HSC Research Report

HSC/14/12

Evaluating the performance of VaR models in energy markets

Saša Žiković¹
Rafał Weron²
Ivana Tomas Žiković¹

¹Faculty of Economics, University of Rijeka, Croatia

²Institute of Organization and Management, Wrocław
University of Technology, Poland

Hugo Steinhaus Center
Wrocław University of Technology
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
<http://www.im.pwr.wroc.pl/~hugo/>

Evaluating the performance of VaR models in energy markets

Saša Žiković, Rafał Weron, and Ivana Tomas Žiković

Abstract In this paper we analyze the relative performance of 13 VaR models using daily returns of WTI, Brent, natural gas and heating oil one-month futures contracts. After obtaining VaR estimates we evaluate the statistical significance of the differences in performance of the analyzed VaR models. We employ the simulation-based methodology proposed by Žiković and Filer [19], which allows us to rank competing VaR models. Somewhat surprisingly, the obtained results indicate that for a large number of different VaR models there is no statistical difference in their performance, as measured by the Lopez size adjusted score. However, filtered historical simulation (FHS) and the semiparametric BRW model stand out as robust and consistent approaches that – in most cases – significantly outperform the remaining VaR models.

1 Introduction

Fossil fuels are commodities that are the primary source of energy for electricity production, transportation and industrial production. Their importance and multiplicative effect on almost all branches of the economy dictates the necessity for deeper insight into the ways of measuring and protecting against risks stemming from changes in their prices. Hedging against price changes of fossil fuels is not only important for the financial industry, where energy serves primarily as an investment opportunity, but also to energy buyers and sellers which have to protect themselves from raising and falling energy prices. In order to be able to hedge against such risks the first step is to correctly evaluate the market risk of a particular energy com-

Saša Žiković, Ivana Tomas Žiković
Faculty of Economics, University of Rijeka, Croatia, e-mail: szikovic@efri.hr

Rafał Weron
Institute of Organization and Management, Wrocław University of Technology, Poland, e-mail: rafal.weron@pwr.edu.pl

modity. Value at Risk (VaR) is an essential tool for this purpose. Within energy markets, VaR can be used to quantify the price risk of energy commodities associated with the desired probability level. The use of the VaR risk models in order to forecast market risk is widely used in the financial industry, where profits and losses are calculated daily, a process known as marking-to-market. VaR methodology is not yet widely used in non-financial companies, one of the reasons being the lack of regulatory requirement to daily mark-to-market their holdings. Although energy companies do not report their daily changes in value, VaR methodology is appropriate for these non-financial companies since their revenues and expenditures often fluctuate daily due to significant volatility in prices of their inputs and outputs.

The volatility of energy commodities has been widely studied in the energy economics literature. In particular, Mohammadi and Su [14] consider weekly oil spot prices in 11 markets and compare the forecasting accuracy of four GARCH-class models under two loss functions using the DM test. Wei et al. [15] conclude that no single model is superior for WTI and Brent across different loss functions. Kang et al. [9] evaluate the out-of-sample performance of four GARCH-class models under two loss functions using the DM test. The CGARCH and FIGARCH models are found to capture the long-memory volatility of crude oil markets and allow to obtain superior performance compared to that of the GARCH and IGARCH models. Agnolucci [1] studies the market volatility of WTI futures and finds extensions of GARCH models with asymmetric effects and different error distributions to outperform implied volatility models in terms of predictive accuracy. Studying electricity spot prices, Weron [16] finds that deseasonalized (with respect to the weekly patterns and long-term trends) returns exhibit heavy or semi-heavy tails, that can be captured well by generalized hyperbolic or Levy-stable laws.

In the risk management context, Hung et al. [8] show that in general the assumption of a Gaussian distribution leads to an underestimation of risk and highlight the importance of selecting the appropriate distribution in a GARCH context. They further find that the Value-at-Risk (VaR) of crude oil and oil products is better captured by fat-tailed distributions. Marimoutou et al. [12] find that extreme value based models perform well in the oil markets and that they offer a major improvement over the traditional (non-parametric and parametric) methods. Žiković and Vlahinić-Dizdarević [20] analyze WTI, natural gas and coal prices at the 95, 99 and 99.5% confidence levels and conclude that the best performance across all of the tested confidence levels is achieved for the EVT-GARCH model. Bunn et al. [2] show that a structural linear quantile regression model outperforms skewed GARCH- t and CAViaR models regarding the accuracy of out-of-sample forecasts of VaR. Overall the findings reported in the literature on the subject of VaR estimation for energy commodities are mixed. A situation that is similar to the one found in the electricity price forecasting literature [17].

