Multidimensional Approaches to Performance Evaluation of Competing Forecasting Models

Bing Xu

PhD in Management

University of Edinburgh

2009

Abstract

The purpose of my research is to contribute to the field of forecasting from a methodological perspective as well as to the field of crude oil as an application area to test the performance of my methodological contributions and assess their merits. In sum, two main methodological contributions are presented.

The first contribution consists of proposing a mathematical programming based approach, commonly referred to as Data Envelopment Analysis (DEA), as a multidimensional framework for relative performance evaluation of competing forecasting models or methods. As opposed to other performance measurement and evaluation frameworks, DEA allows one to identify the weaknesses of each model, as compared to the best one(s), and suggests ways to improve their overall performance. DEA is a generic framework and as such its implementation for a specific relative performance evaluation exercise requires a number of decisions to be made such as the choice of the units to be assessed, the choice of the relevant inputs and outputs to be used, and the choice of the appropriate models. In order to present and discuss how one might adapt this framework to measure and evaluate the relative performance of competing forecasting models, we first survey and classify the literature on performance criteria and their measures – including statistical tests – commonly used in evaluating and selecting forecasting models or methods. In sum, our classification will serve as a basis for the operationalisation of DEA. Finally, we test DEA performance in evaluating and selecting models to forecast crude oil prices. The second contribution consists of proposing a Multi-Criteria Decision Analysis (MCDA) based approach as a multidimensional framework for relative performance evaluation of the competing forecasting models or methods. In order to present and discuss how one might adapt such framework, we first revisit MCDA methodology, propose a revised methodological framework that consists of a sequential decision making process with feedback adjustment mechanisms, and provide guidelines as to how to operationalise it. Finally, we adapt such a methodological framework to address the problem of performance

evaluation of competing forecasting models. For illustration purposes, we have chosen the forecasting of crude oil prices as an application area.

Key words: Forecasting, Performance Measurement, Performance Evaluation, Data Envelopment Analysis (DEA), Multi-Criteria Decision Analysis (MCDA), Crude Oil

Acknowledgement

My PhD has been accompanied and supported by many people. First, I would like to express my special thanks to my supervisor, Dr. Jamal Ouenniche, for his continuous support and encouragement during my PhD studies. He first suggested the research on the performance evaluation of competing forecasting models, and guided me all the way through the research and writing up of this thesis. Without him, this thesis would never become a reality! He has been a great role model for me to follow – integrity and enthusiasm for producing high-quality work. It has been a great pleasure to work with him and I am looking forward for further collaborations. Second, I would like to thank Mr. Alan Brown and Professor Thomas Archibald in Business School at the University of Edinburgh, for their help and valuable advices during my PhD period. Third, I am very grateful for all the supports from my PhD fellows– Yue (Lucy) Liu, Yingfa Liu, Micheal Chang, and Kuangyi Liu, this journey would be so lonely without those good times that we spent together, shared laughs and tears; and a genuinely thanks to many other friends in the U.K. and in China, for taking care of me and helping me when I am mostly needed! Last but not the least, I would like to pay my sincerely thanks to my family – Bo Xu, Yue Hu, Xiaoxiao Wang and Chuan Xu, without their love, understanding and most importantly believing in me, I cannot image how I get where I am today.

Declaration

I declare that the content of this thesis is my own work.

Signature:

Date:

To My Family

Table of Content

Chapter 1:

Introduction

Introduction

The literature on forecasting spreads across a wide range of application areas. Most studies concerned with forecasting the level, the volatility, or both of time series tend to use one or several performance criteria and, for each criterion; e.g., accuracy, one or several metrics to assist in assessing the performance of competing models. Although several performance criteria and measures are most often used in studies, the assessment exercise of competing models is generally restricted to their ranking by measure; thus, the current methodology is unidimensional in nature. Consequently, one may obtain different rankings of models for different measures leading to inconsistent and often confusing results both within and across studies. Therefore, the purpose of my research is to contribute to the field of forecasting from a methodological perspective by proposing multidimensional frameworks to performance evaluation of competing forecasting models. Crude oil is chosen as an application area to illustrate the use of the proposed performance evaluation frameworks. In sum, two main methodological contributions are proposed.

The first contribution consists of proposing a mathematical programming based approach, commonly referred to as Data Envelopment Analysis (DEA), as a multidimensional framework for relative performance evaluation of competing forecasting models or methods. As opposed to other performance measurement and evaluation frameworks, DEA allows one to identify the weaknesses of each model, as compared to the best one(s), and suggests ways to improve their overall performance. DEA is a generic framework and as such its implementation for a specific relative performance evaluation exercise requires a number of decisions to be made such as the choice of the units to be assessed, the choice of the relevant inputs and outputs to be used, and the choice of the appropriate models. In order to present and discuss how one might adapt this framework to measure and evaluate the relative performance of competing forecasting models, we first survey and classify the literature on performance criteria and their measures – including statistical tests – commonly used in evaluating and selecting forecasting models or methods to assist in selecting the appropriate metrics

to measure the criteria under consideration. This classification will serve as a basis for the operationalisation of DEA. In sum, context-dependent DEA is proposed and its application is illustrated in evaluating and selecting models to forecast crude oil prices. The second contribution consists of proposing a Multi-Criteria Decision Analysis (MCDA) based approach as a multidimensional framework for relative performance evaluation of the competing forecasting models or methods. In order to present and discuss how one might adapt such framework, we first revisit MCDA methodology, propose a revised methodological framework that consists of a sequential decision making process with feedback adjustment mechanisms. Second, we provide guidelines as to how to operationalise it. Third, we survey and classify the literature on performance criteria and their measures – including statistical tests – commonly used in evaluating and selecting forecasting models or methods to assist in selecting the appropriate metrics to measure the criteria under consideration. Finally, we discuss how one might adapt such MCDA framework to address the problem of relative performance evaluation of competing forecasting models of crude oil prices. Three outranking methods have been used in our empirical experiments; namely, ELECTRE III, PROMETHEE I and PROMETHEE II.

Our main conclusions may be summarized as follows. First, the proposed two multidimensional frameworks provide valuable tools to apprehend the true nature of the relative performance of competing forecasting models. Second, as far as the evaluation of the relative performance of the valid forecasting models considered in this study is concerned, models such as the linear regression model REG2, REG3, REG5, Holt-Winter Exponential Smoothing with Multiplicative Seasonality (HWESMS), Random Walk (RW) Adjusted to Trend, tend to have ranks that are less sensitive to the performance measures, DEA methods, outranking methods and importance weights, which suggested that the rankings of these models are robust. Third, REG5 is superior to the remaining models. Finally, we recommend that the forecasts produced by models with similar performance should be combined and compared to all forecasts produced by individual models before deciding on the forecasting model to implement.

The remainder of this thesis is organized as follows. In Chapter 2, we survey and classify the literature on modeling crude oil data and forecasting and critically analyze the literature. In Chapter 3, we propose Data Envelopment Analysis (DEA) as a multidimensional framework for relative performance evaluation of competing forecasting models or methods and illustrate the use of the DEA framework to assess competing forecasting models of crude oil prices. In Chapter 4, we propose a Multi-Criteria Decision Analysis (MCDA) based approach as a multidimensional framework for relative performance evaluation of the competing forecasting models or methods and illustrate the use of such framework to assess competing forecasting models of crude oil prices. Finally, Chapter 5 concludes the thesis and outlines future research direction.

Chapter 2:

Survey, Classification and Analysis of the literature on Crude Oil

Survey, Classification and Analysis of the literature on Crude Oil

Abstract

The purpose of this chapter is threefold. First, we overview the oil prices movements over the past 30 years and analyze the potential factors that contribute to some of those spikes or general trend such as booming demand and limited supply. Second, we survey and classify the literatures on data modeling of crude oil prices. Third, we survey and classify the literature emphasis on forecasting both levels and volatilities of crude oil prices. Finally, we critically analyzed these literatures.

Key words: Crude Oil; Literature Review; Data Modeling; Forecasting.

2.1 Introduction

Oil is one of the most important sources of energy and any changes in price are major concerns of governments and governmental organizations, as they lead to a substantial impact on the world economy through various channels; e.g., high oil prices may lead to a lower production outputs, wages, and aggregate demand, higher unemployment and interest rates, and induce inflationary tendencies; whereas low oil prices may result in political instability in oil exporting countries (Mork, 1994; Hamilton, 1996; Barsky and Kilian, 2004; Jones et al., 2004). Therefore, crude oil has attracted the attention of many individuals including investors, analysts, and academic researches. The literature relevant to this paper could be divided into two main categories; namely, data modeling and forecasting. Although our main interest is in the forecasting area, we would like to draw the reader's attention to the fact that research concerned with data modeling provides valuable input to model building in forecasting. Research in data modeling is generally concerned with one of the following three questions; namely, are crude oil markets consistent with the market efficiency hypothesis? Are there any relationships between oil prices and economic variables? Are there any relationships between oil prices and prices of petroleum products? On the other hand, research on forecasting crude oil addresses several crude oil related variables such as prices, returns, supply, and demand. As far as prices and returns are concerned, quantitative forecasting models could be divided into three main categories; namely, non-artificial intelligence models, artificial intelligence models, and hybrid models.

The remainder of this chapter is organized as follows. In section 2.2, we provide an overview of the crude oil market and analyze the major reasons that drive the crude oil prices movements over the past 30 years. In section 2.3, we survey the literature on data modeling of crude oil and classified into three categories depending on the nature of studies and critically analyzed such literature. In section 2.4, we survey and classify the literature on forecasting the levels and/or volatilities of crude oil prices and critically evaluate the literatures. Finally, section 2.5 concludes the chapter.

2.2 Crude Oil Market Overviews

The price of crude oil is highly dependent on its grade as measured by specific gravity and its sulphur content, and its location. There are two main benchmark markets which acting as the references to the price of crude oil; namely, the WTI Crude Oil which is traded on New York Mercantile Exchange (NYMEX); Brent Crude Oil which is traded on the Intercontinental Exchange (ICE). There are several other important markets, which are Dubai, Tapis and the OPEC basket. The Energy Information Administration (EIA) uses the Imported Refiner Acquisition Cost as their world oil price, which is the weighted average cost of all oil imported into the US. Despite of different locations or markets, crude oil prices are now tent to move together and constitute a unified world oil market (Bentzen, 2007).

At first, oil market was controlled with great stability by a group of Western multi-national oil companies, known as the "Seven Sisters". These were British Petroleum, Royal Dutch, Exxon, Mobil, Texaco, Chevron and Gulf Oil. The "Seven Sisters" shared the profits on their concessions with the host countries- host countries received 50% of the value of extracted oil less the cost of production. A joint ownership structure of the holding companies enabled the "Seven Sisters" to plan most of world's production of crude oil and to set the global price – considered to be the reference point for the oil market and known as the "posted price". Then, OPEC (Organization of Petroleum Exporting Countries), a permanent intergovernmental organization was created at the Baghdad Conference on September 10 – 14, 1960 by Iran, Iraq, Kuwait, Saudi Arabia and Venezuela. The five founding members were later joined by eight other members; namely, Qatar, Indonesia, Socialist People Libyan Arab Jamahiriya, United Arab Emirates, Algeria, Nigeria, Ecuador and Gabon (Brief History, www.opec.org). The main objectives for the OPEC are to coordinate and unify petroleum policies among member countries in order to secure fair and stable prices for petroleum products, an efficient, economic and regular supply of petroleum to consuming nations, and a fair return on capital to those investing in the industry (Functions, www.opec.org). The oil prices have been fluctuating up and down since 1970s; for example, the recent lowest

point of crude oil prices was $11 per barrel in January 1999 that might be due to increased oil production from Iraq together with the Asian financial crisis; but from the end of 2001, oil prices starts this steadily upward trend from around $35 per barrel nominal oil price and kept reaching new high records (e.g., $80 per barrel on Sep 2007; and over $100 per barrel at the beginning of 2008 – See Figure 1, data source - EIA).

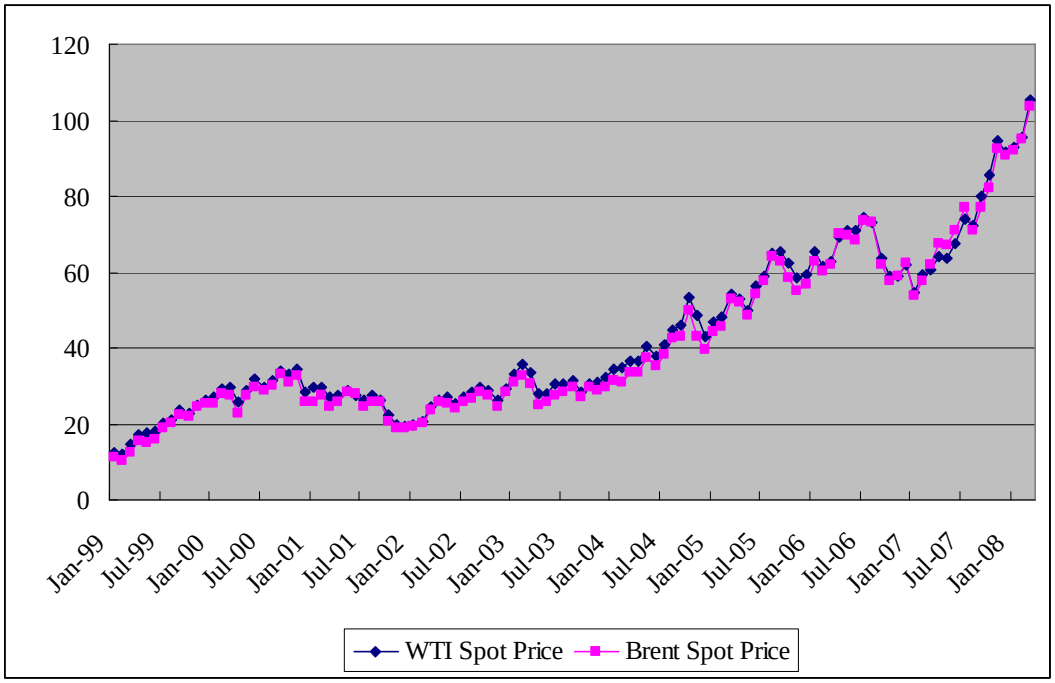


Figure 1: WTI Spot Price and Brent Spot Price

Next, we would like to analyze the potential reasons that attribute to the recent oil prices fluctuations. First, global macroeconomic conditions is believed to have a significant impact on the oil demand and subsequently on its prices (e.g., Barsky and Kilian, 2002; Hamilton, 2008); for example, the recent economic expansion in some emerging economies (e.g., China, India, Brazil and Mexico) have created a huge boost on the oil demand. In particularly, China became the world's second-largest consumer of oil behind the United States, and the third-largest net importer of oil after the U.S. and Japan in 2006; and India was the sixth largest consumer of oil in the world during 2006; and those countries experienced continuous growing demands for 2007 (see Table 1, data source - EIA).

