, *24*(6), 673-685.

Cannon, A. J., 2011. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers & geosciences*, *37*(9), 1277-1284.

Chiu, Y. C., Chuang, I. Y., Lai, J. Y., 2010. The performance of composite forecast models of value-at-risk in the energy market. *Energy Economics*, *32*(2), 423-431.

Christoffersen P., 1998. Evaluating Interval Forecasts. International Economic Review. 39(4), 841–862.

Clemen, R. T., Winkler, R. L., 1986. Combining economic forecasts. *Journal of Business & Economic Statistics*, *4*(1), 39-46.

Danielsson, J., 2013. The new market-risk regulations. *Article, VOX. URL: www. voxeu. org/article/new-market-risk-regulations*.r

Danielsson, J., Morimoto, Y., 2000. *Forecasting extreme financial risk: A critical analysis of practical methods for the Japanese market*. Institute for Monetary and Economic Studies, Bank of Japan.

Dudziński, J., 2016. Ceny w handlu międzynarodowym w drugiej dekadzie XXI wieku. Kierunki zmian i ich czynniki. *International Business and Global Economy*, *35*(2), 249-260.

Duffie, D., Pan, J., 1997. An overview of value at risk. *Journal of derivatives*, *4*(3), 7-49.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of variance of UK inflation. *Econometrica*, 50, 987–1008.

Engle, R.F. S. Manganelli, 1999, CAViaR: Conditional Value at Risk by Quantile Regression, NBER, Working Paper 7341.

Engle R. F., Manganelli S., 2004. CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles*. Journal of Business & Economic Statistics. 22(4),* 367-381.

Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of statistics*, *28*(2), 337-407.

Freund, Y., 1995. Boosting a weak learning algorithm by majority. *Information and computation*, *121*(2), 256-285.

Gardner, M. W., Dorling, S. R., 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14-15), 2627-2636.

Gençay, R., Selçuk, F., Ulugülyağci, A., 2003. High volatility, thick tails and extreme value theory in value-at-risk estimation. *Insurance: Mathematics and Economics*, *33*(2), 337-356.

Giacomini, R., Komunjer, I., 2005. Evaluation and combination of conditional quantile forecasts. *Journal of Business and Economic Statistics*, *23*(4), 416-431.

Gevrey, M., Dimopoulos, I., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3), 249-264.

Glosten, L., Jagannathan, R., Runkle, D., 1993. On the relation between the expected value and the volatility nominal excess return on stocks. Journal of Finance, 46, 1779–1801.

Goodhart, C., 2011. *The Basel Committee on Banking Supervision: A history of the early years 1974–1997*. Cambridge University Press.

González-Rivera, G., Lee, T. H., Mishra, S., 2004. Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *International Journal of forecasting*, *20*(4), 629-645.

Grömping, U., 2009. Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, *63*(4), 308-319.

Halbleib, R., Pohlmeier, W., 2012. Improving the value at risk forecasts: Theory and evidence from the financial crisis. *Journal of Economic Dynamics and Control*, *36*(8), 1212-1228.

Hansen, P. R., Lunde, A., Nason, J. M., 2011. The model confidence set. *Econometrica*, *79*(2), 453-497.

Hardy, Charles O., 1923. Risk and Risk-Bearing, Chicago: University of Chicago Press.

Hastie, T., Tibshirani, R., Friedman, J., 2011. *The elements of statistical learning: data mining, inference, and prediction*. Second ed. Springer.

Haynes, J., 1895. Risk as an economic factor. *The Quarterly Journal of Economics*, *9*(4), 409-449.

Hibon, M., Evgeniou, T., 2005. To combine or not to combine: selecting among forecasts and their combinations. *International journal of forecasting*, *21*(1), 15-24.

Hicks, J. R., 1935. A suggestion for simplifying the theory of money, Economica, 11 (5), 1-19.

Holthausen, D. M., Hughes, J. S., 1978. Commodity returns and capital asset pricing. *Financial Management*, 37-44.

Hsieh, W. W., Tang, B., 1998. Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bulletin of the American Meteorological Society*, *79*(9), 1855-1870.

Huang, H., Lee, T. H., 2013. Forecasting value-at-risk using high-frequency information. *Econometrics*, *1*(1), 127-140.

Ichev, R., Marinč, M., 2018. Stock prices and geographic proximity of information: Evidence from the Ebola outbreak. *International Review of Financial Analysis*, *56*, 153-166.