Apart from known VaR problems that refer to sub-additivity and tail events, there is also an often overlooked systemic problem with risk model comparison and ranking. Backtesting VaR or expected shortfall (ES) figures yields only a comparison for a single realization of the underlying data generating process. The VaR literature is vast but it rarely addresses the issue of statistical significance in performance

of different models. Sometimes it may be only a matter of chance that for some commodity over a particular time horizon a certain model performs better than the competing ones.

The goal of this paper is to evaluate the performance of a wide array of VaR models for WTI, Brent, natural gas (NG) and heating oil (HO) one-month futures contracts during the ongoing global economic crisis, which includes periods of both backwardation and contango. After obtaining the VaR estimates we proceed to evaluate the statistical significance of the differences in performance of the analysed VaR models. We employ the methodology for comparing VaR model performance presented in [19], allowing us to consistently rank competing VaR models. The obtained information is equally important for the risk management purposes, policy making as well as for the pricing of structured commodity derivatives.

The rest of the paper is organized as follows. In Section 2 we present the VaR ranking procedure that we use in our analysis. Next in Section 3 we outline the data, methodology and basic assumptions of the VaR models. VaR backtesting and rankings are presented in Section 4. Finally, in Section 5 we wrap up the results and conclude.

2 Comparing and ranking VaR models

In the risk literature there are a number of methods that test the hypothesis whether a certain model is better than some other model, such as the superior predictive ability (SPA) of Hansen [6], the equal predictive ability (EPA) of Diebold and Mariano [5] and the reality check (RC) of White [18]. The question of interest in all of these tests is whether any alternative forecast is better than the benchmark forecast, or equivalently, whether the best alternative forecasting model is better than the benchmark. This question can be addressed by testing the null hypothesis that the benchmark is not inferior to any alternative forecast. Using such tests is useful for a forecaster who wants to explore whether a better forecasting model than the model currently being used is available. After a search over several alternative models, the relevant question is whether the observed excess performance by an alternative model is significant or not.

To implement forecast evaluation, it is necessary to specify a loss function, and a number of different loss functions have been proposed in the literature. Lopez [11] suggested a size-adjusted loss function that is simple, symmetrical and intuitive:

$$C_t = \begin{cases} 1 + (L_t - VaR_t)^2 & \text{if } L_t > VaR_t, \\ 0 & \text{if } L_t \leq VaR_t. \end{cases}$$

This loss function allows for the sizes of tail losses to influence the final rating of a VaR model. A VaR model that generates higher tail losses would generate higher values under this size adjusted loss function than a VaR model that generates lower

tail losses, *ceteris paribus*. The ranking procedure proposed in [19] consists of five steps:

1. Fitting an ARMA-GARCH model to the time series in order to obtain independent and identically distributed (IID) observations. Then estimating the empirical CDF of each time series (applying it to the non-tail regions of the distribution) with a Gaussian kernel. This allows to smooth the CDF, eliminating the staircase pattern of the sample CDF.
2. Finding the upper and lower thresholds such that a percentage of the residuals are reserved for each tail and fitting the amount by which those extreme residuals in each tail fall beyond the associated threshold to a parametric GPD.
3. Generating N simulated paths for the residuals from the obtained semi-parametric distribution (each path is T observations long) and adding the ARMA-GARCH model to the residuals to obtain $N \times T$ simulated time series returns.
4. Calculating VaR for each of the $N \times T$ simulated returns for each VaR model and Lopez scores for each of the N VaR-simulated return pairs, for each VaR model.
5. Comparing if the mean values of the Lopez scores for different VaR models are significantly different from each other. For this purpose the Kruskal-Wallis test is employed.

The Kruskal-Wallis test is a nonparametric version of the classical one-way analysis of variance (ANOVA), and tests the null hypothesis that all samples are drawn from the same population, or equivalently, from different populations with the same distribution [7]. Unlike a standard one-way ANOVA, the test requires only that the measurements come from a continuous, but not necessarily a normal distribution. The test is based on an analysis of the variance using the ranks of the data values, not the data values themselves. The limitation of this approach is the assumption that the description of the central mass and the tails of the process distribution are adequate, i.e. that the underlying process is well described by the visible realization. This is not an unusual or strong assumption and is made in all the models that are used in practice. By simulating the data generating process in the above described way, stochastic randomness is allowed in the data set.

3 Data and methodology

We analyze the performance of 13 VaR models: a simple moving average (VCV), the RiskMetrics approach, historical simulation (HS 250 and HS 500; the number indicates the number of past observations used to compute VaR), mirrored historical simulation (MHS 250 and MHS 500), BRW semi-parametric simulation with decay factors $\lambda = 0.97$ and 0.99 , a GARCH model, filtered historical simulation (FHS), the unconditional EVT approach using the generalized Pareto distribution (GPD) and the conditional EVT approach (EVT GARCH). In the analysis we employ the log-daily returns of WTI, Brent, natural gas (NG – Henry Hub and New York Harbour) and heating oil (HO – NYMEX) one-month futures contracts. Returns were

collected from the Bloomberg website for the period January 1, 1995 through July 1st, 2014, which includes a period of the latest surge and fall in prices of energy commodities. VaR figures are calculated for a one-day ahead long position and the 99% confidence level.

The analysed VaR models are tested by using: the Kupiec [10] test, the Christoffersen [3] test of independence (IND) and the Lopez [11] size adjusted test. As pointed out in [19], the Christoffersen unconditional coverage (UC) test is problematic since it gives a distorted image of the performance of VaR models. Because it is chi-square distributed with one degree of freedom, deviations from the test's expected value that occur on the conservative side (i.e. with number of exceedances lower than their expected value) are penalized more severely. This characteristic is not compatible with risk-averse or risk-neutral assumptions. Thus, from the regulatory standpoint, the Kupiec [10] binomial test is preferable to the Christoffersen [3] UC test, because it is more desirable to have positive than negative deviations. The same logic extends to the Christoffersen conditional coverage (CC) test since it is the sum of the UC and IND tests. In our two-stage backtesting procedure, the best performing VaR model must first satisfy both the Kupiec [10] and Christoffersen [3] independence (IND) tests and then provide the minimal deviation from the expected value of losses by minimizing the Lopez [11] error statistics.

4 Backtesting results

To secure the same out-of-the-sample backtesting period for all of the examined energy commodities, the out-of-the-sample data sets are formed by removing the 1000 most recent observations from each commodity price series. The remaining observations are used to calculate GPD tail parameters, VaR starting values and calibrate volatility. The length of the tail-loss data set used for backtesting depends on the number of errors generated by each VaR model. Data from all the commodities shows leptokurtosis, asymmetry and significant heteroskedasticity, with pronounced autoregression that shows periodicity.

Based on the Akaike and Bayesian information criterion the best GARCH representation of volatility with GED and Student t distribution was used to capture the dynamics of data-generating processes of each commodity. The asymmetry parameter in the GARCH model was significantly different from zero only for WTI. The asymmetry parameter, which controls the asymmetric impact of positive and negative shocks on conditional variance, indicates significantly higher conditional volatility after positive shocks, i.e. price increases. Estimation of the tail index is crucial in applying EVT models which are directly linked to threshold value u that defines the level above which returns are considered extreme. The threshold value for each index was determined by comparing the Hill estimator with the mean excess plot and the quantile-quantile (QQ) plot [4]. The same procedure of estimating the threshold value was also performed on IID innovations required for the implementation of the EVT-GARCH model [13]. The mean excess and QQ plots, Hill

Table 1 Lopez [11] size adjusted scores for 13 VaR models and four commodities (WTI, Brent, natural gas – NG, and heating oil – HO) at the 99% confidence level are reported in columns 2-5. The observation period covers 1000 days from July 2010 to July 1st, 2014. Scores for models not satisfying the Kupiec [10] test and/or the Christoffersen [3] independence test at the 10% significance level are underlined. The lowest Lopez scores, i.e. the smallest deviations from the expected values, are emphasized in bold. Rankings of VaR models based on the Lopez [11] size adjusted scores are reported in columns 6-10. They were obtained for 2000 simulated trajectories and a forecasting horizon of 1000 days. Lower values indicate better VaR models. If multiple models have the same score there is no statistically significant difference between them.

VaR model	WTI	Brent	NG	HO	WTI	Brent	NG	HO	Total
	<i>Lopez scores</i>				<i>Rankings</i>				
HS 250	3.13	1.10	<u>7.27</u>	2.16	3	2	4	2	3
HS 500	-1.88	-4.92	2.21	-2.87	2	2	2	2	2
MHS 250	-1.93	-5.93	-1.90	-3.87	2	3	1	2	2
MHS 500	-3.92	-6.94	-4.92	-4.89	3	3	4	3	4
BRW =0.97	<u>8.15</u>	<u>7.15</u>	<u>9.26</u>	<u>10.18</u>	4	3	4	4	5
BRW =0.99	-0.92	-1.92	3.20	0.13	1	1	3	1	1
Normal VCV	<u>4.17</u>	<u>6.16</u>	<u>5.26</u>	<u>1.17</u>	5	5	4	4	7
Risk Metrics	<u>10.22</u>	<u>13.20</u>	2.12	<u>11.18</u>	5	5	2	4	6
GARCH	<u>7.17</u>	<u>9.18</u>	0.09	<u>6.17</u>	5	5	1	4	5
HW	<u>8.15</u>	<u>10.16</u>	<u>9.17</u>	<u>16.20</u>	5	5	4	5	8
FHS	-0.92	-0.90	-6.98	-2.88	1	1	3	1	1
EVT GARCH	-5.95	-3.93	-3.96	-0.87	3	2	2	1	2
GPD	-10.00	-10.00	-10.00	-10.00	5	5	4	4	7