Top World Oil Consumption 2006			Top World Oil Consumption 2007		
Rank	Country	Consumption	Rank	Country	Consumption
1	United States	20,687	1	United States	20,680
2	China	7,273	2	China	7,565
3	Japan	5,159	3	Japan	5,007
4	Russia	2,861	4	Russia	2,820
5	Germany	2,665	5	India	2,800
6	India	2,587	6	Germany	2,456
7	Canada	2,264	7	Brazil	2,400
8	Brazil	2,217	8	Canada	2,364
9	Korea, South	2,174	9	Korea, South	2,214
10	Saudi Arabia	2,139	10	Saudi Arabia	2,210
11	Mexico	1,997	11	Mexico	2,119
12	France	1,961	12	France	1,950
13	United Kingdom	1,830	13	United Kingdom	1,740
14	Italy	1,732	14	Iran	1,708
15	Iran	1,686	15	Italy	1,702

Table 1. Top World Oil Consumption 2006 & 2007 (Thousands Barrel Per Day)

Second, crude oil supplies are mainly divided into OPEC and non-OPEC sources (see Table 3 and Figure 2, data source- EIA). According to the EIA, total OPEC (i.e., Algeria, Angola, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates, Venezuela and the Neutral Zone) produced 42.74% of world crude oil production in 2007, of which Saudi Arabia alone accounted for 12%. On the other hand, only few nations have the ability to supply the crude oil equivalent to the OPEC members, such as, Russia, Norway and Mexico (See Table 3). It is believed that the production costs for non-OPEC producers (e.g., S, Russia, Mexico, China, Canada, Norway, UK, and Kazakhstan) tend to be higher than OPEC producers, which makes them more vulnerable to price falls. Furthermore, OPEC has experienced a decreasing trend on the available capacity more recently (see Figure 3, data source - EIA).

Top World Oil Net Importers 2006			Top World Oil Net Importers 2007		
Rank	**Country**	**Imports**	**Rank**	**Country**	**Imports**
1	United States	12,357	1	United States	12,224
2	Japan	5,031	2	Japan	4,874
3	China	3,428	3	China	3,653
4	Germany	2,514	4	Germany	2,310
5	Korea, South	2,156	5	Korea, South	2,184
6	France	1,890	6	India	1,919
7	India	1,733	7	France	1,879
8	Italy	1,568	8	Spain	1,583
9	Spain	1,562	9	Italy	1,533
10	Taiwan	940	10	Taiwan	967
11	Netherlands	935	11	Singapore	907
12	Singapore	825	12	Netherlands	901
13	Turkey	625	13	Turkey	645
14	Thailand	594	14	Belgium	617
15	Belgium	583	15	Thailand	607

Table 2. Top World Oil Net Importers 2006 & 2007 (Thousands Barrel Per Day)

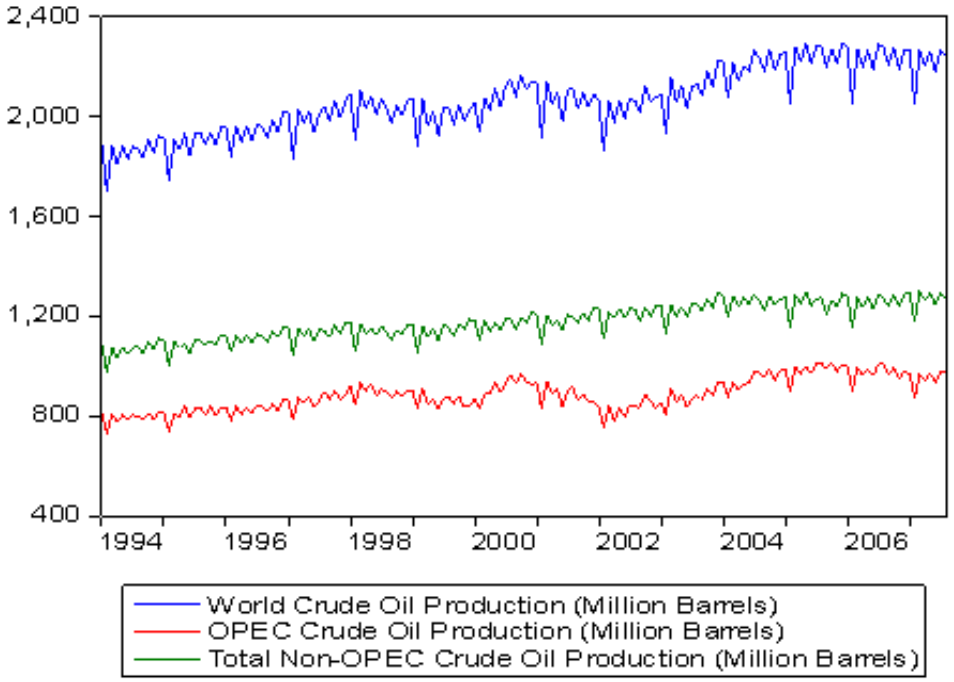


Figure 2: Oil Productions (Million Barrels)

Top World Oil Producers, 2006			Top World Oil Producers, 2007		
Rank	Country	Production	Rank	Country	Production
1	Saudi Arabia	8,525	1	Saudi Arabia	10,248
2	Russia	6,816	2	Russia	9,874
3	United Arab Emirates	2,564	3	United States	8,457
4	Norway	2,551	4	Iran	4,034
5	Iran	2,462	5	China	3,912
6	Kuwait	2,342	6	Mexico	3,500
7	Venezuela	2,183	7	Canada	3,422
8	Nigeria	2,131	8	United Arab Emirates	2,948
9	Algeria	1,842	9	Venezuela	2,670
10	Mexico	1,710	10	Kuwait	2,616
11	Libya	1,530	11	Norway	2,565
12	Iraq	1,438	12	Nigeria	2,353
13	Angola	1,379	13	Brazil	2,253
14	Kazakhstan	1,145	14	Algeria	2,174
15	Qatar	1,032	15	Iraq	2,097

Table 3. Top World Oil Producers 2006 & 2007 (Thousands Barrel Per Day)

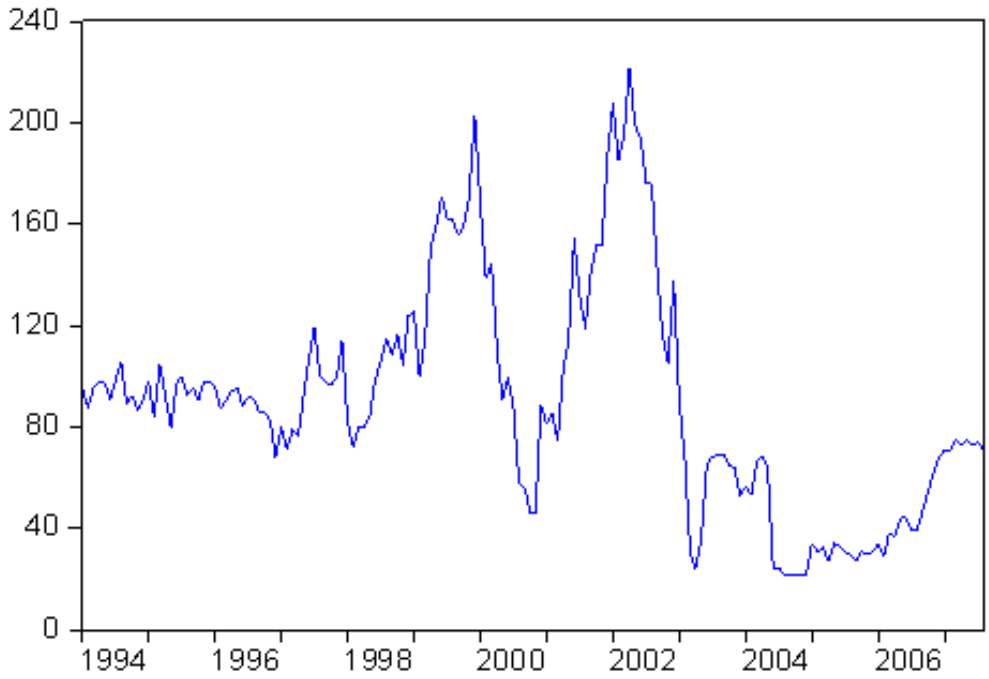


Figure 3: OPEC Excess Capacities (Million Barrels)

Furthermore, there are many indirect factors might affect the prices through different channels (see Figure 4, source EIA). First, political tensions and instabilities within both OPEC countries (i.e., Iraq, Saudi Arabia, Nigeria, and Venezuela) and non-OPEC countries (i.e., Bolivia, Turkey and Russia) have an effect on the supplies of the crude oil and raised concerns about the availability of future oil supplies (BBC News, January 2008). For instance, actions taken by OPEC members to restrict the production of crude oil (e.g., tighten the oil production in April 1999); recent attacks (19th June 2008) on Nigeria pipelines by anti-government militants led to a loss in production for approximately 120,000 barrels per day which equivalently to 6.6% of total production (BBC News, June 2008). However, recent empirical studies were not able to prove that those exogenous supply shocks have a significant impact on the U.S. economy (e.g., Kilian, 2008a). Second, excess speculative trading carried out by hedge funds and brokers might have a further upward pressure on the oil prices, as they bet on the possibility of oil price increases and add extra momentum to the prices (DeLong et al., 1990; Dufour and Engle, 2000; Hamilton, 2009). Third, the lack of spare refining capacity (e.g., the number of refineries in the US has not increased since 1981, and a significant fraction of refinery was closed due to unscheduled maintenance; New York Times, July 2007) and low US gasoline stocks might have an impact on the oil prices (BBC News, September 2004). Nevertheless, Kaufmann et al. (2004) and Dees et al. (2008) failed to find any empirical evidence on that refinery utilization rates have any significant effects on the prices of crude oil. More recently, there were raising concerns on the environmental impacts on the oil prices behaviors (e.g., Henriques and Sadorsky, 2008; Duncan, 2006, 2007). In fact, global warming tends to increase the probability for adverse weather conditions and natural disasters, and subsequently has significant impacts on the oil consumptions and/or production (Glick, 2004); for example, cold winter and/or hot summers lead to increased consumptions (e.g., frozen winter in China 2008); the Hurricane Katrina led to approximately 20% of the Gulf of Mexico's oil and gas production reduction, as refineries were offline in August 2005 (BBC News, August 2005).

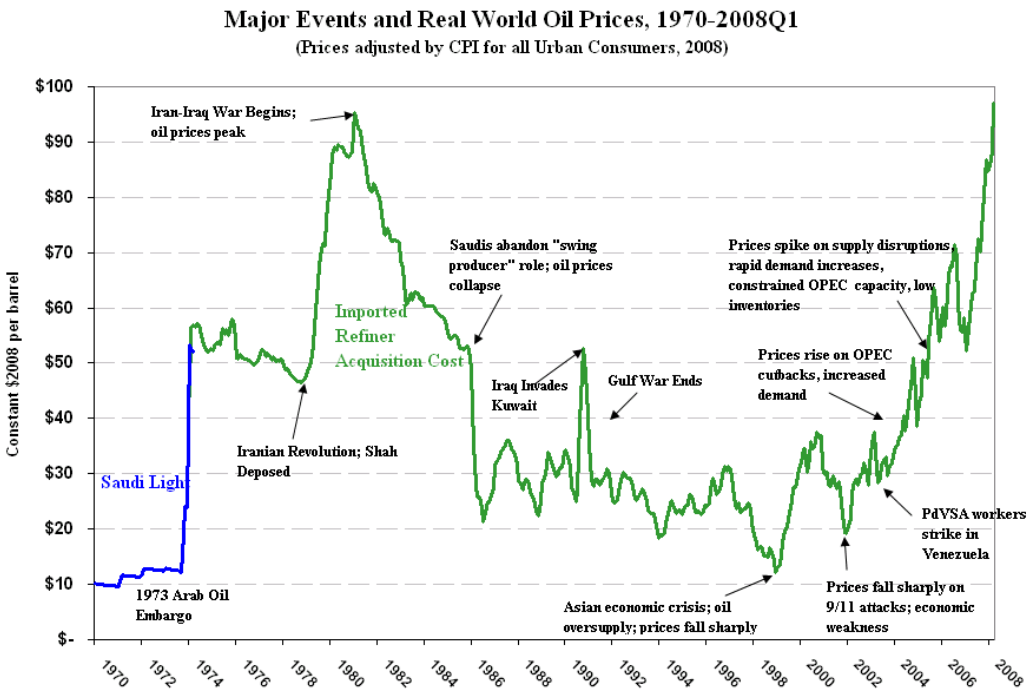


Figure 4: Major Events and Real World Oil Prices from 1970 – 2008 (Source: EIA)

To conclude this part, oil is one of the most important strategic commodities and any changes in prices and their magnitude could have a substantial impact on the economy. For example, high oil prices often lead to an increase in inflation and subsequently hurt economies of oil-importing countries on one hand; whereas the low prices may cause political instability in oil-exporting countries as their economy rely heavily on oil exports on the other hand. The oil market was first controlled by the "Seven Sisters" until the OPEC was formed to secure fair and stable prices of petroleum products in 1960, and the balance of power gradually from the United States to OPEC. Since the early 1970's and up to now, oil prices have experienced ups and downs and this might due to several reasons. First, the booming economies in some emerging markets (e.g., Brazil, China and India) have driven the oil demand significantly and subsequently might cause an upward pressure on crude oil prices. Second, there is a limited supply on the crude oil and high dependences on the OPEC producers. Third, the political instabilities in both OPEC regions (e.g., Iraq and Venezuela) and non-OPEC regions (e.g., Bolivia, Turkey and Russia) have an impact on the production and tend to raise the fears about further disruptions in exports. Fourth, the excess speculative activities tend

to bet on the possibility of oil prices increasing and add extra momentum to the price and further upward pressure on oil. In addition, the insufficient midstream and downstream infrastructures might have an impact on refiners' ability to deliver oil and gas production in a timely manner. Finally, it is believed that the increasing concentration of CO_2 level led to global warming and consequently increasing the likelihood on adverse weather conditions and eventually affects the oil supplies and demand as well as the oil prices.

2.3 Literature Review on Modeling Crude Oil Data

A rich body of studies analyzing oil prices dynamics and relationships with other variables has been published. Depending on the nature of the research questions, we classify the literature on crude oil data modeling into three categories; namely, are crude oil markets consistent with the market efficiency hypothesis? If not, where the inefficiencies lie? Are there any relationships between oil prices and economic variables? (e.g., whether there are any adverse relationships between oil and economy variables that are consistent with economic theories); Are there any relationships between oil prices and prices of petroleum products? (e.g., potential transmissions from crude oil price movements to petroleum products). In order to test these hypotheses, different methods have been used (e.g., factor analysis, regression analysis, and cointegration analysis), and various explanatory variables have been included in these analyses, and could be divided into four main categories; namely, price variables and their lagged values (e.g., WTI, Brent, OPEC, and petroleum products), economic and political variables (e.g., GDP, inflation and interest rates, exchange rates, tax variables), financial variables (e.g., market indices, industrial indices), and specific event-related variables (e.g., dummies for Gulf War, September 11^{th} 2001).

The remaining of this section is organized as follows. In section 2.3.1, we review the literature related to the oil market efficiency tests. In section 2.3.2, we highlight the changing oil and macroeconomy relationships. In section 2.3.3, we present the studies related to oil products. In section 2.3.4, we summarize our findings and conclude this section.