Jeziorski, P., 2014. Wybrane testy nieobciążoności miar ryzyka na przykładzie value at risk. *Studia Ekonomiczne*, *207*, 110-118.

Jeon, J., Taylor, J. W., 2013. Using CAViaR models with implied volatility for Value-at-Risk estimation. *Journal of Forecasting*, *32*(1), 62-74.

Jorion, P., 2010. Risk Management. *Annual Review of Financial Economics*, Vol. 2, pp. 347-365.

Kuester, K., Mittnik, S., Paolella, M. S., 2006. Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, *4*(1), 53-89.

Kupiec, P., 1995. Techniques for Verifying the Accuracy of Risk Management Models. Journal of Derivatives. 3(2), 73-84.

Laurent, S., Rombouts, J. V., Violante, F., 2012. On the forecasting accuracy of multivariate GARCH models. *Journal of Applied Econometrics*, *27*(6), 934-955.

Leavens, Dickson H., (1945). Diversification of investments, Trusts and Estates, 80 (5), 469-473.

Linsmeier, T. J., Pearson, N. D., 2000. Value at risk. *Financial Analysts Journal*, *56*(2), 47-67.

Liu, H. C., Hung, J. C., 2010. Forecasting S&P-100 stock index volatility: The role of volatility asymmetry and distributional assumption in GARCH models. *Expert Systems with Applications*, *37*(7), 4928-4934.

Marimoutou, V., Raggad, B., Trabelsi, A., 2009. Extreme value theory and value at risk: application to oil market. *Energy Economics*, *31*(4), 519-530.

Makridakis, S., Winkler, R. L., 1983. Averages of forecasts: Some empirical results. *Management science*, *29*(9), 987-996.

Markowitz, Harry, M., 1952. Portfolio Selection, Journal of Finance, 7 (1), 77-91.

McAleer, M., Jimenez-Martin, J. A., Perez Amaral, T., 2010. Has the Basel II Accord encouraged risk management during the 2008-09 financial crisis? *Available at SSRN 1397239*.

McAleer, M., Jiménez-Martín, J. Á., Pérez-Amaral, T., 2013. International evidence on GFC-Robust forecasts for risk management under the Basel Accord. *Journal of Forecasting*, *32*(3), 267-288.

Meinshausen, N., Ridgeway, G., 2006. Quantile regression forests. *Journal of Machine Learning Research*, *7*(6).

Mensi, W., Sensoy, A., Vo, X. V., Kang, S. H., 2020. Impact of COVID-19 outbreak on asymmetric multifractality of gold and oil prices. *Resources Policy*, *69*, 101829.

Phillips, P. C., Yu, J., 2011. Dating the timeline of financial bubbles during the subprime crisis. *Quantitative Economics*, *2*(3), 455-491.

Rossignolo, A. F., Fethi, M. D., Shaban, M., 2012. Value-at-Risk models and Basel capital charges: Evidence from Emerging and Frontier stock markets. *Journal of Financial Stability*, *8*(4), 303-319.

Rossignolo, A. F., Fethi, M. D., Shaban, M., 2013. Market crises and Basel capital requirements: Could Basel III have been different? Evidence from Portugal, Ireland, Greece and Spain (PIGS). *Journal of Banking & Finance*, *37*(5), 1323-1339.

Roy, Arthur D., 1952. Safety first and the holding of assets, Econometrica, 20 (3), 431- 449.

Sharpe, W. F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, *19*(3), 425-442.

Szakmary, A. C., Shen, Q., Sharma, S. C., 2010. Trend-following trading strategies in commodity futures: A re-examination. *Journal of Banking & Finance*, *34*(2), 409-426.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Taylor, J. W., 2020. Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, *36*(2), 428-441.

Terui, N., Van Dijk, H. K., 2002. Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting*, *18*(3), 421-438.

Timmermann, A., 2006. Forecast combinations. *Handbook of economic forecasting*, *1*, 135-196.

Tsay, R. S., 2005. *Analysis of financial time series* (Vol. 543). John wiley & sons.

Tse, Y.,2016. Asymmetric volatility, skewness, and downside risk in different asset classes: Evidence from futures markets. *Financial Review*, *51*(1), 83-111.

Touzani, S., Granderson, J., Fernandes, S., 2018. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533-1543.

Youssef, M., Belkacem, L., Mokni, K., 2015. Value-at-Risk estimation of energy commodities: A long-memory GARCH–EVT approach. *Energy Economics*, *51*, 99-110.

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl