Performance Evaluation of Prediction Models Under Multiple Criteria:

An application on crude oil prices volatility forecasting models

Jamal Ouenniche
Business School, University of Edinburgh
29 Buccleuch Place, Edinburgh, EH8 9JS, UK
jamal.ouenniche@ed.ac.uk

Bing Xu
School of Management and Languages, Heriot-Watt University
Mary Burton Building, Edinburgh, EH14 4AS, UK
b.xu@hw.ac.uk

Kaoru Tone
National Graduate Institute for Policy Studies
7-22-1 Roppongi, Minato-ku, Tokyo 106-8677, Japan
tone@grips.ac.jp

Abstract: With the increasing number of quantitative models available to forecast the crude oil prices and its volatility, the assessment of the relative performance of competing models becomes a critical task. So far, competing forecasting models are compared to each other using a single criterion at a time, which often leads to different rankings for different criteria – a situation where one cannot make an informed decision as to which model performs best when taking all criteria into account. In order to overcome this methodological problem, we proposed a multidimensional framework based on Data Envelopment Analysis models to rank order competing forecasting models.

Keyword: Forecasting crude oil prices' volatility, performance evaluation, data envelopment analysis (DEA), commodity and energy markets.

1. INTRODUCTION

The design of quantitative models for forecasting continuous variables in a wide range of application areas has attracted the attention of a large number of academics and professionals for some time; however, the performance evaluation of competing forecasting models has not received as much attention. Nowadays, although most published research involve using several performance criteria and measures to compare models, the performance evaluation exercise remains of a unidimensional nature; that is, models are ranked by a single measure and typically the obtained rankings are conflicting. Therefore, one cannot make an informed decision as to which model performs better under several criteria and their measures. In order to illustrate the problem with the current unidimensional approach, we shall use the literature on forecasting crude oil prices' volatility as an example.

Oil is an important source of energy that drives modern economies and large swings in its price can place a substantial adverse impact on both oil importers and exporters. For example, higher oil prices may lead to lower aggregate demand and production outputs, induce

inflationary tendencies and higher interest rates for importing countries; whereas a sustained decline in oil prices supports the so-called "resource curse" hypothesis for commodity abundant emerging economies. Therefore, a proactive knowledge of future movements of oil prices and their volatility can lead to better decisions in various areas such as macroeconomic policy making, risk management, options pricing, and portfolio management.

Recent dramatic surges and declines in oil prices before and after the global financial crisis has been a catalyst for an increased attention on studying the nature of oil prices' volatility and their determinants, or to propose better volatility forecasting models. With the increasing number of models available to forecast the volatility of crude oil prices, despite the fact that the assessment of the relative performance of competing forecasting models becomes a critical task, it has not attracted as much attention as it deserves. To be more specific, our survey of the literature revealed that most studies tend to report inconsistent results about the performance of specific forecasting models in that some models perform better than others with respect to a specific criterion but worse with respect to other criteria – see, for example [1-3]. In this paper, we overcome this methodological issue by proposing a slacks-based context-dependent DEA (CDEA) framework for assessing the relative performance of competing forecasting models [e.g., 4-10]. Although all DEA models could be used to classify competing forecasting models into efficient and inefficient ones and rank them according to their scores, our proposed approach are motivated by the following reasons. First, in many applications such as the ranking of forecasting models, the choice of an orientation is irrelevant. Second, under the variable returns-to-scale assumption, input-oriented scores can be different from output-oriented ones, which may lead to different rankings. Third, radial DEA models could only take account of technical efficiency and ignore potential slacks in inputs and outputs and thus may over-estimate efficiency scores. Furthermore, most DEA models cannot differentiate between efficient decision making units (DMUs) as they all receive a score of 1. The super-efficiency DEA models allow one to do so, but radial super-efficiency DEA models maybe infeasible for some efficient ones and would lead to unresolved ties. In addition, the reference set changes from one efficient DMU evaluation to another, which in some contexts might be viewed as "unfair" benchmarking.

