Forecasting the Term Structure of Volatility of Crude Oil Price Changes

Ercan Balaban* Shan Lu†

University of Aberdeen Business School, Dunbar Street, Aberdeen, AB24 3QY, UK.

February 2016

Abstract

This is a pioneering effort to test the comparative performance of two competing models for out-of-sample forecasting the term structure of volatility of crude oil price changes employing both symmetric and asymmetric evaluation criteria. Under symmetric error statistics, our empirical model using the estimated growth factor of volatility through time is overall superior, and it beats in most cases the benchmark model of the square-root-of-time ($\sqrt{T}$) for holding periods between one and 250 days. Under asymmetric error statistics, if over-prediction (under-prediction) of volatility is undesirable, the empirical (benchmark) model is consistently superior. Relative performance of the empirical model is much higher for holding periods up to fifty days.

JEL classification: C53; G12; G17.

Keywords: Volatility term structure; Square-root-of-time rule; Forecasting; Forecasting evaluation; Oil prices.

*Email: Ercan.Balaban@abdn.ac.uk
†Email: r02sl14@abdn.ac.uk

1. Introduction

It is well documented that volatility of financial asset returns is out-of-sample predictable (Poon and Granger, 2003), with extensions to commodities markets, including energy markets [Wang and Wu (2012) and their Table 1; Haugom, Langeland, Molnár, and Westgaard (2014)]; albeit there is no consensus on universal superiority of a specific forecasting technique; and model performance is both investment horizon and loss-function dependent. However, to the best of our knowledge, there is no explicit published research on out-of-sample forecasting of the term strcuture of volatility, and subsquent forecast evaluation (Balaban, 1998); although there is primarily substantial theoretical and empirical analysis of the term structure of volatility implied by option prices and credit markets, and of time scaling of risk and volatility. It should be noted that among other areas of its use, most notably; traditional option pricing models such as the Black-Scholes formula, and the regulatory market risk management using the Value-at-Risk technique under a chosen confidence level advocate the use of time scaling of volatility based on the square-root-of-time ($\sqrt{T}$) model. It asserts that volatility in terms of return standard deviation grows proportionally by the square root of time, allowing to compound volatility for longer horizons based on shorter-horizon in-sample estimates of volatility. However, there is ample empirical evidence against the validity of this $\sqrt{T}$ rule, which assumes identically, and independently returns and a normal distribution (Engle, 2004). Among these, Diebold, Hickman, Inoue, and Schuermann (1998) show that the common practice of converting one-day volatility estimates to T-day estimates by scaling by the $\sqrt{T}$ model is inappropriate, and produces overestimates of the variability of long-horizon volatility. Danielsson and Zigrand (2006) find that by applying the $\sqrt{T}$ model to the time scaling of quantiles of return distributions, risk is underestimated at an increasing rate as the extrapolation horizon is extended, if the probability of a jump increases, or the confidence level is raised. It is concluded that in particular, as the scaling horizon increases, the bias introduced by the $\sqrt{T}$ rule grows at a rate faster than time. By examining 47 markets globally, Wang, Yeh, and Cheng (2011) find

the $\sqrt{T}$ rule to be lenient, in that it generally yields downward-biased 10-day and 30-day Value-at-Risk estimates; particularly in Eastern Europe, Central-South America, and the Asia Pacific.

This is a pioneering effort aiming at filling the empirical gap in forecasting the out-of-sample term structure of volatility using the crude oil price changes. To this end, we first empirically estimate the growth rate of volatility through time for holding periods one to 250 days. Then, we construct volatility forecasts for each holding period using both the empirical model and the $\sqrt{T}$ rule. The predictive ability of both models is compared employing both symmetric and asymmetric error statistics.