2.3.1 Oil Markets Efficiency Tests

The literature on testing oil markets efficiency could be divided into three sub-categories depends on the nature of the research questions. First, if the crude oil market is semi-strong efficient, then futures prices should neither lead nor lag the underlying spot price. In general, common methodologies have applied to test for the lead-lag relationships were cointegration analysis, Granger causality analysis, and linear regression analysis; and the empirical evidence were mixed. For examples, Serletis and Banack (1990) tested market efficiency hypothesis for futures markets on crude oil, gasoline and heating oil with cointegration analysis, they found evidence to support the hypothesis. Crowder and Hamid (1993) also used cointegration analysis and obtained similar result. However, Quan (1992) tested for a cointegration relationship between crude oil spot prices and futures contracts, and showed that futures prices do not play an important role for futures contracts longer than three months; but Schwartz and Szakmary (1994) found the contradictory results, by examining the cointegration, arbitrage and price discovery in crude oil, heating oil and gasoline futures prices. Gulen (1999) applied cointegration tests for WTI spot prices and NYMEX one-, three- and six- month ahead futures contracts, and identified a structural break after the crash of early 1986 and light sweet crude oil future prices plays a central role in price discovery. Moreover, Silvapulle and Moosa (1999) tested for both linear and nonlinear causalities on daily spot and futures prices of WTI crude oil and reached different results (e.g., linear Granger causality test suggested that futures market leads the spot market, but non-linear causality test suggested a bi-directional effect). Bekiros and Diks (2008) also examined for both linear and nonparametric nonlinear causality (by Diks and Panchenko) relationships between daily spot and future prices, and found evidences on nonlinear effects. Furthermore, Green and Mork (1991) used a linear regression model to test for the spot market efficiency and failed to support the null hypothesis for period of 1978 – 1985, but suggested an increased efficiency in 1980s. Moosa and AI-Loughani (1994) tested for the market efficiency and unbiasedness jointly with linaer regression analysis, by using monthly data on WTI crude oil spot and three- and six- month ahead futures prices, they suggested futures prices were neither unbiased nor efficient predictors of spot prices.

Nevertheless, Peroni and McNown (1998) found evidences for a speculative efficiency hypothesis for the WTI crude oil prices during the period of 1984 to 1996 and questioned Moosa and Al-Loughani (1994)'s credibility of their conclusion. Hammoudeh et al. (2003) used VECMs to test the causality and volatility spillovers among petroleum prices of WTI, gasoline, and heating oil spot and futures prices, and found long term relationships among spot and future prices.

Second, if crude oil market is semi-strong efficient, then oil prices in different markets should move closely together with more or less constant price differential (Weiner, 1991), and the regionalization hypothesis should then be rejected. The empirical results on testing the regionalization hypothesis were inconsistent. For instance, Weiner (1991) failed to reject the regionalization hypothesis. Ewing and Harter (2000) tested the null hypothesis of a unified market, and they found out crude oil prices of Alaska North Slope and Brent share a long run common trend. Kleit (2001) concluded that oil markets became more unified. Lin and Tamvakis (2001) investigated information transmission between the NYMEX and IPE, and found NYMEX was the true leader in the crude oil market. Bentzen (2007) used vector error correction (VEC) model to test for the co-movements among WTI, Brent and OPEC prices and rejected the regionalization hypothesis of the global oil market.

Third, if crude oil market is weak-form efficient, then prices should follow a random walk process (e.g., evidence on chaotic, nonlinear structures and/or long memory from the past prices cannot be used to explore the future prices movements). The empirical results were mixed. Recent studies by Panas and Ninni (2000) tested for the chaotic behavior of the oil markets; they found that oil products (e.g., Naphtha, Mogas Prem, gasoil) were chaotic by using correlation dimensions, entropies and Lyapunoy exponents. However, Adrangi et al. (2001) only found non-linear dependencies for crude oil, heating oil and unleaded gasoline futures prices, and failed to find any evidence on chaotic behavior. Based on Lyapunov test and generalized BDS statistic along with Kaplan's test, Matilla-Garcia (2007) found evidences for both nonlinearity and chaotic behavior for crude oil, natural gas, and unleaded gasoline futures prices. Furthermore,

several studies have tested for the long – range dependence in crude oil prices with persistent structure. For example, Alvarez-Ramirez et al. (2002) found crude oil market has a persistent process with long-run memory effects by using Brent, WTI and Dubai crude oil prices for the period of 1981 to 2001 with multifractal analysis; and this study has been further extended with a Zipf-type analysis methodology to obtain additional information on oil price dynamics by Alvarze – Ramirez et al. (2003). Serletis and Andreadis (2004) found empirical evidence of long-range dependences in WTI crude oil prices and Henry Hub natural gas prices with persistent structure. Tabak and Cajueiro (2007) analyzed the Brent and WTI crude oil prices for the period of 1983 to 2004, based on the time-varying Hurst exponents, they revealed that crude oil market has became more efficient over time and also suggested oil prices volatility is highly persistence and volatility models such as GARCH or EGARCH should not be used to estimate and forecast its volatility. In addition, Shambora and Rossiter (2007) used an artificial neural network (ANN) model with moving average crossover inputs to test for efficiency of crude oil futures market, and rejected the efficiency hypothesis for the NYMEX crude oil futures market.

The overall results on testing the crude oil markets efficiency were mixed. The potential reasons for causing this inconsistence might because of the use of different data set and/or period. More importantly, various methodologies have also been reported in the literature, and Peroni and McNown (1998) criticized some earlier studies for which they have included nonstationary time series in their regression models and questioned the reliability of their conclusions.

2.3.2 Oil – Economy Relationships

Oil prices shocks tend to influence the economy through various transmission channels, for example, it tend to have a significant impact on the supply side of the economy, as rising oil prices will reduce the availability of input to production and lead to a reduced output (see Brown and Yücel, 1999, Abel and Bernanke, 2001); another channel is from the demand side, as higher oil prices tend to reduce oil-importing countries' purchasing power of firms and households, and subsequently lead a wealth transfer from oil –

importing countries to oil-exporting countries (e.g., Olson, 1988; Ferderer, 1996); according to the real balance effect, counter inflationary monetary policy will respond to oil prices increases and subsequently led to losses on real output (e.g., Bohi, 1991; Bernanke et al., 1997; Bernanke and Mihov, 1998); furthermore, oil prices shocks might raise uncertainty and operating costs for certain durable goods and contributed to a reduced or postponed demand for durables and investments (see Bernanke, 1983; Pindyck and Rotemberg, 1984; Hamilton, 1988; Lee and Ni, 2002). Besides the oil prices, volatility of oil prices are also have a significant impact on the economy (Ferderer, 1996).

In this section, we review empirical studies on oil and economy relationships and report their main results. We classify the literature into two main categories; namely, are there any relationships between oil prices and macroeconomic variables? (e.g., to test for the impact for oil prices shocks and various aggregate economic activities, GDP, inflation, interest rate, exchange rates, and financial markets); are there any relationships between oil prices and microeconomic activities (e.g., individual industries, firms, or workers)? While not exhausted, we highlight the key trends and critically evaluate the main results.

2.3.2.1. Oil Prices Shocks and Macroeconomy Relationships

There have been an increased numbers of empirical studies on testing the impacts of a oil price shock on the macroeconomy. The overall results were mixed. Hamilton (1983, 1985) used a vector autoregressive (VAR) methodology and suggested that oil prices had preceded seven out of eight post-war US recessions between 1948 and 1980; and his empirical results of an inverse relationship between oil prices and US macroeconomic aggregates has been further confirmed by some other studies (e.g., Gisser and Goodwin, 1987; Hickman et al., 1987; Darby, 1982; Bruno and Sachs, 1982). Furthermore, these inverse relationships have also been tested for other economies. For instance, Burbidge and Harrison (1984) found mixed but overall reinforcing evidence for UK, Canada, France, Germany, and Japan; Chang and Wong (2003) only found marginal impact on the Singapore economy; Rautava (2004) studied oil prices shocks and their relationship with Russian macroeconomic variables.

Moreover, a number of studies concerned with the interactions between oil prices and exchange rates and suggested a positive relationship in general. For example, Amano and Norden (1998), Olomola and Adejumo (2006) and Benassy – Quere et al. (2008) all found a positive relationship between real oil prices and US dollar based on cointegration and causality tests; Camarero and Tamarit (2002) applied panel cointegration techniques and also suggested a positive relationship between oil prices shocks and the real exchange rate of the Spanish peseta; Zalduendo (2006) used VECM methodology and noticed that Brent real oil prices became a significant cause on the real exchange rate of Venezuela. Huang and Guo (2007) constructed a structural VAR model and also confirmed the positive relationship between real oil price and real exchange rate for China. Narayan et al. (2008) found the positive relationship between oil price and the Fiji – US exchange rate based on GARCH and EGARCH models. Nonetheless, Chen and Chen (2007) investigated the long-run relationship between real oil prices and real exchange rates for G7 countries for the period ranging from January 1972 to October 2005, and they found negative relationships.

Another interesting area that attracted many researches was analyzing the relationships between oil prices shocks and financial markets (i.e., developed and/or emerging markets), and the empirical results were mixed overall. For example, Jones and Kaul (1996) used quarterly data from 1947 to 1991, and found out oil prices movements have a significant impacts on the US and Canadian stock prices but not for Japan and UK. Huang et al. (1996) investigated the daily oil futures returns and the US stock returns with VAR methodology, and failed to prove that oil futures returns have any significant impacts on the S&P 500. Papapetrou (2001) studied the interaction between oil prices, real stock prices, interest rates, and real economic activity in Greece with a multivariate VAR methodology, and found oil prices was an important factor for explaining stock prices movements. On the other hand, there has been a increased attention on the studying the emerging markets. For example, Hammoudeh and Eleisa (2004) examined the relationships between daily oil prices and stock prices for five members of the Gulf Cooperation Council (i.e., Bahrain, Kuwait, Oman, Saudi Arabia and the United Arab Emigrates); they suggested that only the Saudi Arabia stock market has a bi-directional

relationship between oil prices and stock prices. Malik and Hammoudeh (2007) used VECH and BEKK to examine the volatility and shock transmission mechanism among US equity, oil prices and equity markets of Saudi Arabia, Kuwait and Bahrain, for the period of February 1994 to December 2001, and suggested that only Saudi Arabia market has an bi-direction effects to spill-over volatility to the oil market. Maghyereh (2004) explored 22 emerging stocks markets with the VAR methodology for the period ranging from January 1998 to April 2004, and did not find any relationships between oil prices and those economies. Basher and Sadorsky (2006) applied international multi-factor models to study the impact of daily, weekly, and monthly oil prices on emerging stock market returns for 21 emerging stock markets over a period spanned from January 1993 to October 2005, and found mixed evidences on the relationships between oil prices shocks and stock markets.

While there has a general acceptance for an adverse relationship between oil prices and macroeconomy, Hooker (1996), Huntington (1998) among others, highlighted that oil prices failed to Granger cause most key macroeconomic variables (e.g., the real GDP, unemployment rate, aggregate employment, and industrial production) from the mid 1980s. A number of authors (e.g., Loungani, 1986, Hamilton, 1996) argued that the changing relationship between oil prices and outputs were due to the misspecification of the functional form rather than a weakened relationship. Ciner (2002) also believed that for those studies who failed to detect any relationship between oil prices and the stock market could be due to the use of linear causality test; and when he conducted both linear and nonlinear causality tests between oil futures returns and stock index returns with Huang et al. (1996)'s data, he have successfully found a nonlinear relationship between oil price shocks and stock index returns. Furthermore, Mork (1989) first suggested the asymmetries effects (e.g., increasing oil prices have more impacts on the aggregate economic activity than decreasing oil prices of the same magnitudes), and found that only the positive prices has a significant impact on the U.S economy; this asymmetry effect has been further confirmed on other industrialized countries by Mork et al. (1994). What is more, Lee et al. (1995) and Hamilton (1996) defined a number of proxy variables for oil price shocks (e.g., interanual changes of oil prices, oil price

increases, net oil price increases), the empirical results showed the asymmetric effect for which the positive shocks have a statistically significant impact on the US economic activities. Those proxy variables have been further used by several studies test for other economies; for example, Cunado and Perez de Gracia (2003, 2005) investigated fifteen European countries, and Six Asian countries, respectively; Jimenze – Rodriguez and Sanchez (2005) compared both linear and non-linear approaches to test the impacts of an oil price shock on main industrialized countries. On the other hand, Gately and Huntington (2002), Huntington (2004) decomposed the response to oil prices into three separate components, the maximum price change, the price cut and the price recovery. Moreover, Balke and Fomby (1997) first proposed a threshold cointegration methodology which considered non-linearity and cointegration simultaneously; this methodology has been further extended by Enders and Dibooglu (2001), Lo and Zivot (2001) and Schorderet (2004). Lardic and Mignon (2006) decomposed positive and negative oil prices according to Schorderet (2004) and investigated the existence of a long-term relationship between oil prices and GDP in 12 European countries; the same methodology has been used by Lardic and Mignon (2008) to test for the G7, US, and Europe and Euro area. In addition, Huang et al. (2005) used a multivariate threshold autoregressive (TAR) model to test for relationships between oil price shocks and macroeconomic aggregates (e.g., industrial production and stock returns, based on monthly data of the US, Canada and Japan for the period ranging from 1970 – 2002, they suggested the existence of the threshold value.

More recently, Kilian (2008a) proposed a new measure of exogenous oil supply shocks and found a sharp drop in real GDP growth and a spike in CPI inflation series after five weeks for an exogenous oil supply shocks, but he suggested that exogenous supply shocks did not have much impact on the U.S. economy since the 1970s; Kilian (2008b) extended his previous work in 2008a, and compared the effects of exogenous shocks to oil production on CPI inflation and real GDP on seven major industrialized countries. Killian (2009) proposed a new measure on global real economic activity and designed a new structural VAR methodology which allows one to identify how different shocks (e.g., oil supply shocks, all other oil supply shocks, aggregate demand shocks) will

affect oil price shocks, the empirical results suggested that no two oil price shocks were alike and these shocks have different impacts on the U.S. real growth and CPI inflation depending on the causes of oil price increases (e.g., booming world economy; a disruption of global crude oil production, or shifts in precautionary crude oil demand that reflect the concerns for future oil supply shortage); this structural VAR methodology have been further apply to other contexts, e.g. Kilian et al. (2009) studied the relationship between oil prices and external balance; Kilian and Park (forthcoming) analyzed the aggregate stock market fluctuations associated with oil price shocks.

In sum, a large number of studies have examined the impact of an oil price shock on the macroeconomy, and the results obtained were far from a consensus. This might be due to first, the changing relationship after mid- 1980s. Second, the definition of an oil price shock may have changed dramatically (Huntington, 2007), and different types of asymmetric and non-linear transformation have been proposed to evaluate the oil price and macroeconomy relationships. Third, Hamilton (2001, 2003) highlighted the problem for choosing an appropriate nonlinear specification and proposed a flexible parametric model to investigate the nature of the nonlinearity.

2.3.2.2 Oil Prices Shocks and Microeconomy Relationships

At a less aggregate level, a large body of studies focused on analyzing the interactions among oil prices and specific industrials such as gas, oil sectors, or commodities (e.g., metals) or firms. It is often believed that oil price shocks tend to raise the uncertainty and operating costs of certain durable goods and therefore might lead to a reduced or postponed demand for durable and investments (e.g., Pindyck and Rotemberg, 1984; Hamilton, 1988, 1999; Lee and Ni, 2002).

In general, most empirical studies tend to find a positive relationship between oil prices and energy related industries and/or companies' stock prices and/or returns. For example, Al-Mudhaf and Goodwin (1993) examined the impact of oil price shocks on 29 oil companies listed on the New York Stock Exchange and found a positive relationship between oil price shocks and firms with significant assets in domestic oil production.

Faff and Brailsford (1999) investigated the sensitivity of Australian industry equity returns to an oil price factor with monthly data ranging from 1983 to 1996, based on the conventional multifactor framework; they found negative relationships on equity returns for all industries except mining, oil and gas industries. Sadorsky (2001) confirmed the positive relationships between oil price shocks and stock returns of Canadian oil and gas companies. El-Sharif et al. (2005) revealed positive interactions between oil prices shocks and oil, gas sector in the UK. Boyer and Filion (2007) also suggested a positive relationship between energy stock returns and appreciation of oil and gas prices. Furthermore, Lee and Ni (2002) used VAR models to analyze the effects of oil price shocks on demand and supply in various industries, they discovered that oil price shocks tend to induce recessions. Based on linear regression methods, Nandha and Faff (2008) found the adverse relationships between oil prices shocks and 35 industry indices for the period spanned from April 1983 to September 2005. Baffes (2007) found mixed results on the effect of real crude oil prices movements on 35 real commodities price indices. Nevertheless, there has been relatively less empirical works to examine the relationships between oil prices and alternative energy companies, except Henriques and Sadorsky (2008), they have used a VAR methodology to examine the impact of oil prices shocks on the stock prices of alternative energy companies, between the stocks prices for the period ranging from Jan 2001 to May 2007, and they found that technology stock prices have a significant impact on alternative energy stock prices.