The reminder of this paper is organized as follows. In Section 2, we describe the proposed slacks-based context-dependent DEA framework to evaluate the relative performance of competing forecasting models. Section 3 discusses how one might adapt the proposed framework to evaluate competing forecasting models for crude oil prices volatility. In Section 4, we present and discuss our empirical results. Section 5 concludes the paper.

2. A SLACKS-BASED CDEA MODEL FOR ASSESSING FORECASTING MODELS

In this paper, we propose a slacks-based CDEA framework to assess the relative performance of competing forecasting models. The proposed framework is a three-stage process which could be summarized as follows:

Stage 1 – Returns-to-scale (RTS) Analysis: Perform RTS analysis to find out whether to solve a DEA model under constant returns-to-scale (CRS) conditions, variable returns-to-scale (VRS) conditions, increased returns-to-scale (IRS) conditions, or decreased returns-to-scale (DRS) conditions – see [11] for details.

Stage 2 – Classification of DMUs: Use the following algorithm to partition the set of DMUs into several levels

of best-practice frontiers or evaluation contexts, say L :

Initialization Step

Initialize the performance level counter ℓ to 1 and the set of *DMUs* to evaluate at level ℓ, say J^{ℓ}, to $\{DMU_k, k=1,...,n\}$.

Use the *relevant DEA model* to evaluate J^{ℓ} and set the ℓ th-level best-practice frontier E^{ℓ} accordingly; that is, $E^{\ell} = \left\{k \in J^{\ell} \middle| \textit{Efficiency Score } \rho_k^{\ell} = 1\right\}$.

Exclude the current performance level best-practice frontier E^{ℓ} from the set of *DMUs* to evaluate next; that is, set $J^{\ell+1} = J^{\ell} - E^{\ell}$, increment ℓ by 1 and proceed to the iterative step.

Iterative Step

While $J^{\ell} \neq \varnothing$ **Do**

{

Use the relevant DEA model to evaluate J^{ℓ}, set the ℓ th-level best-practice frontier E^{ℓ} accordingly, set $J^{\ell+1} = J^{\ell} - E^{\ell}$, and increment ℓ by 1;

}

---

where the relevant DEA model to use is the slacks-based measure (SBM) model [12]:

$$\begin{aligned} \text{Min} \quad & \rho_k^{\ell} = \left(1 - \frac{1}{m}\sum_{i=1}^{m}\frac{s_{i,k}^{-}}{x_{i,k}}\right) \Big/ \left(1 + \frac{1}{s}\sum_{r=1}^{s}\frac{s_{r,k}^{+}}{y_{r,k}}\right) \\ \text{s.t.:} \quad & \sum_{j\in J^{\ell}} \lambda_j x_{i,j} + s_{i,k}^{-} = x_{i,k}; \forall i \\ & \sum_{j\in J^{\ell}} \lambda_j y_{r,j} - s_{r,k}^{+} = y_{r,k}; \forall r \\ & \lambda_j \geq 0, \forall j \in J^{\ell}; s_{i,k}^{-} \geq 0, \forall i; s_{r,k}^{+} \geq 0, \forall r \end{aligned} \qquad (1)$$

where the i^{th} input and r^{th} output of $DMU_j (j=1,...,n)$ are denoted by $x_{i,j}(i=1,...,m)$ and $y_{r,j}(r=1,...,s)$, respectively, λ_j is the weight assigned to DMU_j in constructing its ideal benchmark, $s_{i,k}^{-}$ and $s_{r,k}^{+}$ are slack variables associated with the first and the second sets of constraints, respectively, and ρ_k^{ℓ} denotes the SBM efficiency score of DMU_k achieved at performance level ℓ. If the optimal value of $\rho_k^{\ell} = 1$, then DMU_k is classified as efficient; otherwise DMU_k is classified as inefficient. Note that model 1 above is solved as it is if stage 1 reveals that the CRS conditions hold; otherwise, one would have to impose one of the following additional constraints depending on whether VRS, IRS, or DRS conditions prevail, respectively:

$$\sum_{j\in J^{\ell}} \lambda_j = 1; \sum_{j\in J^{\ell}} \lambda_j \geq 1; \sum_{j\in J^{\ell}} \lambda_j \leq 1 \qquad (2)$$

Obviously, once DMUs have been partitioned into L efficient frontiers with different levels of performance, one could rank order them from best to worst starting with 1st-level efficient frontier DMUs as best and ending with the L^{th}-level efficient frontier DMUs as worst. Note that ties might exist between DMUs on the same efficient frontier and the next stage is designed to break those ties.