2. Data and methodology

Daily spot prices for WTI-crude oil are used to calculate the T-day realized standard deviation (σ_T^R) of T-day non-overlapping continuously compounded changes between January 1986 and November 2014, with 7542 observations. The full sample is equally divided into the estimation and forecast periods. The empirical growth factor (β) of volatility through time is obtained in the estimation period by

$$\ln(\sigma_T^R) = \alpha + \beta \ln(T) + \varepsilon_T \tag{1}$$

where holding period $T = 1, 2, 3, ..., 250$ days. The out-of-sample T-days volatility forecasts (σ_T^F) in the forecasting period are generated by using estimated $\hat{\beta}$ from (1) for the empirical model; while the benchmark $\sqrt{T}$ model uses $\beta = 0.5$. Both models use one-day realized standard deviation (σ_1^R) in the estimation period as a base to produce out-of-sample forecasts.

$$\sigma_T^F = \sigma_1^R \times T^{\hat{\beta}} \tag{2}$$

The out-of-sample predictive ability of both models is evaluated by, first through the

standard and most widely used symmetric error statistics; namely, the mean error (ME), the mean absolute error (MAE), the root mean squared error (RMSE), and the mean absolute percentage error (MAPE); and then through the relatively less employed asymmetric error statistics; namely, the mean mixed error for under-predictions and over-predictions, MME(U) and MME(O) as in Brailsford and Faff (1996), Balaban, Bayar, and Faff (2006), and Balaban (2004); and the mean logarithmic error (MLE) as in Pagan and Schwert (1990).[1] Note that the MME(U) and MME(O) penalize under-predictions and over-predictions more heavily, respectively. The T-day error statistic, forecast volatility minus realized volatility for horizon T in the forecast period, is given by $E_T = \sigma_T^F - \sigma_T^R$.

$$MME(U) = \frac{\sum_{T=1}^{n} E_T D_{T,O} + \sum_{T=1}^{n} |E_T|^{0.5} D_{T,U}}{n} \tag{3}$$

$$MME(O) = \frac{\sum_{T=1}^{n} E_T^{0.5} D_{T,O} + \sum_{T=1}^{n} |E_T| D_{T,U}}{n} \tag{4}$$

where $D_{T,O} = 1$ if $\sigma_T^F > \sigma_T^R$, i.e., σ_T^F is an over-prediction, and 0 otherwise; and similarly, $D_{T,U} = 1$ if $\sigma_T^F < \sigma_T^R$, i.e., σ_T^F is an under-predicion, and 0 otherwise. Note that $|E_T| < 0$, and n is the number of forecasts.

$$MLE = \frac{\sum_{T=1}^{n} \left(\ln \frac{\sigma_T^F}{\sigma_T^R} \right)^2}{n} \tag{5}$$

All evaluation criteria are applied for the full sample, where $n = 250$ and holding period $T = 1, 2, 3, ..., 250$; and for five different sub-sets of holding periods, with $n = 50$ in each.

3. Empirical results

The estimated growth factor (β) of volatility of crude oil price changes through time in the estimation period January 1986 to mid-June 2000 for holding periods from one day to 250 days is 0.448, which is statistically smaller than 0.5 as assumed by the benchmark model

[1]The utility-based and profit-based loss functions are out of the scope of the paper.

of $\sqrt{T}$ for the term structure of volatility (). The corresponding figure in the forecasting period mid-June 2000 to November 2014 is 0.470; again, statistically smaller than 0.5.[2] This implies that the volatility in the crude oil market grows slower than the theoretical model assumes. Hence, the use of the $\sqrt{T}$ model may overall result in over-prediction of volatility, compared to what it should be.

Table 2 presents the results for the symmetric error statistics, which do not differentiate between the positive and negative prediction errors. The ME shows that the empirical (benchmark) model under-predicts (over-predicts), on the average, in all subset of holding periods. Under the MAE, RMSE and MAPE criteria, it is consistently reported that the empirical model is overall superior for holding periods from one day to 250 days; and holding period sub-groups one to 50 days, 151-200 days, and 201-250 days. Most notably, and calculated as $(1 - \text{Relative})$; the empirical model is 65.8%, 58.3% and 68.4% better than the benchmark model for the one to 50-day sub-period under the MAE, RMSE and MAPE, respectively. In this group, it is superior in 44 horizons out of fifty. On the other hand, for holding periods 51-100 days, and 101-150 days; the benchmark model is better than the empirical model by between a minimum of 4.9% (51-100, MAPE) and a maximum of 25.5% (101-150, RMSE) depending on the error criterion chosen. Note that the empirical (benchmark) model is superior for 59.4% (40.6%) of the total 250 horizons. As expected, the prediction errors increase with the increasing length of holding period.