Besides the oil prices levels, the uncertainty also drives the volatility of oil prices and subsequently influences other industries or firms. In particular, several studies have tested whether oil prices were more volatile than other commodities; for example, Plourde and Watkins (1998) discovered that crude oil prices were significantly more volatile than over 93% of other commodities during the period 1985 to 1994; Regnier (2007) also found that energy prices (e.g., crude oil, refined petroleum, and natural gas) were more volatile than prices for about 95% of products sold by U.S. producers over the period ranging from January 1945 to August 2005.

Overall, depending on whether oil is an input or output for an industry or company, oil prices shocks tend to have different impacts on their returns. To be more specially, oil companies might be able to pass on higher costs to their customers, and therefore higher crude oil prices will not be necessary have significant effects on their profitability, and most empirical studies observed a positive relationship between oil price shocks and their returns (Nandha and Faff, 2007).

2.3.2.3 Conclusion

Oil price increases alone is not a sufficient condition for recessions (e.g., Hamilton, 2001; Jones et al., 2004), but it normally affects many aspects of the economy (e.g., reduce production output, wages, and aggregate demand; raise interest rates and unemployment etc). Thus, a large number of empirical studies attempted to investigate the interactions among oil prices and economic variables. The key findings from the literature could be summarized as follows. First, Hamilton (1983, 1985)'s early studies have gained general acceptance on the negative linear relationships between oil and major macroeconomic variables. However, this strong liner inverse relationship seems to be weakened after mid- 1980s (e.g., Hooker, 1996). On the other hand, a number of authors argued that the changing relationships between oil prices and outputs were due to the misspecification of the oil price. Therefore, different asymmetric and non-linear transformations of oil prices have been proposed in the literature, and most studies found evidences on that economic activities responded asymmetrically to oil price shocks. Furthermore, the empirical results on examining the financial markets returns were inconsistent – at the aggregate level, most studies confirmed with the negative impacts of oil prices shocks on economic growth; but at the less aggregate level, depending on whether oil is acting as an input or output for the specific industry or firm, their returns tend to react differently to the oil prices shocks.

Furthermore, most common methodologies that have been used in the literature on analyzing the oil and economy relationships could be divided into two categories; namely, single equation regression analysis and multiple equation regression analysis. In terms of the single equation regression analysis, both linear and nonlinear regression

models have been used to analyze the relationships among oil prices on economic activities (e.g., Hamilton, 2001, 2003; Huntington, 2004). One of the key assumptions for the single equation regression analysis is that all dependent variables are exogenous and no interrelationships among those variables. Another approach that has been commonly used in the literature is the multiple equation regression analysis, it is a convenient way to model systems of interrelated variables, and Granger Causality tests and innovation accounting (i.e., impulse responses and variance decomposition analysis) could be used to obtain further information on the interactions among the variables (Brooks, 2002). In this category, two approaches have most often been used in the literature. First, the unrestricted VARs (respectively, the structural VARs) are commonly used in the literature to estimate dynamic relationships among crude oil prices and other economic variables without (respectively, with) impose prior restrictions (e.g., structural relationships, or exogeneity of some of the variables; see Hamilton, 1983, 1985; Huang et al. 2005; Lardic and Mignon, 2006, 2008; Kilian, 2008a, 2008b, 2009). Second, vector error correction models (VECMs) is another popular approach to test for long term relationships among oil prices and economic variables and most often the Johansen Maximum likelihood was applied to test for cointegration in a multivariate context.

2.3.3 Oil Prices and Petroleum Products

The literature on examining oil prices and its refinery products (e.g., retail gasoline prices, heating oil, naphtha and kerosene) could be divided into upstream and downstream prices transmissions. To be more specific, upstream relationships reveal the impact of inputs variations (crude oil prices) on outputs variations (i.e., wholesale or retail prices of petroleum products); whereas downstream relationships consider the interaction among outputs (e.g., gasoline and heating oil), and/or impact of wholesales on retail prices of outputs (Karrenbrock, 1991; Duffy-Deno, 1996; Davis and Hamilton, 2004). In this chapter, we mainly focused on reviewing studies on analyzing upstream relationships on other refinery products in response to crude oil prices changes.

Most studies found empirical evidences on long term relationships between crude oil and its refined products. For instance, Serletis (1994) used Johansen's cointegration test

to check for common stochastic trends for crude oil, heating oil, and unleaded gasoline futures returns for the period of December 1984 and April 1993, and only reported one common trend. Asche et al. (2003) also used Johansen's test and found the existence of long term relationships between monthly Brent crude oil prices and gas oil, naphtha and kerosene for the period ranging from 1992 – 2000. Siliverstovs et al. (2005) used principal components analysis (PCA) and Johansen cointegration test to investigate the degree of integration of natural gas markets in Europe, North America and Japan for the period from 1993 to 2004 and found co-movements within those markets except for the European (respectively, Japanese) and the North American markets. Bachmeier and Griffin (2006) tested for cointegration relationships between crude oil and gas prices and suggested a weak integration in the U.S. market. Asche et al. (2006) also used Johansen cointegration test and found the UK energy market was an integrated market. Furthermore, Akarca and Andrianacos (1998) investigated the dynamic relationship between monthly crude oil prices and gasoline prices for the period of 1976 to 1996, and based on a regression analysis and Box-Tiao Statistic they identified the changing relationship since February 1986. Panagiotidis and Rutledge (2007) tested the relationship between UK wholesale gas prices and the Brent oil prices over the period of 1996 - 2003, and confirmed gas prices and oil prices were not decoupled. Based on VAR and bivariate GARCH models, Adrangi et al. (2001) only found a uni-directional relationship between the daily returns of Alaska North Slop crude oil and L.A. diesel fuel price.

On the other hand, there has been an increased attention on analyzing the asymmetric responses of petroleum products to crude oil prices movements. In particular, Bacon (1991) first referred to this asymmetric pattern as the "rockets and feathers" hypothesis, where petroleum products prices rise/drop at a rocket's/feather's speed by following an increase/decrease in upstream prices (crude oil prices), respectively. The empirical studies on analyzing the relationships between crude oil prices and gasoline prices (and/or other petroleum products) have generally shown an asymmetric behaviour with some exceptions (e.g., Bachmeier and Griffin, 2003, only found a symmetric speed of adjustment when daily data is used to capture the dynamics of price pass-through within

a week). In the U.S. markets, Borenstein et al. (1997) used an error correction model (ECM) to exam the asymmetric behaviour for weekly gasoline market for the period of 1986 - 1992; and Balke et al. (1998) extended Borenstein et al. (1997)'s study with two different model specifications with weekly data from 1987 – 1997. Chen et al. (2005) investigated asymmetric gasoline prices adjustment to crude oil prices with threshold cointegration by employing the weekly data from January 1991 to March 2003, and found asymmetric transmission for both spot and futures markets of crude oil and refinery gasoline. Radchenko (2005) examined the response of retail gasoline prices to the crude oil prices movements in the US, based on the Markov-Switching models, he observed significant differences between the long term and short term shocks to the gasoline prices. Kaufmann and Laskowski (2005) found evidence for asymmetric behaviour between gasoline and crude oil prices, heating oil and crude oil prices. Based on Threshold autoregression (TAR) model, momentum-threshold autoregressive (M-TAR) model and VECM, Al-Gudhea et al. (2007) found the asymmetric price adjustments in the U.S. gasoline prices to the crude oil prices for the period from December 1998 to January 2004. Blair and Rezek (2008) used an asymmetric ECM to analyze the impact of WTI crude oil prices on the gasoline prices in the Gulf Region for the period immediately after the Hurricane Katrina, and they observed the deviation from historic price pass through patterns for the immediate post-Hurricane Katrina period. Furthermore, a number of papers examined the asymmetric effects for other OECD countries. For example, Bacon (1991) used semi-monthly data for the period of 1982 and 1989, and found evidences on a rocket response of the UK gasoline retail prices to oil prices increases. Reilly and Witt (1998) also analyzed the UK market and revisited the evidence of Bacon (1991) by applying a restricted ECM with monthly data from 1982 to 1995, and observed the asymmetry behaviours for gasoline prices when crude oil increases. Galeotti et al. (2003) examined five European countries (i.e., Germany, France, UK, Italy and Spain) with monthly data range from 1985 to 2000 based on ECM; they confirmed the asymmetric price adjustment mechanism of gasoline markets. Grasso and Manera (2007) investigated above five European countries again with threes asymmetric ECM and found evidence of asymmetry for all countries under

consideration. However, Godby et al. (2000) applied TAR methodologies to test for 13 Canadian cities gasoline's asymmetric response to the crude oil prices for the period range from January 1990 to December 1996, but failed to find any evidence on the asymmetric effects. Bettendorf et al. (2003) investigated the retail price adjustments in the Dutch gasoline market by applying an asymmetric ECM on weekly price changes for period starting from 1996 to 2001, and reported that the asymmetric behaviours various depending on the choice of the day for which the prices were observed. Furthermore, several studies aimed to explain the potential causes for the asymmetry responses on output prices (e.g., Meyer and von Cramon-Taubadel, 2004), such as market power, production lags and inventory costs and search theory (see Borenstein et al. 1997; Brown and Yücel, 2000; Peltzman, 2000; Johnson, 2002; Kaufmann and Laskowski, 2005).

Overall, most studies have found the existence of the long term relationships between crude oil and petroleum products; and most often, there were also empirical evidences on asymmetric price adjustment mechanism of petroleum products (e.g., gasoline) to oil prices increases.

2.3.4 Conclusion

Empirical research in the area of data modeling is generally concerned with one of the following three questions; namely, are crude oil markets consistent with the market efficiency hypothesis? Are there any relationships between oil prices and economic variables? Are there any relationships between oil prices and prices of petroleum products? On the other hand, research on forecasting crude oil addresses several crude oil related variables such as prices, returns, supply, and demand. As far as prices and returns are concerned, quantitative forecasting models could be divided into three main categories; namely, non-artificial intelligence models, artificial intelligence models, and hybrid models. The overall results were mixed for all three categories.

Several issues have drawn my attention. First most macroeconomic variables are only available at quarterly frequency and this would cause a potential problem when include

more exogenous variables and their lagged variables. Second, most studies tend to pay little attention on the robustness of the results for selecting the appropriate lag length for a VAR. Third, most studies did not provide much information on the validity of their models. For example, there might be a problem of multicollinearity when applying for certain asymmetry tests was addressed by Houck (1977), Gathier and Zapata (2001). Fourth, most studies found that economic activities tend to respond asymmetrically to oil price shocks (e.g., Balke and Fomby 1997; Goodwin and Holt, 1999; Huntington, 2007); and different approaches have proposed in the literature to capture the impact from positive and negative increments on the oil prices, and this led to a concern in terms of how to select the appropriate thresholds and the significance of these thresholds. In addition, there is a lack of rigorous comparison and analysis of the strengths and weaknesses of the available methods (Meyer and von Cramon-Taubadel, 2004).

2.4 Literature Review on Forecasting Crude Oil Levels and Volatilities

Oil price is a key factor which affects economic plans and decisions of governments and commercial firms, so a proactive knowledge of its future movements can lead to better decisions at various governmental and managerial levels. Therefore, there is a rich body of literature to forecast the oil prices levels and/or its volatilities. Depending on the nature of the methodologies, we classify the literature on the levels and volatilities into three sub-categories; namely, Non-Artificial Intelligence (AI) methods, AI methods and hybrid methods. To be more specifically, the Non-AI approaches include conventional models such as univariate and multivariate time series models, regression models (e.g., linear regression models, nonlinear regression models); error correction models (e.g., ECM and VECM) etc. Most often, research in this category are mainly assuming the existences of the relationships with other related variables based on equilibrium theories (e.g., return to storages; future market equilibrium; supply and demand equilibrium), and research on the data modeling provide valuable input to model building in this type of forecasting. On the other hand, AI based models (e.g., Artificial Neural Networks, Genetic Programming) and hybrid approaches (i.e., combination of different methods), are tend to rely less on the theories and there is a trend for applying more complex

methodologies to predict the highly volatile and chaotic oil prices series in the forecasting literature recently. Furthermore, various explanatory variables have been used in these analyses, and could be divided into six main categories; namely, price variables and their lagged values (e.g., WTI, Brent, OPEC, and Crude oil derivatives), supply and demand variables (i.e., world, OPEC, Non-OPEC, OECD, US demand and/or supply), inventories (i.e., strategic reserves, industrial stocks, and relative inventory levels), economic variables (e.g., GDP, inflation and interest rates, exchange rates, tax variables, and market share), specific event-related variables and structural change (e.g., September 11, 2001, Gulf Wars, and OPEC Quota Tightening – April 1999), and political variables (i.e., Amicability among OPEC members, and political environment).

The remaining of this section is organized as follows. In section 2.4.1, we review and evaluate the literature related to forecast crude oil prices. Section 2.4.2, we survey and evaluate the studies on forecasting volatilities of crude oil prices. Finally, we conclude the section.

2.4.1. Forecasting Crude Oil Levels

The literature on forecasting the crude oil levels could be divided into three main categories depending on the nature of the methodologies; namely, Non- AI methods, AI methods and hybrid methods. First, most empirical studies use non AI methods are mainly based their forecast on equilibrium theories (e.g., returns to storage, futures markets equilibrium) to explore the dynamic behavior of crude oil prices with other variables. For example, Zeng and Swanson (1998) incorporated the cost of carry theory into their error correction models and compared the performances with other models (e.g., RW, VAR), by using various statistical measures (e.g., trace of MSE, MAPE) and statistical tests (e.g., Diebold & Mariano test, and Confusion Matric based test) and obtained inconsistent results. Sequeira and McAleer (2000) studied two Brent crude oil futures with Unbiased Expectations Hypothesis, Cost-of-Carry, ECM models, and ARMA – GARCH (1, 1), the relative performances of those competing forecasting models were mixed based on RMSE and MAE. Longo et al. (2007) forecasted the crude

oil prices with different models (e.g., AR, ECM, regression models), as well as different data frequencies (i.e., daily, weekly, monthly, and quarterly) for the period ranging from Jan 1986 to Dec 2005, the overall results were mixed based on various statistical measures (e.g., MAPE, MAE, Theil, and RMSE). Knetsch (2007) proposed a convenience yield based forecasting model by using both monthly spot and futures of Brent crude oil prices for the period of April 1991 and October 2006, and he computed the recursive out-of-sample forecasts between 1 and 11 months ahead and compared its performances with two benchmarking models (RW, Unadjusted futures prices approach), convenience yield based model was not be able to outperform the benchmark models based on the RMSE, ME and direction-of-change prediction. Coppola (2008) used VECM to explore the dynamic relationship between spot and futures prices based on cost of carry theory, but based on different statistical measures (e.g., MSE) and statistical tests (e.g., DM, Market Timing), VECM did not constantly beat the random walk. Murat and Tokat (2009) recently examined the relationship between oil prices and crack spread futures prices traded on NYMEX with a VECM, based on RMSE, Theil Inequity coefficient and DM tests, VECM was not clearly beaten the benchmark model (i.e., RW).