Stage 3 – Break Efficiency Ties: First, for each efficient frontier $E^{\ell}(\ell = 2,...,L)$, compute *relative progress scores* δ_k^1s with respect to the best evaluation context[1], E^1, by solving the following model for each $DMU_k \in E^{\ell}$ and rank order DMUs on efficient frontier E^{ℓ} according to the values of these scores:

$$\begin{aligned} \text{Min} \quad & \delta_k^1 = \left(1 - \frac{1}{m}\sum_{i=1}^{m}\frac{t_{i,k}^{-}}{x_{i,k}}\right) \Big/ \left(1 + \frac{1}{s}\sum_{r=1}^{s}\frac{t_{r,k}^{+}}{y_{r,k}}\right) \\ \text{s.t.:} \quad & \sum_{j\in E^1} \lambda_j x_{i,j} \geq x_{i,k} - t_{i,k}^{-}; \forall i \\ & \sum_{j\in E^1} \lambda_j y_{r,j} \leq y_{r,k} + t_{r,k}^{+}; \forall r \\ & \lambda_j \geq 0, \forall j \in E^1; t_{i,k}^{-} \geq 0, \forall i; t_{r,k}^{+} \geq 0, \forall r \geq 0 \end{aligned} \qquad (3)$$

where $t_{i,k}^{-}$ (respectively, $t_{r,k}^{+}$) denotes the amount by which input i (respectively, output r) of DMU_k should be decreased (respectively, increased) to reach the efficient frontier corresponding to evaluation context E^1. Second, for DMUs belonging to the best efficient

[1] The rationale behind this choice is to set a common global target for all lower level efficient frontiers for the sake of fairness in benchmarking.

frontier E^1, compute *relative attractiveness scores* γ_k^2s with respect to the second best evaluation context[2], E^2, by solving the following model for each $DMU_k \in E^1$ and rank order DMUs on the best efficient frontier according to the values of these scores:

$$\begin{aligned} \text{Max} \quad & \gamma_k^2 = \left(1 - \frac{1}{m}\sum_{i=1}^{m}\frac{t_{i,k}^+}{x_{i,k}}\right) \Big/ \left(1 + \frac{1}{s}\sum_{r=1}^{s}\frac{t_{r,k}^-}{y_{r,k}}\right) \\ \text{s.t.:} \quad & \sum_{j\in E^2}\lambda_j x_{i,j} \le x_{i,k} + t_{i,k}^+; \forall i \\ & \sum_{j\in E^2}\lambda_j y_{r,j} \ge y_{r,k} - t_{r,k}^-; \forall r \\ & \lambda_j \ge 0, \forall j \in E^2; t_{i,k}^+ \ge 0, \forall i; t_{r,k}^- \ge 0, \forall r \end{aligned} \tag{4}$$

where $t_{i,k}^+$ (respectively, $t_{r,k}^-$) denotes the amount by which input i (respectively, output r) of $DMU_k \in E^1$ should be increased (respectively, decreased) to reach the frontier corresponding to evaluation context E^2.

In the next section, we use the proposed procedure to rank order competing forecasting models of crude oil prices' volatility and report on our empirical findings.

3. ADAPTING SLACKS-BASED CDEA FRAMEWORK FOR ASSESSING THE RELATIVE PERFORMANCE OF COMPETING FORECASTING MODELS

In order to adapt the proposed slacks-based CDEA as a multidimensional framework for the relative performance evaluation of competing forecasting models, two main decisions need to be made; namely, the choice of DMUs, and the choice of relevant inputs and outputs. Hereafter, we shall briefly report on how these decisions are made in this article – the reader is referred to [7, 9] for detailed descriptions of forecasting models and performance metrics.