estimator and maximum likelihood estimates (MLE) all show that tail indexes for all commodities are equal to or greater than zero, implying fat tails and that the GPD belongs to the Gumbel domain of attraction. This suggests that the normal distribution is not appropriate for describing tail events.

Columns 2-5 in Table 1 present our out-of-sample VaR model ranking according to minimal Lopez size adjusted scores for WTI, Brent, natural gas (NG) and heating oil (HO) at the 99% confidence level. We conduct a two stage ranking since we allow a model to be ranked only after it satisfies both the Kupiec [10] and the Christoffersen [3] independence test.

Satisfactory performance with regards to coverage and independence criteria is recorded for nonparametric models (HS, MHS), as well as FHS and extreme value based approaches. Very weak performance is recorded for GARCH, VCV, RiskMetrics and Hull-White models. Regarding the closeness of fit performance – which we measure by the Lopez [11] score – FHS, EVT GARCH and HS/MHS models provide the best fit.

In the next step, we apply the methodology discussed in Section 2 to test whether there is any statistically significant difference in the performance of the tested VaR models. The data is simulated based on the distribution of returns in the period July 2010 – July 2014. For each commodity 2000 simulated trajectories of 1000 days are considered. The Lopez size adjusted score yields the best performing model as the one that is the closest to zero. After obtaining 2000 Lopez size adjusted scores for each VaR model and for each commodity, we apply the Kruskal-Wallis test to

comapre the differences among the tested VaR models. The results are reported in columns 6-10 in Table 1.

When the simulated mean value of a given VaR model lies outside the 95% confidence bands of all the other tested VaR models, it is ranked according to its relative performance. If a model is not significantly different from the other models it shares the same ranking as the VaR models not significantly different from it. Looking at the ranking results presented in columns 6-10 in Table 1, we see that for a large number of different models there is no statistically significant difference in their performance measured by our loss function.

When considering the overall performance for the tested commodities, the best performing VaR models that are statistically different from other tested models are the filtered historical simulation (FHS) and the semiparametric BRW ($\lambda = 0.99$) model. These models are followed by the conditional EVT GARCH model and two simple nonparametric models (HS and MHS). The worst performance, measured by the distance from expected losses, be it positive or negative is recorded for Hull-White, VCV and unconditional GPD models. Although we tested 13 VaR models, we see that for all the tested commodities the number of statistically different models never surpasses five, and in the case of natural gas all the VaR models are concentrated in just four groups.

5 Conclusions

Our empirical study shows that for a large number of different VaR models there is no statistical difference measured by the Lopez [11] size adjusted score as the loss function. Overall, statistically significant top performers are the filtered historical simulation (FHS) and the semiparametric BRW ($\lambda = 0.99$) model. As our loss function is the size adjusted Lopez score, these VaR models provide the closest fit to the actual level of risk encountered in the analyzed energy commodities. Simpler parametric models, such as the VCV, Hull-White and RiskMetrics, were the worst performers in our VaR comparison. Poor performance with regard to our loss function is also recorded for a very conservative unconditional GPD model. Unlike the aforementioned parametric models this is due to the fact that it constantly overestimated the level of risk that was recorded in the four commodities during the 2010-2014 period, i.e. it was overly conservative.

It is also interesting to note that although historical simulation based models are clearly theoretically inferior to EVT models in VaR estimation, their empirical track record is impressive. This finding may suggest that during the analyzed period there were a large number of extreme events that allowed the simpler nonparametric models to correctly asses the true level of VaR. Advanced models based on conditional EVT and FHS as well as the very simple nonparametric models, such as mirrored historical simulation, yield very robust and consistent results. Models that span between these two extremes are not successful in energy markets and yield inconsistent risk estimates.

Acknowledgements This work was supported by the Croatian Science Foundation under Grant number IP-2013-11-2203.