Table 3 shows the results for the asymmetric error statistics. Recall that the MME(U) and MME(O) penalize under-predictions and over-predictions more heavily, respectively. If under-prediction of volatility is undesirable, the benchmark should be chosen regardless of the length of the holding period group. Its relative performance over the empirical model is better by between a minimum of 35.4% for holding periods 151-200 days and a maximum of 78.9% for holding periods 51-100 days. Note that under the MME(U) the benchmark

[2]The correlation coefficient between the estimation and forecast periods volatility series is 79.8%, significant at < 0.01. The equality of mean, median and standard deviation of natural log of volatility series in two periods cannot be rejected at any conventional level.

(empirical) model is superior for 69.1% (30.9%) of the time if all horizons included. Most notably, the $\sqrt{T}$ rule is superior for 46 horizons within 51 to 100-day group. If, however, over-prediction of volatility is undesirable, the empirical model should be chosen regardless of the length of the holding period. Its relative performance over the benchmark model is better by between a minimum of 45% for holding periods 151-200 days and a maximum of 78.8% for holding periods one to 50 days. On the average, the empirical model is superior for 79.5% of the time if all horizons included. Within one to 50-day group, the empirical model is always superior if over-prediction is undesirable.

Under another asymmetric criterion, the MLE, a proportional measure; the empirical model is overall superior by 28.9% for holding periods one to 250 days. Most notably, it is superior by 83.5% within one to 50-day horizon. It is also superior for the two longer holding periods 151-200 days, and 201-250 days over the benchmark model by 35.6% and 51.3%, respectively. The benchmark model is better than its empirical competitor by between 32.1% (51-100 days) and 43.9% (101 to 150 days).

Note that under the MLE criterion, the empirical (benchmark) model is superior for 57.4% (42.6%) of the time in case of all horizons. Within one to 50-day group, the empirical model is better for 44 holding periods.

4. Conclusion

This is a pioneering effort to compare out-of-sample predictive ability of two competing models for the term structure of volatility of crude oil price changes between mid-June 2000 and November 2014 for holding periods one to 250 days, and five different subsets of 50-day holding periods. Using symmetric error statistics, it is found that our empirical model is overall superior over the benchmark of model of $\sqrt{T}$ model, which tends to over-predict volatility. The superiority of the empirical model is much clearer in the short horizon, or for holding periods one to 50 days; being its relative performance, on the average, 65% better.

The benchmark model consistently generates better results if under-prediction of volatility is undesirable, and this is valid regardless of the length of holding period group. The empirical model is overall superior if volatility over-prediction is undesirable, and in particular, it is better for all horizons within one to 50-day holding period.

Our research contributes to the empirical literature on time scaling of volatility by providing evidence against the $\sqrt{T}$ rule in an out-of-sample forecasting framework. We suggest practical use of empirical growth rate to compound volatility for longer horizons; and reinforce that any T-day volatility should ideally be estimated based on appropriate T-day volatility time series. Our results have important implications, particularly, for financial and real options pricing, regulatory market risk management using Value-at-Risk models, and management of short?long portfolio positions; which each and all require out-of-sample volatility forecasts in the short, medium and long horizons in the crude oil markets. In addition, differentiation of forecasting performance with special reference to under-prediction and over-prediction of volatility is of particular importance to options pricing and trading; as volatility being the only unobservable input with a positive impact on both call and put options. Further research could and should look into the performance of both models in these areas through using utility-based and profit-based evaluation criteria. In addition, we suggest that this forecasting exercise should be extended to analyze the volatility term structure of other financial time series.

References

Balaban, E., 1998. Forecasting the term structure of volatility of stock market returns and foreign exchange changes., working Paper, Central Bank of the Republic of Turkey.

Balaban, E., 2004. Comparative forecasting performance of symmetric and asymmetric conditional volatility models of an exchange rate. Economics Letters 83, 99–105.

Balaban, E., Bayar, A., Faff, R., 2006. Forecasting stock market volatility: Further international evidence. The European Journal of Finance 12, 171–188.

Brailsford, T., Faff, R., 1996. An evaluation of volatility forecasting techniques. Journal of Banking and Finance 20, 419–438.

Danielsson, J., Zigrand, J., 2006. On time-scaling of risk and the square-root-of-time rule. Journal of Banking and Finance 30, 2701–2713.

Diebold, F., Hickman, A., Inoue, A., Schuermann, T., 1998. Scale models. Risk 11, 104–107.

Engle, R., 2004. Risk and volatility: Econometric models and financial practice. The American Economic Review 94, 405–420.