First, DMUs are volatility forecasting models. In our survey of the literature on crude oil prices' volatility forecasting, time series models tend to be the popular ones. We have included the following fourteen time series models that turned out to be valid for our performance evaluation exercise; namely, Random Walk (RW); Historical Mean (HW); Simple Moving Average with averaging periods of 20 and 60 – SMA20 and SMA60; Auto Regressive Moving Average – ARMA(1,1); Auto Regressive with order 1 and 5 - AR(1) and AR(5); Generalized Auto Regressive Conditional Heteroscedasticity models (GARCH(1,1)); GARCH-in-Mean (GARCH-M(1,1)); Exponential GARCH (EGARCH (1,1)); Threshold GARCH (TGARCH(1, 1)); Power ARCH (PARCH(1,1)); Component GARCH (CGARCH(1,1)).

Second, inputs and outputs are the relevant performance criteria, along with their measures, to be used for assessing forecasting models. Our review of the literature on forecasting the volatility of crude oil prices has revealed that three performance criteria have typically been used; namely, *goodness-of-fit*, *biasedness*, and *correct sign*. Note that depending on the application context, the data features, and the decision makers' preferences as to how to penalize large, small, positive, and negative errors, different metrics could be used. In this study, measures of *biasedness* and *goodness-of-fit* are used as input, whereas measures of *correct sign* are used as output. Note that the choice of our inputs (respectively, outputs) is motivated by the principle of "the less the better" (respectively, "the more the better"). Note also that we have chosen to consider several measures for each criterion to find out about the robustness of multidimensional rankings with respect to different measures. To be more specific, *Goodness-of-fit* is measured by one of the following metrics: MSE, Mean Squared Volatility Scaled Error (MSVolScE), MAE, Mean Absolute Volatility Scaled Error (MAVolScE), Mean Mixed Error Under-estimation penalized (MMEU)

[2] The rationale behind this choice is to compare the most efficient DMUs with those that have the closest performance.

and Mean Mixed Error Over-estimation penalized (MMEO); *biasedness* is measured by one of the following metrics: ME or Mean Volatility Scaled Error (MVolScE); and the *correct sign* is measured by Percentage of correct direction change predictions (PCDCP).

4. EMPIRICAL INVESTIGATION AND RESULTS

In this study, we focus on WTI crude oil daily spot prices and our data covers the period ranging from January 2nd 1986 to May 28th 2010 resulting in a total of 6,157 observations. As crude oil prices are level non-stationary, in the literature there is a tendency to study their level stationary equivalent; namely, returns. We compute daily WTI crude oil returns R_t. Since volatility is not directly observable, we use daily squared returns (R_t^2) as a proxy of volatility – the reader is referred to [7, 9] for discussions on different volatility proxies. Note that all chosen volatility forecasting models are tested out-of-sample and the specific implementation we performed is the one with rolling origin and fixed window.

Our RTS analysis revealed that VRS conditions hold for our dataset and therefore models 1, 3 and 4 are augmented with the following constraint: $\sum_{j\in J'} \lambda_j = 1$.

Table 1 provide the unidimensional rankings of fourteen forecasting models of crude oil prices' volatility based on 9 measures of 3 criteria: *biasedness*, *goodness-of-fit* and *correct sign* – this is a typical output presented by most existing forecasting studies (see for example, [1-3]). These unidimensional rankings are devised as follows: models are ranked from best to worst using the relevant measure of each of the criteria under consideration. Notice that different criteria led to different unidimensional rankings, which provides evidence of the problem resulting from the use of a unidimensional approach in a multi-criteria setting as discussed in Section 1. For example, CGARCH(1,1) outperforms SMA20 on measures of goodness-of-fit based on squared errors, whereas SMA20 performs better with respect to the biasedness criterion, as measured by both Mean Error (ME) and Mean Volatility-Adjusted or Scaled Errors (MVolScE), and with respect to the correct sign criterion, as measured by Percentage of correct direction change predictions (PCDCP). In order to remedy to these mixed performance results, one would need to a single ranking that takes account of multiple criteria, which we provide using the proposed DEA framework.