Furthermore, petroleum products are distillated from the crude oil, so it is common to forecast the crude oil prices by examining the long term relationships among crude oil prices and petroleum products prices. For example, Lanza et al. (2005) analyzed the price dynamics between crude oil and petroleum products (e.g., premium gasoline, gasoil, low and high sulphur fuel oil) by using ECM and VAR methodologies, a number of statistical measures (e.g., MAPE, Theil's inequality coefficient and the success ratio) have been used to evaluate the performance of those competing forecasting models, but they failed to reach a consensus. Pindyck (1999) examined the long-run behavior of real energy prices (e.g., oil, coal and natural gas) in the U.S. with Kalman Filtering methods and obtained mixed results. Benard et al. (2004) tested the statistical significance of Pindyck's (1999) models and found statistically significant instabilities for coal and natural gas prices, but not for crude oil prices, the results were mixed by comparing their results based on two different forecasting periods. Chantziara and Skiadopoulos (2008)

applied Principal Components Analysis (PCA) to forecast the futures prices (i.e., WTI crude oil, Brent crude oil, heating oil, and unleaded gasoline) over the period of 1993 - 2006. The out of sample performance for five models (i.e., AR (1), VAR (1), ARMA (1, 1), the PCA and joint PCA) were mixed based on different criteria and measures (i.e., RMSE, Theil Coefficients, MAE, direction of change).

In addition, several studies based their forecasts on other types of relationships. For example, Ye et al. (2002, 2005, 2006a, b) observed a negative relationship between the oil prices and oil inventories, and they proposed their forecasting models based on OECD petroleum inventory series and its derivate series (e.g., high and low relative inventory levels), the results were mixed for the competing forecasting models based on different performance measures such as (MAE, MAPE, Theil U etc). Dees et al. (2008) first tested for the effect of downstream condition (e.g., refinery utilization rates, OPEC capacity utilization) on crude oil prices, and failed to find any evidence to support a negative relationship between the refining capacity and crude oil prices for the period of 1986 – 2006; they second performed out of sample one step ahead forecast for the period ranging from Q2 1999 to Q1 2007 by using error correction model and random walk, but received different rankings based on Diebold & Mariano test, and Clark & McCracken encompassing test.

Second, there has been an increased popularity for using, adopting and/or proposing more sophisticated AI models to forecast the highly volatile and chaotic oil prices. For instance, Abramson and Finizza (1991) and Abramson (1994) introduced Belief networks (i.e., knowledge-based models) that could incorporate a large number of explanatory variables (e.g., tax, GDP, supply, demand, and production of the world oil market). Abramson and Finizza (1995) then applied the combination of Belief Network and Probabilistic Model to forecast probabilistic forecast in the oil markets. Kaboudan (2001) used Genetic Programming (GP), Artificial Neural Networks (ANNs) to forecast the monthly crude oil prices, GP and ANNs performed better than the RW model based on MSE. More recently, wavelet transformations have been adopted and modified to the area of forecasting; for example, Yousefi et al. (2005) proposed a wavelet-based

prediction procedure to investigate the market efficiency in WTI crude oil prices and NYMEX futures prices over 1, 2, 3 and 4 step ahead forecasting horizons, they noticed that the predictive power of wavelet-based procedure was highly sensitive to the underlying sample size. Furthermore, Xie et al. (2006) proposed a new approach to forecast crude oil prices for the period of Jan 1970 to Dec 2003 based on the support vector machine (SVM), and SVM outperformed ARIMA and Back propagation neural network (BPNN) based on the RMSE and Directional Change statistics. Fernandez (2006) forecasted the crude oil and natural gas spot prices with ARIMA, ANN and SVM, and found mixed results for different forecasting horizons based on MSPE and encompassing tests. Matilla-Garcia (2007) forecasted three futures returns (i.e., natural gas, unleaded gasoline and light crude oil) by using GA, random walk and GARCH (1, 1), but GA did not performance superior than the other two benchmark models based on MSE, MAD and Diebold and Mariano tests.

Last but not the least, another popular approach to forecast the crude oil prices was based on hybrid forecasting models. For example, Wang et al. (2004, 2005) designed a novel hybrid model, which was a systematic integration of ANN, rule based expert system (RES) together with web-based text mining (WTM) techniques, to forecast monthly WTI crude oil prices for the period of Jan 1970 up to Dec 2003; based on RMSE and Directional change rate measures, they confirmed their TEI@I methodology framework was satisfactory. Liu et al. (2007) applied a fuzzy neural network model which combined RBF neural network, Markov Chain based semi-parametric model and Wavelet analysis based model, to forecast the daily Brent crude oil prices for the period ranging from 1987 to 2006, and found their proposed hybrid model produced the smallest MSE compared to the three individual models. Amin-Naseri and Gharacheh (2007) introduced a hybrid ANN model which combined k-means technique and genetically-evolved feed-forward neural network to forecast the WTI crude oil prices for the period of Jan 1983 and Dec 2006, they compared their hybrid ANN model with other models which have been used in other papers (e.g., TEI@I methodology from Wang et al., 2004, 2005; GP used by Kaboudan 2001); but based on various statistical measures and tests, their proposed hybrid ANN model was worse than other benchmark models.

Fang et al. (2008) introduced a generalized pattern matching based on genetic algorithm (GPMGA) method to forecast both daily Brent and WTI crude oil prices, and compared with two other standard AI models (i.e., Elman Networking, Pattern modelling and recognition system), based on RMSE and MAPE, their model performed best. Ghaffari and Zare (2009) proposed a Adaptive Neural Fuzzy Inference Systems (i.e., a combination of ANN and fuzzy logic) to forecast the WTI crude oil prices, based on the correct sign prediction, they found out 66% of daily variation predictions were consistent with the actual values.

Overall, a large number of studies have been carried out in the field of forecasting crude oil levels, but there have been inconsistent and often confusing results reported for both within and across studies. The potential reason for causing this problem is that most studies tend to use one or several performance criteria (i.e., goodness-of-fit, correct sign, informational efficiency) and for each criterion, one or several statistical measures (e.g., MSE, MAPE) and/or statistical tests (e.g., Diebold & Mariano test, and Clark & McCracken encompassing test) are used to assess the performance of competing models; the assessment exercise of competing models is generally restricted to their ranking by measure. Thus, the current methodology is unidimensional in nature and there is a need for the multidimensional framework.

2.4.2 Forecasting Crude Oil Volatilities

Volatility is an important input in diverse areas, such as asset pricing, option pricing, portfolio selection models, risk management, policy making and financial regulation among others. Therefore, it has attracted increasing attention on forecasting the volatility of crude oil prices and/or returns. For example, Day and Lewis (1993) used GARCH(1,1), EGARCH(1,1), implied volatility and historical volatility models to forecast crude oil volatilities based on the daily data for the period ranging from November 1986 to March 1991. They have used several performance criteria (e.g., Goodness-of-fit and Biasedness) and statistical tests (e.g., ME, MAE, RMSE) to evaluate those competing forecasting models, but the out-of-sample results were mixed for which model has outperformed the rest. Duffie and Gray (1995) used implied

volatility, GARCH (1, 1), EGARCH (1, 1), Bi-Variant GARCH and regime switching models to forecast the crude oil, heating oil, and natural gas volatilities for period of May 1988 to July 1992, and found implied volatility model performed the best based on RMSE. Morana (2001) proposed a semiparametric bootstrap approach to forecast the volatility of Brent oil prices over the period of November 1998 to January 1999, and suggested that one-month forward prices did not help much for forecasting the spot Brent oil prices. Fong and See (2002) examined the temporal behavior of volatility on the WTI with Markov regime switching model, GARCH (1, 1) and constant variance models for the period spanned from January 1992 to December 1997, the rankings were different when using different statistical measures (e.g., MSE, MAE, and R-Squares). Sadorsky (2005) compared the forecasting performances of range – based stochastic volatility model with other univariate models (e.g., RW, MA, ES, AR, linear regression) based on various statistical measures (e.g., MSE, MAD, MPE, and MAPE) and statistical tests (e.g., DM, MDM, regression tests on biasness), and obtained mixed results. Sadorsky (2006) forecasted volatility of several petroleum futures returns (e.g., WTI crude oil, heating oil No. 2, unleaded gasoline and nature gas) by using a large number of models (e.g., RW, historical mean, MA, ES, linear regression, AR, GARCH, TGARCH, GARCH-M, state space, VAR, and Bi-Variant GARCH models). He obtained inconsistent results by evaluating those competing forecasting models with different statistical measures (e.g., MSE, MAD, Theil U Coefficients) and statistical tests (e.g., DM, market timing tests). Marzo and Zagaglia (2007) compared the forecasts of oil volatility using of GARCH, EGARCH and TGARCH models, using daily futures prices on crude oil traded in NYMEX from January 1995 to November 2005, they did not find an constant superior model based on different statistical measures (e.g., MSE, MAD, Success Ratio, Heteroskedasticity adjusted MSE) and statistical test (e.g., Direction accuracy, Diebold and Mariano). Kang et al. (2009) forecasted the volatility of three crude oil markets - Brent, Dubai and WTI, with GARCH, Integrated GARCH (IGARCH), component GARCH (CGARCH) and Factional Integrated GARCH (FIGARCH) models. They evaluated the performance of those competing forecasting models by using MSE, MAE and Diebold & Mariano test statistic, and suggested that

CGARCH and FIGARCH models were useful for modelling and forecasting persistence in the volatility of crude oil prices. Agnolucci (2009) compared the predictive ability of WTI futures returns' volatilities from 1992 – 2005 with different GARCH types of models (e.g., GARCH, APARCH, EGARCH, CGARCH, and TGARCH) and Implied volatility models (e.g., Black Scholes Model), but their relative performances were inconsistent with respect to different statistical measures (e.g., MAE, MSE) and statistical test (e.g., regression-based test for biasdness).

To sum it up, the literature on forecasting the crude oil volatility are mainly using the non-AI models. For example, univariate and multivariate time series models (e.g., ARMA, VAR models); Autoregressive Conditional Heteroscedasticity (ARCH) models and their variants (e.g., univariant GARCH types and multivariate GARCH models), Stochastic Volatility (SV) models, Regime switching models and threshold models. Therefore, it would be interesting to adopt or propose new AI and hybrid models to forecast the crude oil volatility.

2.4.3 Conclusion

In conclusion, various methodologies have been adapted, proposed and used in the literature of forecasting crude oil prices levels and volatilities. The empirical results were mixed both within and across studies and might due to one or more of the following aspects: different time frequency (daily, weekly, monthly etc), different period of the data, data transformations (deflated, and/or deseasonlized, and/or logarithmic, and/or returns, and/or normalized), forecasting models (e.g., Non-AI models, AI models and hybrid models), explanatory variables (e.g., demand and supply variables, inventories, economic variables, dummy variables for specific event-related variables and structural change), performance criteria (i.e., goodness-of-fit, correct sign, informational efficiency), statistical measures (e.g., MSE, MAPE), and statistical tests (e.g., Diebold & Mariano test, and Clark & McCracken encompassing test).

Several issues have drawn my attention. First, there were more studies carried out on analyzing and forecasting the level rather than the volatility. Second, most studies on

forecasting crude oils volatilities tend to use traditional statistical models (e.g., ARIMA, Implied Volatility, GARCH etc), where less artificial intelligences models or hybrid models seem to be used. Third, most studies did not discuss much on the choice of their chosen statistical measures and statistical tests. In fact, the choice of a particular measure might attribute to the inconsistent results within the studies. For example, the MSE tend to penalize the large errors (e.g., greater than 1) and soothe the small errors (e.g., errors between 0 and 1), whereas the RMSE penalize the small errors and soothe the large errors. Finally, although several performance criteria and measures are used in some papers (e.g., Sadorsky, 2005, 2006), the assessment exercise of competing models is generally restricted to their ranking by measure. Therefore, the current methodology is unidimensional in nature and there is a lack of multidimensional framework for relative performance evaluation of competing forecasting models.

References

Abosedra S, Baghestani H. On the predictive accuracy of crude oil future prices. Energy Policy 2004; 32; 1389–1393.

Abel A.B., Bernanke B.S. Macroeconomics. Addison Wesley Longman Inc: Boston; 2001.

Abramson B. The design of belief network-based systems for price forecasting. Computers and Electrical Engineering 1994; 20; 163–180.

Abramson B, Finizza A. Using belief networks to forecast oil prices. International Journal of Forecasting 1991; 7; 299–315.

Abramson B, Finizza A. Probabilistic forecasts from probabilistic models: a case study in the oil market. International Journal of Forecasting 1995; 11; 63–72.

Adrangi B, Chatrath A, Dhanda K.K., Raffiee K. Chaos in Oil prices? Evidence from future markets. Energy Economics 2001; 23; 405-425.

Adrangi B, Chatrath A, Raffiee K, Rippe R.D. Alaska North Slope crude oil price and the behavior of diesel prices in California. Energy Economics 2001; 23; 29-42.

Akarca A.T., Andrianacos D. The relationship between crude oil and gasoline prices. International Advances in Economic Research 1998; 4; 282–288.

Al-Gudhea S, Kenc T, Dibooglu S. Do retail gasoline prices rise more readily than they fall? A threshold cointegration approach. Journal of Economics and Business 2007; 59; 560–574.

Al-Mudhaf A, Goodwin T.H. Oil shocks and oil stocks: evidence from the 1970s, Applied Economics 1993; 25; 181–190.

Alvarez-Ramirez J, Cisneros M, Ibarra-Valdez C, Soriano A. Multifractal Hurst Analysis of crude oil prices. Physica A 2002; 313; 651-670.

Alvarez-Ramirez J, Soriano A, Cisneros M, Suarez R. Symmetry/anti-symmetry phase transitions in crude oil markets. Physica A 2003; 322; 583–596.

Amano R.A., Norden S. Oil prices and the rise and fall of the US real exchange rate. Journal of International Money and Finance 1998; 17; 299–316.

Asche F, Osmunddsen P, Sandssmark M. The UK market for natural gas, oil and electricity: are the prices decoupled? The Energy Journal 2006; 27; 27–40.

Asche F, Gjolberg O, Volker T. Price relationships in the petroleum market: an analysis of crude oil and refined product prices. Energy Economics 2003; 25; 289-301.

Bachmeier L.J., Griffin J.M. New evidence on asymmetric gasoline price responses. Review of Economics and Statistics 2003; 85; 772–776.

Bachmeier L.J., Griffin J.M., 2006. Testing for market integration: crude oil, coal, and natural gas. The Energy Journal 27; 2; 55–71.

Bacon R. Modelling the price of oil. Oxford Review of Economic Policy 1991; 7; 17–34.

Baffes J. Oil spills on other commodities. Resources Policy 2007; 32; 126-134.

Balke N.S., Brown P.A. Yücel M.K. Crude oil and gasoline prices: an asymmetric relationship? Federal Reserve Bank of Dallas Economic Review 1998; 2–11.

Balke N.S., Fomby T.B. Threshold cointegration. International Economic Review 1997; 38; 627–645.

Barsky R.B., Kilian L. Oil and the Macroeconomy Since the 1970s. Journal of Economic Perspectives 2004; 18; 115-134.

Basher S.A., Sadorsky P. Oil price risk and emerging stock markets. Global Finance Journal 2006; 17; 224–251.

BBC News. Katrina set to raise oil prices. 28th August 2005; http://news.bbc.co.uk/1/hi/business/4192528.stm.

BBC News. What is driving Oil Price so high? 2nd January 2008; http://news.bbc.co.uk/go/pr/fr//1/hi/business/7048600.stm.

BBC News. What is Keeping Oil Prices so high? 23rd June 2008; http://news.bbc.co.uk/1/hi/business/7469124.stm.