Haugom, E., Langeland, H., Molnár, P., Westgaard, S., 2014. Forecasting volatility of the u.s. oil market. Journal of Banking and Finance 47, 1–14.

Pagan, A., Schwert, G., 1990. Alternative models for conditional stock volatility. Journal of Econometrics 45, 267–290.

Poon, S., Granger, C., 2003. Forecasting volatility in financial markets: A review. Journal of Economic Literature 41, 478–539.

Wang, J., Yeh, J., Cheng, N., 2011. How accurate is the square-root-of-time rule in scaling tail risk: A global study. Journal of Banking and Finance 35, 1158–1169.

Wang, Y., Wu, C., 2012. Forecasting energy market volatility using garch models: Can multivariate models beat univariate models? Energy Economics 34, 2167–2181.

Table 1: **Empirical growth factor (β) for term structure of volatility**

Period	α	β	R^2	Wald (χ^2)
Estimation	-3.591***	0.448***	0.939	50.729***
	-106.590[a]	61.681		
Forecast	-3.678***	0.470***	0.867	6.646***
	-67.979	40.274		

Notes: The Wald test gives the χ^2 statistic for the difference of β from 0.5.
*** Significance at 1% level.
[a] t-statistic is shown below each coefficient.

Table 2: **Symmetric error statistics**

Holding period (days)	ME		MAE			RMSE			MAPE			% superior[a]
	Empirical	Benchmark	Empirical	Benchmark	Relative	Empirical	Benchmark	Relative	Empirical	Benchmark	Relative	
1-250	-0.0236	0.0366	0.0392	0.0457	0.857	0.0540	0.0643	0.840	0.1451	0.2073	0.700	59.4
1-50	-0.0028	0.0161	0.0055	0.0161	0.342	0.0075	0.0180	0.417	0.0477	0.1507	0.316	87.8
51-100	-0.0249	0.0196	0.0252	0.0216	0.855	0.0298	0.0251	0.840	0.1184	0.1127	0.951	44.9
101-150	-0.0451	0.0183	0.0494	0.0378	0.765	0.0619	0.0461	0.745	0.1684	0.1516	0.900	38.8
151-200	-0.0272	0.0525	0.0571	0.0696	0.820	0.0703	0.0835	0.841	0.1914	0.2906	0.659	63.3
201-250	-0.0178	0.0763	0.0587	0.0834	0.703	0.0698	0.1031	0.676	0.1995	0.3310	0.603	67.3

Notes: The highlighted model is superior. Mean error (ME), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). Relative is the ratio between the actual error statistic of a model and that of the worst performing model.

[a] Percent of time where the empirical model is superior compared to the benchmark model.

Table 3: **Asymmetric error statistics**

Holding period (days)	MME(U)				MME(O)				% of over-predictions		MLE			
	Empirical	Benchmark	Relative	% superior[a]	Empirical	Benchmark	Relative	% superior[b]	Empirical	Benchmark	Empirical	Benchmark	Relative	% superior[b]
1-250	0.1415	0.0658	0.465	69.1	0.0712	0.1695	0.420	79.5	29.6	83.6	0.0387	0.0544	0.711	57.4
1-50	0.0454	0.0161	0.355	53.1	0.0260	0.1224	0.212	100.0	44.0	100.0	0.0036	0.0219	0.165	87.8
51-100	0.1455	0.0308	0.211	91.8	0.0284	0.1289	0.221	85.7	8.0	88.0	0.0223	0.0152	0.679	40.8
101-150	0.1893	0.0783	0.414	83.7	0.0647	0.1391	0.465	65.3	16.0	70.0	0.0534	0.0300	0.561	38.8
151-200	0.1625	0.1050	0.646	55.1	0.1145	0.2082	0.550	69.4	40.0	74.0	0.0580	0.0901	0.644	63.3
201-250	0.1649	0.0987	0.598	59.2	0.1223	0.2490	0.491	83.7	40.0	86.0	0.0560	0.1149	0.487	61.2

Notes: Mean mixed error for under-predictions (MME(U)), mean mixed error for over-predictions (MME(O)), root mean squared error (RMSE), and mean logarithmic error (LE). Relative is the ratio between the actual error statistic of a model and that of the worst performing model.

[a] Percent of time where the benchmark model is superior compared to the empirical model.

[b] Percent of time where the empirical model is superior compared to the benchmark model.