Table 2 summarizes efficient frontiers with different performance levels, denoted by E^t. As the same single measures of correct sign are used throughout, in the rest of the paper, we only put measures of the goodness-of-fit criterion and biasedness criterion in the tables to differentiate between results. In Table 2, forecasting models belonging to the first-level efficient frontier performs better than those belonging to the second-level efficient frontier, models belonging to the second-level efficient frontier performs better than those belonging to the third-level efficient frontier, and so on. For example, under the set of performance measures {ME, MMEU and PCDCP}, model 3 (i.e., SMA20) and model 14 (i.e., CGARCH (1, 1)) outperform model 5 (i.e., SES). These results suggest that the best and the worst efficient frontiers are insensitive to adjusting biasedness measures for volatility. Note that any rankings based on these efficient frontiers would lead to a large number of ties.

In order to break these ties, we use relative progress and attractiveness scores obtained by solving models 3 and 4, respectively, which result in the multidimensional rankings provided in Table 3 where models are ranked from best to worst based on these relative scores. Notice, for example, that the unidimensional ranking corresponding to ME, MAE PCDCP are different, on one

hand, and have ties, on the other hand, as compared to the multidimensional ranking corresponding to {ME, MAE, PCDCP} where ties have been resolved. In general, multidimensional rankings seem to have less or no ties – see multidimensional rankings corresponding to the remaining combinations of measures in Table 3. In addition, multidimensional rankings generally differ from unidimensional ones whenever these later ones are different, which confirms that the proposed multidimensional framework provides a valuable tool to apprehend the true nature of the relative performance of competing forecasting models.

Table 1: Unidimensional Rankings of Competing Forecasting Models

	Measures	Ranked from Best to Worst
Biasedness	ME	3 → 5 → 11 → 9 → 14 → 10 → 12 → 4 → 13 → 6 → 2 → 8 → 7 → 1
	MVolScE	3 → 5 → {9, 10, 11, 12, 14} → 4 → {6, 13} → 2 → 8 → 7 → 1
Goodness-of-Fit	MAE	8 → 5 → 6 → 3 → {9, 10} → 11 → 12 → 7 → 14 → 13 → 4 → 1 → 2
	MAVolScE	{5, 8} → {3, 6} → {9, 10, 11, 12} → {7, 14} → 13 → 4 → 1 → 2
	MSE	14 → 13 → 10 → 12 → 11 → 9 → 5 → 3 → 6 → 4 → 8 → 7 → 2 → 1
	MSVolScE	14 → 13 → 10 → {11, 12} → 9 → 5 → 3 → 6 → 4 → 8 → 7 → 2 → 1
	MMEU	14 → 3 → 5 → 10 → 13 → 12 → 9 → 11 → 4 → 6 → 8 → 7 → 2 → 1
	MMEO	1 → 2 → 7 → 8 → 6 → 11 → 12 → 9 → 13 → 10 → 14 → 4 → 5 → 3
Correct Sign	PCDCP	3 → 5 → {10, 12} → 9 → 14 → 4 → 13 → 6 → 8 → 11 → {1, 7} → 2

*[1]RW; [2] HM; [3]SMA20; [4]SMA60; [5]SES; [6]ARMA (1, 1); [7]AR (1); [8]AR (5); [9]GARCH (1, 1); [10]GARCH-M(1, 1); [11]EGARCH (1, 1); [12]TGARCH (1, 1); [13]PARCH (1, 1); [14]CGARCH(1,1)

Table 2: Efficient frontiers with different performance levels

Efficient Frontiers	ME & MAE	ME &MAVolScE	ME & MSE; ME & MSVolScE	ME & MMEU	ME & MMEO
E^1	{3,5,8}	{3,5}	{3,5,14}	{3,14}	{1,2,3,5,6,8,11}
E^2	{6,9,10,11}	{6,8,9,10,11}	{9,10,11,13}	{5}	{7,9,1012}
E^3	{12,14}	{12,14}	{12}	{9,10,11}	{13,14}
E^4	{4,7,13}	{4,7,13}	{4,6}	{12,13}	{4}
E^5	{1,2}	{1,2}	{2,8}	{4}	
E^6			{7}	{6}	
E^7			{1}	{2,8}	
E^8				{7}	
E^9				{1}	

Panel B: Combinations of Performance Measures used as Inputs along with Output PCDCP - Continues