BBC News. Why are oil prices so high? 28th September 2004; http://news.bbc.co.uk/go/pr/fr/-/1/hi/business/3708951.stm.

Bekiros S, Diksa C. The relationship between crude oil spot and futures prices: Cointegration, linear and nonlinear causality. Energy Economics 2008; 30; 2673-2685.

Benassy-Quere A, Mignon V, Penot A. China and the relationship between the oil price and the dollar. Energy Policy 2007; 35; 5795–5805

Bentzen J. Does OPEC influence crude oil prices? Testing for co-movements and causality between regional crude oil prices. Applied Economics 2007; 39; 1375-1385.

Bernanke B.S. Irreversibility, Uncertainty, and Cyclical Investment. Quarterly Journal of Economics 1983; 98; 85-106.

Bernard J.T., Khalaf L, Kichian M. Structural change and forecasting long-run energy prices. Bank of Canada 2004; Working Papers 04-5.

Bettendorf L, Van der Geest S.A., Varkevisser M. Price asymmetry in the Dutch retail gasoline market. Energy Economics 2003; 25; 669–689.

Blair B.F., Rezek J.P. The effects of Hurricane Katrina on price pass-through for Gulf Coast gasoline. Economics Letters 2008; 98; 229-234.

Bohi D.R. On the Macroeconomic Effects of Energy Price Shocks. Resources and Energy 1991; 13; 145-162.

Bopp A.E., Sitzer, S. Are petroleum futures prices good predictors of cash value. The Journal of Futures Markets 1987; 7; 705–719.

Bopp A. E., Lady G.M. A comparison of petroleum futures versus spot prices as predictors of prices in the future. Energy Economics 1991; 13; 274-282.

Borenstein S, Cameron A.C., Gilbert R. Do gasoline prices respond asymmetrically to crude oil price changes? Quarterly Journal of Economics 1997; 112; 305–339.

Boyer M.M., Filion D. Common and fundamental factors in stock returns of Canadian oil and gas companies. Energy Economics 2007; 29; 428–453.

Brown S.P.A., Yücel M.K. Oil prices and U.S. aggregate economic activity: A question of neutrality. Federal Reserve Bank of Dallas Economic and Financial Review 1999; 16–53.

Brown S.P.A., Yücel M.K. Energy prices and aggregate economic activity: an interpretative survey. Quarterly Review of Economics and Finance 2002; 42; 193–208.

Bruno M, Sachs J. Input Price Shocks and the Slowdown in Economic Growth: The Case of U.K. Manufacturing. Review of Economic Studies 1982; 49; 679-705.

Burbidge J, Harrison A. Testing for the Effects of Oil-Price Rises Using Vector Autoregression. International Economic Review 1984; 25; 459-484.

Camarero M. Tamarit C. Oil prices and Spanish competitiveness: A cointegrated panel analysis. Journal of Policy Modeling 2002; 24; 591–605.

Ciner C. Energy shocks and financial markets: nonlinear linkages, Studies in Nonlinear Dynamics and Econometrics 2001; 5; 203–212.

Chang Y. Wong J.F. Oil price fluctuations and Singapore economy. Energy Policy 2003; 31; 1151–1165.

Chen S.S., Chen H.C. Oil prices and real exchange rates. Energy Economics 2007; 29; 390–404.

Chen L.H., Finney M. Lai K.S. A threshold cointegration analysis of asymmetric price transmission from crude oil to gasoline prices. Economics Letters 2005; 89; 233–239.

Coppola A. Forecasting oil price movements: exploiting the information in the futures market. Journal of Futures Market 2008; 28; 34–56.

Crowder W. J., Hamid A. A cointegration test for oil futures market efficiency. Journal of Futures Markets 1993; 13; 933–941.

Cuñado J, Perez de Gracia F. Do Oil Price Shocks Matter? Evidence from Some European Countries. Energy Economics 2003; 25; 137-154

Darby M.R. The price of oil and world inflation and recession. The American Economic Review 1982; 72; 738–751.

Davis M.C., Hamilton J.D. Why Are Prices Sticky? The Dynamics of Wholesale Gasoline Prices. Journal of Money, Credit, and Banking 2004; 36; 17-37.

Day T.E., Lewis C.M. Forecasting futures market volatility. The Journal of Derivatives 1993; 1; 33–50.

DeLong J.B., Shleifer A, Summers L.H., Waldmann R.J. Positive Feedback investment strategies and destabilizing rational speculation. Journal of Finance 1990; 45; 379-395.

Dufour A, Engle R. Time and the Price Impact of a Trade. Journal of Finance 2000; 55; 2467-2498.

Duffy-Deno K.T. Retail price asymmetries in local gasoline markets. Energy Economics 1996; 18; 81–92.

Duncan E. The heat is on: a survey of climate change. The Economist 2006; 380; 8494.

Duncan E. Cleaning up: a survey of climate change. The Economist 2007; 383; 8531.

El-Sharif I, Brown D, Burton B, Nixon B, Russell A. Evidence on the nature and extent of the relationship between oil prices and equity values in the UK. Energy Economics 2005; 27; 819–830.

Enders W, Dibooglu D. Long-run purchasing power parity with asymmetric adjustment. Southern Economic Journal 2001; 68; 433–445.

Energy Information Administration (EIA). Annual Energy Review 2006; DOE/EIA-0384; http://tonto.eia.doe.gov/FTPROOT/multifuel/038406.pdf

Energy Information Administration (EIA). Short Term Energy Outlook; February 2008; http://www.eia.doe.gov/pub/forecasting/steo/oldsteos/feb08.pdf.

Energy Information Administration (EIA). EIA Oil market chronology; 2008; http://www.eia.doe.gov/emeu/cabs/AOMC/Overview.html.

Ewing B.T., Harter C.L. Co-movements of Alaska North Slope and UK Brent crude oil prices. Applied Economics Letters 2000; 7; 553–558.

Faff R, Brailsford T.J. Oil price risk and the Australian stock market. Journal of Energy Finance and Development 1999; 4; 69–87.

Fan Y, Liang Q, Wei Y.M. A generalized pattern matching approach for multi-step prediction of crude oil price. Energy Economics 2006; 30; 889–904.

Ferderer J.P. Oil price volatility and the macroeconomy: a solution to the asymmetry puzzle. Journal of Macroeconomics 1996; 18; 1–16.

Fernandez V. Forecasting crude oil and natural gas spot prices by classification methods. No 229, Documentos de Trabajo from Centro de Economía Aplicada, Universidad de Chile.

Fong W.M., See K.H. A Markov switching model of the conditional volatility of crude oil futures prices. Energy Economics 2002; 24; 71–95.

Galeotti M, Lanza A, Manera M. Rockets and feathers revisited: an international comparison on European gasoline markets. Energy Economics 2003; 25; 175–190.

Gately D, Huntington H.G. The asymmetric effects of changes in price and income on energy and oil demand. Energy Journal 2002; 23; 19–55.

Gauthier W.M., Zapata H. Testing Symmetry in Price Transmission Models. Louisiana State University, Department of Agricultural Economics and Agribusiness, Working Paper 2001.

Gisser M, Goodwin T.H. Crude oil and the macroeconomy: Tests of some popular notions. Journal of Money, Credit and Banking 1986; 18; 95–103.

Ghaffari A, Zare S. A novel algorithm for prediction of crude oil price variation based on soft computing. Energy Economics 2009; 31; 531-536

Godby R, Lintner A.M., Stengos T. Wandschneider B. Testing for asymmetric pricing in the Canadian retail gasoline market, Energy Economics; 2000; 22; 349–368.

Goodwin B, Holt M. Price transmission and asymmetric adjustment in the US beef sector. American Journal of Agricultural Economics 199; 81; 630–637.

Green S.L., Mork K.A. Toward efficiency in the crude oil market. Journal of Applied Econometrics 1991; 6; 45–66.

Gulen S.G. Regionalization in world crude oil markets: Further evidence. The Energy Journal 1999; 20; 125–139.

Hamilton J.D. Oil and the macroeconomy since World War II. Journal of Political Economy 1983; 91; 228–248.

Hamilton J.D. Historical causes of postwar oil shocks and recessions. Energy Journal 1985; 6; 97–116.

Hamilton J.D. A neoclassical model of unemployment and the business cycle. Journal of Political Economy 1988; 96; 593–617.

Hamilton J.D. This is what happened to the oil price–macroeconomy relationship. Journal of Monetary Economics 1996; 38; 215–220.

Hamilton J.D. A parametric approach to flexible nonlinear inference. Econometrica 2001; 69; 537–573.

Hamilton J.D. What Is an Oil Shock? Journal of Econometrics 2003; 113; 363-398.

Hamilton J.D. Oil and the Macroeconomy. In Durlauf S, Blume L (Eds), The New Palgrave Dictionary of Economics. Palgrave MacMillan Ltd. 2008.

Hamilton J.D. Understanding Crude Oil Prices. Energy Journal 2009; 30; 179-206.

Hammoudeh S, Eleisa L. Dynamic relationships among GCC stock markets and NYMEX oil futures. Contemporary Economic Policy 2004; 22; 250–269.

Hammoudeh S, Li H, Jeon B. Causality and volatility spillovers among petroleum prices of WTI, gasoline and heating oil in different locations. North American Journal of Economics and Finance 2003; 14; 89–114.

Hansen H, Johansen S. Some tests for parameter constancy in cointegrated VAR-models. Econometrics Journal 1999; 2; 306–333.

Henriques I, Sadorsky P. Oil prices and the stock prices of alternative energy companies. Energy Economics 2008; 30; Pages 998-1010.

Hickman B, Huntington H, Sweeney J. Macroeconomic Impacts of Energy Shocks. North-Holland: Amsterdam; 1987.

Hooker M. What happened to the oil price–macroeconomy relationship? Journal of Monetary Economics 1996; 38; 195–213.

Houck J.P. An Approach to specifying and estimating nonreversible Functions. American Journal of Agricultural Economics 1977; 59; 570-572.

Huang B.N., Hwang M. J., Peng H.P. The Asymmetry of the impact of oil price shocks on economic activities: an application of the multivariate threshold model. Energy Economics 2005; 27; 455-476.

Huang Y, Guo F. The role of oil price shocks on China's real exchange rate. China Economic Review 2007; 18; 244–265.

Huang R.D., Masulis R.W., Stoll H.R. Energy shocks and financial markets. Journal of Futures Markets 1996; 16; 1−27.

Huntington H. Crude oil prices and U.S. economic performance: Where does the asymmetry reside? Energy Journal 1998; 19; 107–132.

Huntington H. Shares, gaps and the economy's response to oil disruptions. Energy Economics 2004; 26; 415–424.

Huntington H. Industrial natural gas consumption in the United States: an empirical model for evaluating future trends. Energy Economics 2007; 29; 743–759.

Jiménez-Rodríguez R, Sánchez M. Oil price shocks and real GDP growth: empirical evidence for some OECD countries. Applied Economics 2005; 37; 201–228.

Johnson R.N. Search costs, lags and prices at the pump. Review of Industrial Organization 2002; 20; 33–50.

Jones C.M., Kaul G. Oil and stock markets. Journal of Finance 1996; 51; 463-91.

Jones D.W., Leiby P.N., Paik I.K. Oil price shocks and the macroeconomy: what has been learned since 1996. The Energy Journal 2004; 25; 1-32.

Kaboudan M.A. Compumetric forecasting of crude oil prices. Proceedings of the 2001 Congress on Evolutionary Computation 2001; 1; 283-287.

Kang S.H., Kang S.M., Yoon S.M. Forecasting volatility of crude oil markets. Energy Economics. Energy Economics 2009; 31; 119–125.

Karrenbrock J.D. The behavior of retail gasoline prices: symmetric or not? Federal Reserve Bank of St. Louis Review 1991; July; 19–29.

Kaufmann R.K., Laskowski C. Causes for an asymmetric relation between the price of crude oil and refined petroleum products. Energy Policy 2005; 33; 1587–1596.

Knetsch T.A. Forecasting the price of crude oil via convenience yield predictions. Journal of Forecasting 2007; 6; 527-549.

Kilian L. Exogenous oil supply shocks: how big are they and how much do they matter for the U.S. economy? Review of Economics and Statistics 2008a; 90; 216–240.

Kilian L. A comparison of the effects of exogenous oil supply shocks on output and inflation in the G7 countries. Journal of the European Economic Association 2008b; 6; 78–121.

Kilian L. Not all oil price shocks are alike: disentangling demand and supply shocks in the crude oil market. American Economic Review 2009; 99; 1053-1069.

Kilian L, Park C. The Impact of Oil Price Shocks on the U.S. Stock Market. Forthcoming: International Economic Review.

Kilian L, Rebucci A, Spataford N. Oil shocks and external balances. Journal of International Economics 2009; 77; 181-194.

Kim I.M. Loungani P. The role of energy in real business cycle models. Journal of Monetary Economics 1992; 29; 173–189.

Kleit A.N. Are regional oil markets growing closer together? An arbitrage cost approach. The Energy Journal 2001; 22; 1–15.

Lalonde R, Zhu Z, Demers F. Forecasting and analyzing world commodity prices. Bank of Canada 2003; No. 24.

Lanza A, Manera M. Grasso M, Giovannini M. Long-run models of oil stock prices. Environmental Modelling and Software 2005; 20; 1423-1430.

Lardic S, Mignon V. The impact of oil prices on GDP in European countries: an empirical investigation based on asymmetric cointegration. Energy Policy 2006; 34; 3910–3915.

Lardic S, Mignon V. Oil prices and economic activity: An asymmetric cointegration approach. Energy Economics 2008; 30; 847 – 855.

Lee K, Ni S, Ratti R. Oil shocks and the macroeconomy: The role of price variability. The Energy Journal 1995; 16; 39–56.

Lin S.X., Tamvakis M.N. Spillover effects in energy futures markets. Energy Economics 2001; 23; 43–56.

Liu J.L., Bai Y, Li B. A New approach to forecast crude oil price based on Fuzzy Neural Network. Fourth International Conference on Fuzzy Systems and Knowledge Discovery 2007; 3; 273-277.

Lo M, Zivot E. Threshold cointegration and nonlinear adjustment to the law of one price. Macroeconomic Dynamics 2001; 5; 533–576.

Longo C, Manera M, Markandya A, Scarpa E. Evaluating the empirical performance of alternative econometric models for oil price forecasting. FEEM Working Paper 2007; No 4.

Loungani P. Oil price shocks and the dispersion hypothesis. Review of Economics and Statistics 1986; 68; 536–539.

Maghyereh A. Oil price shocks and emerging stock markets: a generalized VAR approach. International Journal of Applied Econometrics and Quantitative Studies 2004; 1; 27-40.

Malik F, Hammoudeh S. Shock and volatility transmission in the oil, US and Gulf equity markets. International Review of Economics and Finance 2007; 16; 357–368.

Meyer J, von Cramon-Taubadel S. Asymmetric price transmission: A survey. Journal of Agricultural Economics 2004; 55; 581 – 611.

Mirmirani S, Li H.C. A comparison of VAR and neural networks with genetic algorithm in forecasting price of oil. Advances in Econometrics 2004; 19; 203–223.

Murat A, Tokat E. Forecasting oil price movements with crack spread futures. Energy Economics 2009; 31; 85-90.

Moosa I.A., Al-Loughani N.E. Unbiasedness and time varying risk premia in the crude oil futures market. Energy Economics 1994; 16; 99–105.

Morana C. A semiparametric approach to short-term oil price forecasting. Energy Economics 2001; 23; 325–338.

Moshiri S, Foroutan F. Forecasting nonlinear crude oil futures prices. The Energy Journal 2006; 27; 81–95.