References

1. Agnolucci P (2009) Volatility in crude oil futures: A comparison of the predictive ability of GARCH and implied volatility models. *Energy Econ* 31:316–321.
2. Bunn D, Andresen A, Chen D and Westgaard S (2013) Analysis and forecasting of electricity price risks with quantile factor models. London Business School Working Paper. Available from http://www.ceem-dauphine.org/assets/dropbox/Derek_BUNN.pdf.
3. Christoffersen PF (1998) Evaluating interval forecasts. *Int Econ Rev* 39(4):841–862.
4. Danielsson J and de Vries C (1997) Tail index and quantile estimation with very high frequency data. *J Emp Finan* 4:241–257.
5. Diebold FX and Mariano R (1995) Comparing predictive accuracy. *J Bus Econ Statist* 13:253–263.
6. Hansen PR (2005) A test for superior predictive ability. *J Bus Econ Statist* 23(4):365–380.
7. Hollander M and Wolfe DA (1999) *Nonparametric Statistical Methods*. Wiley, Hoboken, NJ.
8. Hung JC, Lee MC and Liu HC (2008) Estimation of Value-at-Risk for energy commodities via fat-tailed GARCH models. *Energy Econ* 30(3):1173–1191.
9. Kang SH, Kang SM and Yoon SM (2009) Forecasting volatility of crude oil markets. *Energy Econ* 31:119–125.
10. Kupiec P (1995) Techniques for verifying the accuracy of risk management models. *J Derivatives* 3:73–84.
11. Lopez AJ (1998) Methods for evaluating Value-at-Risk estimates. Federal Reserve Bank of New York, *Econ Policy Rev* 2:3–17.
12. Marimoutou V, Raggad B and Trabelsi A (2009) Extreme value theory and Value at Risk: Application to oil market. *Energy Econ* 31:519–530.
13. McNeil AJ and Frey R (2000) Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *J Emp Finan* 7:271–300.
14. Mohammadi H and Su L (2010) International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models. *Energy Econ* 32(5):1001–1008.
15. Wei Y, Wang Y and Huang D (2010) Forecasting crude oil market volatility: Further evidence using GARCH-class models. *Energy Econ* 32(6):1477–1484.
16. Weron R (2009) Heavy-tails and regime-switching in electricity prices. *Math Meth Oper Res* 69(3):457–473.
17. Weron R (2014) Electricity price forecasting: A review of the state-of-the-art with a look into the future. *Int J Forecasting* 30(4):1030–1081.
18. White H (2000) A reality check for data snooping. *Econometrica* 68:1097–1126.
19. Žiković S and Filer RK (2013) Ranking of VaR and ES models: Performance in developed and emerging markets. *Czech J Econ Finan* 63(4):327–359.
20. Žiković S and Vlahinić-Dizdarević (2011) Similarities between expected shortfall and Value at Risk: Application to energy markets. *Int J Management Cases* 13(3):386–399.

HSC Research Report Series 2014

For a complete list please visit <http://ideas.repec.org/s/wuu/wpaper.html>

- 01 *Diffusion and adoption of dynamic electricity tariffs: An agent-based modeling approach* by Anna Kowalska-Pyzalska, Katarzyna Maciejowska, Katarzyna Sznajd-Weron and Rafał Weron
- 02 *A review of electricity price forecasting: The past, the present and the future* by Rafał Weron
- 03 *Merging quantile regression with forecast averaging to obtain more accurate interval forecasts of Nord Pool spot prices* by Jakub Nowotarski and Rafał Weron
- 04 *A note on using the Hodrick-Prescott filter in electricity markets* by Rafał Weron and Michał Zator
- 05 *Fundamental and speculative shocks, what drives electricity prices?* by Katarzyna Maciejowska
- 06 *Modeling consumer opinions towards dynamic pricing: An agent-based approach* by Anna Kowalska-Pyzalska, Katarzyna Maciejowska, Katarzyna Sznajd-Weron and Rafał Weron
- 07 *Electricity price forecasting: A review of the state-of-the-art with a look into the future* by Rafał Weron
- 08 *Modelling price spikes in electricity markets – the impact of load, weather and capacity* by Rangga Handika, Chi Truong, Stefan Trück and Rafał Weron
- 09 *Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging* by Katarzyna Maciejowska, Jakub Nowotarski and Rafał Weron
- 10 *Probabilistic load forecasting via Quantile Regression Averaging of independent expert forecasts* by Tao Hong, Katarzyna Maciejowska, Jakub Nowotarski and Rafał Weron
- 11 *Forecasting the occurrence of electricity price spikes in the UK power market* by Paweł Maryniak and Rafał Weron
- 12 *Evaluating the performance of VaR models in energy markets* by Saša Žiković, Rafał Weron and Ivana Tomas Žiković