Efficient Frontiers	MVolScE & MAE	MVolScE & MAVolScE	MVolScE& MSE; MVolScE& MSVolScE	MVolScE & MMEU	MVolScE & MMEO
E^1	{3,5,8}	{3,5}	{3,5,14}	{3,14}	{1,2,3,5,6,8,1}
E^2	{6,11}	{6,8,10,12}	{10,13}	{5}	{7,12}
E^3	{9,12}	{9}	{12}	{10}	{9,10}
E^4	{4,7,13}	{11,14}	{9,11}	{12,13}	{13,14}
E^5	{1,2}	{4,7,13}	{4,6}	{9}	{4}
E^6		{1,2}	{2,8}	{4,11}	
E^7			{7}	{6}	
E^8			{1}	{2,8}	
E^9				{7}	
E^{10}				{1}	

*[1]RW; [2] HM; [3]SMA20; [4]SMA60; [5]SES; [6]ARMA (1, 1); [7]AR (1); [8]AR (5); [9]GARCH (1, 1); [10]GARCH-M(1, 1); [11]EGARCH (1, 1); [12]TGARCH (1, 1); [13]PARCH (1, 1); [14]CGARCH(1,1)

Last, but not least, we have considered several measures of the goodness-of-fit criterion and the biasedness criterion to find out about the robustness of multidimensional rankings with respect to different measures. Our empirical results reveal that whether one measures biasedness by ME (Panel A, Table 3) or MVolScE (Panel B, Table 3), and measures of goodness-of-fit by MAE, MAVolScE, MSE or MSVolScE, the ranks of the best models (e.g., SMA20, SES) and the worst models (e.g.,HM, RW) remain the same; i.e., they are robust to changes in measures. Finally, whether one measures biasedness by ME or MVolScE, and measures goodness-of-fit by MMUO or MMEO, the ranks of the best and worst models differ significantly as compared to other goodness-of-fit measures combinations (e.g., RW, HM, CGARCH(1,1), which suggest that the performance of models such as RW, HM, CGARCH(1,1) is very sensitive to whether one penalizes negative errors more than positive ones or vice versa. Finally, notice that given the data set and the measures under consideration, our numerical results suggest that, with the exception of CGARCH, the family of GARCH models scored less as compared to smoothing models such as SMA20 and SES, which suggests that the data generation process has a relative long memory.

Table 3: Slacks-based Context-dependent DEA model scores-based multidimensional rankings of volatility forecasting models

Inputs	Output	Models Ranked from Best to Worst
ME; MAE	PCDCP	3→ 5→8→6→9&10→11→12→14→13→4→7→2→1
ME; MAVolScE	PCDCP	3→ 5→11→9→10→6→8→14→12→4→13→7→2→1
ME; MSE	PCDCP	3→ 5→14→9→11→10→13→12→4→6→8→2→7→1
ME; MSVolScE	PCDCP	3→ 5→14→9→11→10→13→12→4→6→8→2→7→1
ME; MMEU	PCDCP	3→ 14→5→9→10→11→12→13→4→6→8→2→7→1
ME; MMEO	PCDCP	3→5→2→11→1→6→8→9,10&12→14→13→4
Panel B: Combinations of Performance Measures used as Inputs along with Output PCDCP – Continues		
Inputs	**Output**	**Models Ranked from Best to Worst**
ME; MAE	PCDCP	3→ 5→8→6→10→9→12→14→11→13→4→7→2→1
ME; MAVolScE	PCDCP	3→ 5→10&12→6→8→9→14→11→4→13→7→2→1
ME; MSE	PCDCP	3→ 5→14→10→13→12→9→11→4→6→8→2→7→1
ME; MSVolScE	PCDCP	3→ 5→14→10→13→12→9→11→4→6→8→2→7→1
ME; MMEU	PCDCP	3→ 14→5→10→12→13→9→11→4→6→8→2→7→1
ME; MMEO	PCDCP	3→ 5→2→1→11→6→8→12→7→9&10→14→13→4