Mork K.A. Oil and the macroeconomy when prices go up and down: An extension of Hamilton's results. Journal of Political Economy 1989; 91; 740–744.

Mork K.A. Business cycles and the oil market. The Energy Journal 1994; 15; 15–38.

Mork K.A., Olsen O, Mysen H.T. Macroeconomic responses to oil price increases and decreases in seven OECD countries. Energy Journal 1994; 15; 19-35.

Nandha M, Faff R. Does oil move equity prices? A global view. Energy Economics 2008; 30; 986–997.

Narayan P.K., Narayan S. Modelling oil price volatility. Energy Policy 2007; 35; 6549–6553.

Narayan P.K., Narayan S, Smyth R. Are oil shocks permanent or temporary? Panel data evidence from crude oil and NGL production in 60 countries. Energy Economics 2008; 30; 919–936.

New York Times. Record failures at oil refineries raise gas prices. 22nd July 2007. http://www.nytimes.com/2007/07/22/business/22refine.html.

Olomola P.A., Adejumo A.V. Oil price shocks and macroeconomic activities in Nigeria. International Research Journal of Finance and Economics 2006; 3; 28–34.

Olson M. The productivity slowdown, the oil shocks, and the real cycle. Journal of Economic Perspectives 1988; 2; 43-69.

Panagiotidis T, Rutledge E. Oil and Gas Markets in the UK: A cointegrating approach. Energy Economics 2007; 29; 329-347.

Papapetrou E. Oil price shocks, stock markets, economic activity and employment in Greece. Energy Economics 2001; 23; 511−532.

Panas E, Ninni V. Are oil markets chaotic? A non-linear dynamic analysis. Energy Economics 2000; 22; 549–568.

Peltzman S. Prices rise faster than they fall. The Journal of Political Economy 2000; 108; 466–502.

Peroni E, McNown R. Non-informative and informative tests of efficiency in three energy futures markets. Journal of Futures Markets 1998; 18; 939–964.

Pindyck R. S. The long-run evolution of energy prices. The Energy Journal 1999; 20; 1-27.

Pindyck R.S., Rotemberg J.J. Energy shocks and the macroeconomy. In Alm AL, Weimer R.J. (Eds) Oil Shock: Policy Response and Implementation. Harper and Row Ballinger: Cambridge. 1984. pp. 97–120.

Plourde A, Watkins G, Crude oil prices between 1985 and 1994: How volatile in relation to other commodities? Resource and Energy Economics 1998; 20; 245–262.

Organization of the Petroleum Exporting Countries (OPEC). Brief History. http://www.opec.org/aboutus/history/history.htm

Organization of the Petroleum Exporting Countries (OPEC). Functions. http://www.opec.org/aboutus/functions/functions.htm

Quan J. Two-step testing procedure for price discovery role of futures prices. The Journal of Futures Markets 1992; 12; 139–149.

Radchenko S. Lags in the response of gasoline prices to changes in crude oil prices: the role of short-term and long-term shocks. Energy Economics 2005; 27; 573–502.

Rautava J. The role of oil prices and the real exchange rate in Russia's economy - a cointegration approach. Journal of Comparative Economics 2004; 32; 315-327.

Regnier E. Oil and energy price volatility. Energy Economics 2007; 29; 405-427.

Reilly B, Witt R. Petrol Price Asymmetries Revisited. Energy Economics 1998; 20; 297–308.

Sadorsky P. Oil price shocks and stock market activity. Energy Economics 1999; 21; 449−469.

Sadorsky P. The empirical relationship between energy futures prices and exchange rates. Energy Economics 2000; 22; 253–266.

Sadorsky P. Risk factors in stock returns of Canadian oil and gas companies. Energy Economics 2001; 23; 17–28.

Sadorsky P. The macroeconomic determinants of technology stock price volatility. Review of Financial Economics 2003; 12; 191–205.

Sadorsky P. Stochastic volatility forecasting and risk management. Applied Financial Economics 2005; 15; 121–135.

Sadorsky P. Modeling and forecasting petroleum futures volatility. Energy Economics 2006; 28; 467–488.

Schorderet Y. Asymmetric cointegration. Department of Econometrics. University of Geneva; Working Paper 2004.

Schwartz T.V., Szakmary A.C. Price discovery in petroleum markets: Arbitrage, cointegration and the time interval of analysis. The Journal of Futures Markets 1994; 14; 147–167.

Sequeira J.M., McAleer M. A market-augmented model for SIMEX Brent crude oil futures contracts. Applied Financial Economics 2000; 10; 543–552.

Serletis A. A cointegration analysis of petroleum futures prices. Energy Economics 1994; 16; 93–97.

Serletis A, Andreadis I. Random fractal structures in North American energy markets. Energy Economics 2004; 26; 389–399.

Serletis A, Banack D. Market efficiency and cointegration: an application to petroleum markets. The Review of Futures Markets 1990; 9; 372–385.

Shambora W.E., Rossitera R. Are there exploitable inefficiencies in the futures market for oil? Energy Economics 2007; 29; 18-27.

Siliverstovs B, Neuman A, L'Hégaret G, von Hirschhausen C. International Market Integration for Natural Gas? A Cointegration Analysis of Prices in Europe, North America and Japan. Energy Economics 2005; 27; 603–615.

Silvapulle P, Moosa I. The relationship between spot and futures prices: evidence from the crude oil market. Journal of Futures Markets 1999; 19; 175–193.

Suriya K. Forecasting Crude Oil Price Using Neural Networks. CMU Journal 2006; 5; 377-386.

Tabak B.M., Cajueiro D.O. Are the crude oil markets becoming weakly efficient over time? A test for time-varying long-range dependence in prices and volatility. Energy Economics 2007; 29; 28-36.

Weiner R.J. Is the world oil market ‘One Great Pool’? The Energy Journal 1991; 12; 95–107.

Wang S.Y., Yu L, Lai K.K. A novel hybrid AI system framework for crude oil price forecasting. Lecture Notes in Computer Science 2004; 3327; 233–242.

Wang S.Y., Yu L, Lai K.K. Crude oil price forecasting with Tei@I methodology, Journal of Systems Science and Complexity 2005; 18; 145–166.

Xie W, Yu L, Xu S.Y., Wang S.Y. A new method for crude oil price forecasting based on support vector machines. Lecture Notes in Computer Science 2006; 3994; 441–451.

Ye M, Zyren J. Shore J. Forecasting crude oil spot price using OECD petroleum inventory levels. International Advances in Economic Research 2002; 8; 324–334.

Ye M, Zyren J. Shore J. A monthly crude oil spot price forecasting model using relative inventories. International Journal of Forecasting 2005; 21; 491–501.

Ye M, Zyren J. Shore J. Forecasting short-run crude oil price using high and low-inventory variables. Energy Policy 2006a; 34; 2736–2743.

Ye M, Zyren J. Shore J. Short-run crude oil price and surplus production capacity. International Advances in Economic Research 2006b; 12; 390–394.

Yousefi S, Weinreich I. Reinarz D. Wavelet-based prediction of oil prices. Chaos, Solitons and Fractals 2005; 25; 265–275.

Yu L, Lai K.K., Wang S, He K. Oil price forecasting with an EMD-based multiscale neural network learning paradigm. Lecture Notes in Computer Science 2007; 4489; 925–932.

Yu L, Wang S, Lai K.K. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Economics 2008; 30; 2623–2635.

Zalduendo J. Determinants of Venezuela's equilibrium real exchange rate. IMF Working Paper 2004; 0674.

Zeng T, Swanson N.R. Predictive evaluation of econometric forecasting models in commodity futures markets. Studies in Nonlinear Dynamics and Econometrics 1998; 2; 1037-1037.

Chapter 3:

A Multidimensional Framework for Performance Evaluation of Forecasting Models: Context-Dependent DEA

A Multidimensional Framework for Performance Evaluation of Forecasting Models: Context-Dependent DEA

Abstract

The purpose of this paper is to propose a mathematical programming based approach, commonly referred to as Data Envelopment Analysis (DEA), as a multidimensional framework for relative performance evaluation of competing forecasting models. We first survey and classify the literature on performance criteria and their measures commonly used in evaluating and selecting forecasting models or methods and discuss the limitations of the current practices. Then, we use this classification as a basis for discussing the DEA framework and its implementation issues. Finally, for illustration purposes, we have chosen forecasting of crude oil prices as an application area.

Key words: Forecasting, performance criteria, performance measures, data envelopment analysis, crude oil

3.1. Introduction

The literature on forecasting spreads across a wide range of application areas. Most studies concerned with forecasting the level, the volatility, or both of time series tend to use one or several performance criteria and, for each criterion (e.g., accuracy), one or several metrics to assist in assessing the performance of competing models. Although several performance criteria and measures are used in most papers, the assessment exercise of competing models is generally restricted to their ranking by measure; thus, the current methodology is unidimensional in nature. Consequently, one may obtain different rankings of models for different measures leading to inconsistent and often confusing results both within and across studies. In this paper, we intend to contribute, from a methodological perspective, to the field of forecasting by proposing a multidimensional framework for relative performance evaluation of competing forecasting models. This multidimensional framework is known in the Operations Research and Management Science community as Data Envelopment Analysis (DEA) and is mathematical programming based. As opposed to other performance evaluation frameworks, DEA allows one to identify the weaknesses of each model, as compared to the best one(s), and suggests ways to improve their overall performance. DEA is a generic framework and as such its implementation for a specific relative performance evaluation exercise requires a number of decisions to be made such as the choice of the units to be assessed, the choice of the relevant inputs and outputs to be used, and the choice of the appropriate model(s). In order to present and discuss how one might adapt this framework to measure and evaluate the relative performance of competing forecasting models, we first survey and classify the literature on performance criteria and discuss how continuous and discrete metrics could be designed to measure these criteria. In sum, such a classification will assist with the operationalisation of DEA. Finally, we illustrate the use of DEA in evaluating and selecting models to forecast crude oil prices.

The remainder of this paper is organized as follows. In section 3.2, we survey and classify the performance criteria most used in forecasting and the metrics commonly

used to measure them. In addition, we discuss how continuous and discrete metrics could be designed to measure these criteria. In section 3.3, we present DEA concepts and discuss how one might adapt a context-dependent DEA (CDEA) framework to evaluate the relative performance of competing forecasting models. In section 3.4, we illustrate the use of CDEA in evaluating and selecting models to forecast crude oil prices. Finally, section 3.5 concludes the paper.

3.2. Performance Criteria and Measures in Forecasting

Forecasting researchers in various academic disciplines as well as practitioners in both private and public organizations are commonly faced with the problem of evaluating and selecting forecasting models or methods; such a process requires the specification of criteria and their measures as well as statistical tests based on which decisions will be made. The remainder of this section is divided into two sub-sections. In the first sub-section, we propose a classification of those performance criteria commonly used in forecasting research and provide some definitions/terminology that reflect our own views as well as what we believe to reflect the current needs of both academics and professionals. Then, in the second sub-section, we provide a list of metrics to measure the criteria presented in the previous sub-section along with a discussion of their limitations. In addition, we discuss how continuous and discrete metrics could be designed to measure these criteria. We also discuss how the conclusions reached by some statistical tests could be used for measurement purposes.

3.2.1. Performance Criteria in Forecasting

Based on our survey of the literature, we identified several criteria used by both academics and practitioners in evaluating and selecting forecasting models or methods for actual use or implementation (see, for example, Dalrymple, 1975; 1987; Carbone and Armstrong, 1982; Mentzer and Cox, 1984a, 1984b; Mahmoud, 1984; Armstrong, 1985; Mahmoud et al., 1986; Sanders and Manrodt, 1994; Mentzer and Kahn, 1995; Yokum and Armstrong, 1995; Armstrong, 2001a; 2001b; Armstrong et al., 2001; McCarthy et al., 2006). In this paper, we classify these criteria into six categories; namely, reliability,

costs, benefits, complexity, universality, and ability to incorporate managerial judgmental (see Figure 1). Obviously, these criteria might take on different meanings over time. In this paper, we divide the *reliability* criterion into five sub-criteria; namely, theoretical relevance, validity, accuracy, informational efficiency, and degree of uncertainty of output of a model or forecast.

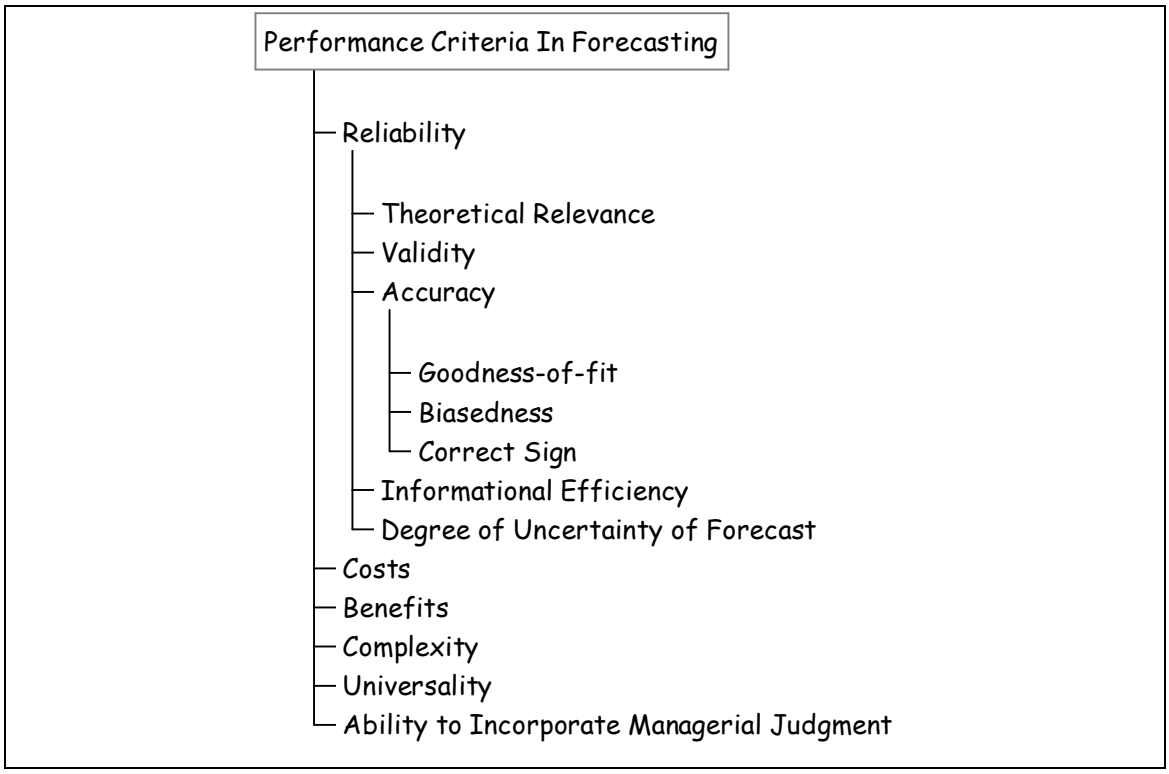


Figure 1: Classification of Performance Criteria in Forecasting

The first sub-criterion reliability is made of is *theoretical relevance* and refers to the degree of suitability of a model or method for a given data set; that is, the ability of the model or method to predict patterns and turning points, if any, to adapt to new conditions, sometimes referred to as robustness or flexibility, such as a structural change, if any, and to account for interventions, if any, as well as its suitability for a specific forecasting horizon; i.e., short-, medium-, and long horizons – the relevance of the forecasting horizon lies in the fact that, by design, some forecasting methods are more suitable for a specific time horizon than others. In sum, the theoretical relevance of a method or model refers to its ability, from a design perspective, to take into account all

elements of information within the data set under consideration. The second sub-criterion is *validity*, which refers to whether the assumptions underlying a model hold or not. From a methodological perspective, invalid models should be discarded from further consideration; however, in practice things aren't either black or white. For example, within a regression analysis framework, one needs to test whether the residuals are normally distributed or not using several normality tests, which may lead to different conclusions; in this case, one may conclude that the normal distribution is a reasonable approximation and consider the model as valid – details about how to measure the extent to which a model is valid will be provided in the next sub-section. The third sub-criterion is *accuracy* and refers to the ability of a model or method to reproduce the past. Accuracy is a multidimensional construct that has three main facets; namely, the goodness-of-fit dimension, the bias dimension, and the correct sign dimension, where the *goodness-of-fit* dimension or sub-criterion refers to how close the forecasts are from the actual values, the *biasedness* sub-criterion refers to whether the model or method tends to systematically over-estimate or under-estimate the forecasts, and the *correct sign* sub-criterion refers to the ability of a model to forecast the correct sign; that is, to produce forecasts that are consistent with actuals in that forecasts reveal increase (resp. decrease) in value when actuals increase (resp. decrease) in value – this criterion is particularly important in investment environments. The fourth sub-criterion is *informational efficiency* and refers to the ability of a model to capture all elements of information within the data. The fifth and last sub-criterion of reliability is the *degree of uncertainty of the output of a model*, or equivalently, the degree of uncertainty of a forecast; that is, the likelihood that the forecast and the actual values will be close to each other.