*[1]RW; [2] HM; [3]SMA20; [4]SMA60; [5]SES; [6]ARMA(1, 1); [7]AR(1); [8]AR(5); [9]GARCH(1,1); [10]GARCH-M(1,1); [11]EGARCH(1,1); [12]TGARCH(1,1); [13]PARCH(1, 1); [14]CGARCH(1,1)

5. CONCLUSION

Nowadays, forecasts play a crucial role in driving our decisions and shaping our future plans in many application areas such as economics, finance and investment, marketing, and design and operational management of supply chains, among others. Obviously, forecasting problems differ with respect to many dimensions; however, regardless of how one defines the forecasting problem, he or she needs to assess the relative performance of competing forecasting models and finds out which ones have the potential of doing a good "prediction job". Although most studies tend to use several performance criteria, and for each criterion, one or several metrics to measure each criterion, the assessment exercise of the relative performance of competing forecasting models is generally restricted to their ranking by measure, which usually leads to different unidimensional rankings.

In this study, we proposed a slack-based context-dependent DEA-based methodology, and used forecasting crude oil prices' volatility as an application area to illustrate the use of the proposed framework. The main conclusions of this research may be summarized as follows. First, the proposed multidimensional framework provides a valuable tool to apprehend the true nature of the relative performance of competing forecasting models. Second, models that are on the efficient frontier and have zero slacks regardless of the performance measures used (e.g., SMA20) maintain their ranks. Third, the multi-criteria rankings of the best and the worst models seem to be relatively robust to changes in most performance measures. Furthermore, when under-estimated forecasts are penalized, most GARCH types of models tend to perform well – suggesting that they often produce forecasts that are over-estimated. On the other hand, when over-estimated

forecasts are penalized, averaging models such as RW, HM, SES tend to perform very well – suggesting that these models often produce forecasts that are under-estimated. Finally, our empirical results seem to suggest that, with the exception of CGARCH, the family of GARCH models have an average performance as compared to smoothing models such as SMA20 and SES, which suggests that the data generation process has a relatively long memory.

6. REFERENCES

[1] Sadorsky, P. Stochastic Volatility Forecasting and Risk Management, Applied Financial Economics, 15, 121-135 (2005).

[2] Sadorsky, P. Modelling and forecasting petroleum futures volatility, Energy Economics, 28, 467-488 (2006).

[3] Agnolucci, P. Volatility in Crude Oil Futures: A Comparison of the Predictive Ability of GARCH and Implied Volatility Models, Energy Economics, 31, 316-321 (2009).

[4] Morita, H., Hirokawa, K., Zhu, J. A slack-based measure of efficiency in context-dependent data envelopment analysis, Omega, 33, 357-362 (2005).

[5] Seiford, L.M., Zhu, J. Context-dependent data envelopment analysis – measuring attractiveness and progress, Omega, 31, 397-408 (2003).

[6] Xu, B. Ouenniche, J. A multidimensional framework for performance evaluation of forecasting models: context-dependent DEA, Applied Financial Economics, 21, 1873-1890 (2011).

[7] Xu, B. Ouenniche, J. A Data Envelopment Analysis-based Framework for The Relative Performance Evaluation of Competing Crude Oil Prices' Volatility Forecasting Models, Energy Economics, 34, 576-583 (2012).

[8] B. Xu, J. Ouenniche, Performance evaluation of competing forecasting models: A multidimensional framework based on MCDA, Expert Systems with Applications, 39, 8312-8324 (2012).

[9] Ouenniche J., Xu B. and Tone K. Relative performance evaluation of competing crude oil prices' volatility forecasting models: a slacks-based super efficiency DEA model, American Journal of Operations Research, 4(4), 235-245 (2014).

[10] Ouenniche J., Xu B. and Tone K. Forecasting models evaluation using a slacks-based context-dependent DEA framework, Journal of Applied Business Research, 30(5), 1477-1484 (2014).

[11] Banker, R. D., Cooper, W.W., Seiford, L.M., Thrall, R.M., Zhu, J. Returns to scale in different DEA models, European Journal of Operational Research, 154, 345-362 (2004).

[12] Tone, K. A slacks-based measure of efficiency in data envelopment analysis, European Journal of Operational Research, 130, 498-509 (2001).