The second criterion commonly used in assessing and selecting forecasting models or methods for implementation is *cost*. The cost of a forecasting model or method refers to the extent to which it is relatively cheap to acquire and use. It includes several categories of costs; for example, the development/purchase and maintenance costs, the data purchase/collection, storage, and pre-processing costs, the costs of training analysts to use a model/method effectively, the costs related to the time required to obtain a forecast,

and the costs of repeated applications of the method. These cost elements do vary in importance and magnitude depending on whether the forecasting method is quantitative or qualitative. In fact, as far as quantitative methods are concerned, nowadays most forecasting software are purchased rather than in-house developed, because the purchasing price has become negligible as compared to the in-house development cost. In addition, forecasting software typically provide a relatively wide range of models along with multiple facilities to handle and analyze data; nevertheless, one might still need to write code for specific implementations; e.g., a rolling horizon implementation, or analyses not provided by the software as ready-to-use options. Furthermore, the continual use of a quantitative forecasting model typically requires some maintenance work that comes at a relatively low cost; e.g., updating the parameters of the model. Given the current state of technology, the cost of storing data has become negligible for the average user; thus, data collection or purchase costs have become the main expenses. However, the relative availability of a wide range of data from both public and private sources along with the competitiveness of these data providers have driven costs to the customer advantage. Obviously, different models or methods require different amounts of historical data, which makes some models/methods cheaper to use than others. Another cost related to data could be incurred if any preliminary processing of data is required before it can be used by a method or model. As reflected by the widely rehearsed adage “time is money”, in practice the cost related to the time required to obtain a forecast may be an important factor to take into account depending on whether the model or method is qualitative or quantitative; in fact, most often the time required to obtain a forecast using a quantitative model or method is limited to the running time of the method/speed on a computer machine, which in general is negligible, whereas the time required to obtain a forecast using a qualitative method may be substantial depending on the nature of the process adopted. Finally, the repeated applications or uses of a quantitative model do not in general require much additional cost. On the other hand, qualitative forecasting methods tend to be more expensive as they are more human resources intensive and require rather expensive inputs; e.g., opinions provided by

experts. Furthermore, the repeated applications or uses of a qualitative model do not in general differ much in cost from the first use.

The third criterion refers to the expected *benefits* that would result from the use of a model or method in generating forecasts such as cost savings and improved decisions. The fourth criterion refers to the *complexity* of a model or method – also referred to in the literature as the ease of use or the ease of implementation of the model/method. In this paper, the complexity of a forecasting model or method refers to the extent to which it is easy to understand by users/managers and to interpret its results, or equivalently, the level of conceptual and technical knowledge/expertise required for an effective use of the model/method. The fifth criterion refers to the *universality* of a model or method; that is, the extent to which the model/method is widely used in practice, or equivalently, the familiarity of the audience with it. Note however that, for quantitative models, this factor may largely depend on the availability of a model or method within popular software packages. The sixth and last criterion refers to the *ability of a model or method to incorporate managerial judgment*; that is, the integration of subjective information to produce a forecast.

Our literature survey reveals that most academic studies in the area of forecasting mainly use the first criterion; namely, reliability or one of its sub-criteria, to assess the performance of a model or method. This reality could be explained by the fact that, with the exception of complexity and universality, the remaining criteria could only be effectively measured after the actual implementation of the forecasts, which tends to limit their use in the academic context. Regardless of the context of use and the relative weights assigned to each of these criteria, one needs to measure them. In the next sub-section, we survey and classify the different metrics that could be used to measure these criteria and suggest different metrics for those criteria for which not much seems to be available in the published literature. We also present some statistical tests commonly used to test a single forecasting ability of a model and discuss how their outcome could be used for measurement purposes.

3.2.2. Performance Measures in Forecasting

In the previous section, we proposed a classification of performance criteria that one could use within a multidimensional framework to assess the relative performance of several forecasting models (see Figure 1). In this section, we present and discuss how one might measure these criteria. In addition, a classification of those measures commonly used by both academics and professionals are provided for "popular" criteria.

The first criterion of our classification is referred to as the *reliability* criterion. It is decomposed into several sub-criteria; namely, *theoretical relevance, validity, accuracy, informational efficiency, and degree of uncertainty of a forecast*.

As far as *theoretical relevance* or relevance by design is concerned, several metrics could be used; to be more specific, measuring the theoretical relevance of a model depends on the level of detailed information one wants to take account of. For example, one might only be interested in knowing whether a model is relevant by design for the specific data set under consideration or not; i.e., the design of the model provides it or not with the ability to take account of, or handle better, the different features of the data, in which case a score of zero or one could be used as a measure. On the other hand, if more information is to be taken into account, one may use a categorical variable to represent the extent to which a model is relevant by design and such a variable may take, for example, three values, say 1, 2, and 3, to reflect low, medium, and high degrees of theoretical relevance, respectively. Finally, one might use the proportion of data features that a model is by design capable of handling as a continuous measure of theoretical relevance. For example, after one has performed preliminary/exploratory analysis of the data set under consideration and identified the main features of the data, the ratio of the number of such features that a model can take account of by design to the total number of features of the data could be used to measure the theoretical relevance of a model. Features of the data could include trend, seasonality, cycles, structural change, planned events, etc.

The second sub-criterion of reliability is *validity*. Recall that, from a methodological perspective, invalid models should be discarded; however, in practice such decisions are

not straightforward. For example, within a regression analysis framework, one needs to test whether the residuals are white noise, which involves testing several hypotheses; e.g., residuals are statistically independent, normally distributed with mean zero and a constant standard deviation. But, for some data sets, depending on which test is used to test each of these hypotheses, one might end up with a different conclusion. In order to avoid discarding a model based on a single statistical test, one might use several tests, whenever available, and reject the specific hypothesis under consideration only if all tests reject it; otherwise, the hypothesis should not be automatically rejected because each statistical test has its own limitations. To be more specific, one should consider a measure of validity that reflects the proportion of tests that do not find enough statistical evidence to reject the null hypothesis under consideration – the basic idea behind this approach is that, in practice, one often finds out that some tests reject normality whereas others don't and analysts often conclude that the normal distribution is a reasonable approximation and consider the model as valid. Whenever such conditions prevail, we suggest the use of an average of these proportions as a measure of validity.

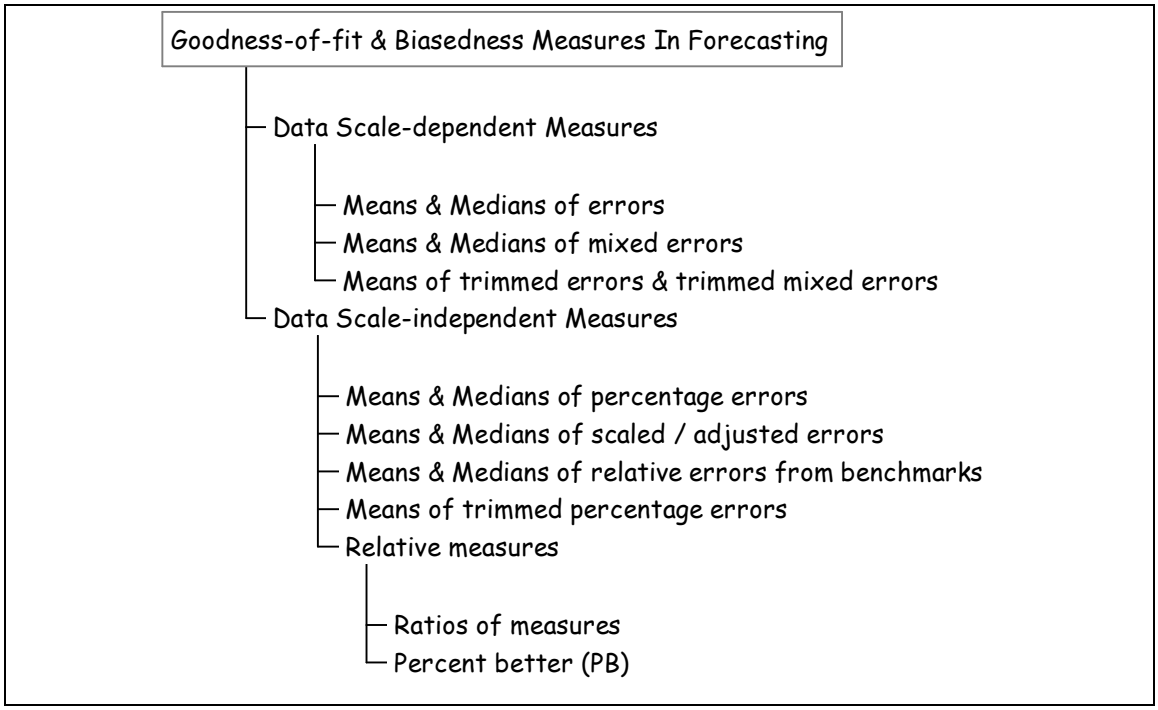


Figure 2: Classification of Goodness-of-fit and Biasedness Measures

The third sub-criterion of reliability is *accuracy*. In this paper, the accuracy criterion is decomposed into three sub-criteria; namely, goodness-of-fit, biasedness, and correct sign. Accuracy is without doubt the most popular criterion used by both academics and professionals in almost all studies and hence a substantial effort has been dedicated to design and study metrics for accuracy. As far as the first two sub-criteria are concerned, their metrics may be divided into two broad categories depending on whether they are scale-dependent or independent (see Figure 2).

The most commonly used *scale-dependent measures of goodness-of-fit and biasedness* are the means of errors $e_t = Y_t - \hat{Y}_t$, absolute errors $|e_t|$, and squared errors e_t^2, where the means of errors are used to measure bias and the means of absolute and squared errors are used to measure goodness-of-fit. Several mean metrics (e.g., arithmetic mean, quadratic mean, geometric mean, harmonic mean) could be used to provide the correct notion of "average" depending on the purpose of use, where the arithmetic mean seems to be the most popular. Note however that using statistics; such as a mean, based on e_ts, $|e_t|$s, and e_t^2s assume that errors of the same magnitude are assigned the same weight regardless of their signs, which may not be the sensible thing to do in some applications; to overcome this problem and be in a position to express users' preferences or to penalize under- or over-predictions, whatever needs to be penalized, one may use mixed errors; that is, the errors resulting from a transformation that penalizes either under- or over-predictions. For example, if one wants to penalize under-predictions (i.e., positive errors), he or she might use the square root of an error if the error is less than or equal to one, a squared error if the error is greater than one, and either the absolute or the squared error for negative errors depending on whether one is interested in computing statistics based on mixed absolute errors or mixed squared errors:

$$\text{mixed } |e_t| = \begin{cases} |e_t| & \text{if } e_t \prec 0 \\ \sqrt{e_t} & \text{if } 0 \le e_t \le 1 \\ e_t^2 & \text{if } e_t \succ 1 \end{cases}, \quad \text{mixed } e_t^2 = \begin{cases} e_t^2 & \text{if } e_t \le 0 \\ \sqrt{e_t} & \text{if } 0 \le e_t \le 1 \\ e_t^2 & \text{if } e_t \succ 1 \end{cases}$$

Note that means, as measures of central tendency, are affected by the presence of outliers, which would distort numerical figures and lead to inappropriate decisions.

Therefore, whenever the errors' distribution is skewed, one may use other measures of central tendency such as medians, which are not affected by outliers. However, using medians to discriminate between competing models could only provide the researcher with information as to which model is "better" but not by how much, which from a practical perspective could be misleading. An alternative to the use of medians would be to use trimmed errors, also referred to as winsorized errors, where one replaces extreme values with certain limits to temper the impact of outliers. Obviously, the use of trimmed errors introduces some degree of arbitrariness as the amount of trimming must be specified by the researcher, but retains some information about the high and low values of errors as compared to medians, whether one uses e_ts, $|e_t|$s, e_t^2s, or mixed e_ts.

On the other hand, the most commonly used *scale-independent measures of goodness-of-fit and biasedness* are based on percentage errors $PE_t = \frac{e_t}{Y_t} \times 100$, absolute percentage errors $APE_t = \left| \frac{e_t}{Y_t} \times 100 \right|$, and squared percentage errors $SPE_t = \left(\frac{e_t}{Y_t} \times 100 \right)^2$. Note however that PE_ts and, consequently, APE_ts and SPE_ts, penalize positive errors – or equivalently, under-predictions – more than negative ones; i.e., over-predictions. To overcome this problem, one may adjust PE_ts as follows:

$$\text{Adjusted } PE_t = \frac{e_t}{\left(Y_t + \hat{Y}_t \right) / 2} \times 100 .$$

In the literature, measures based on $\text{Adjusted } PE_t$s are sometimes referred to as symmetric measures, because they assign the same penalty to errors of the same magnitude regardless of their sign. Referring to these measures as symmetric may however be misleading as argued by some researchers (e.g., Goodwin and Lowton, 1999; Koehler, 2000), especially if Y_t and $\hat{Y}_t$ are of opposite signs. Expressing errors in terms of percentages is one way of making forecasting models comparable across different time series or periods of the same series. A second way of achieving comparability is to scale errors (i.e., e_ts, $|e_t|$s, e_t^2s, or mixed e_ts) or adjust them for, say trend or volatility, either in a static or in a dynamic fashion. In fact, static scaling

could be achieved by dividing errors by the same time-independent quantity such as a specific measure of goodness-of-fit of a benchmark (e.g., MAE of random walk proposed by Hyndman and Koehler, 2006) or a measure of trend or volatility of the series over the whole period under consideration. On the other hand, dynamic scaling could be achieved by dividing errors by time-varying quantities such as a specific measure of goodness-of-fit of a benchmark over different sub-periods or by dividing errors by a measure of trend or volatility that varies over time; e.g.,

$$\text{Trend Scaled } SE_{t_1,t_2} = \frac{e_t^2}{\overline{T}_{t_1,t_2}} \text{ and } \text{Volatility Adjusted } SE_{t_1,t_2} = \frac{e_t^2}{S^2_{t_1,t_2}}$$

$$\text{where } \overline{T}_{t_1,t_2} = \frac{1}{t_2 - t_1}\sum_{k=t_1+1}^{t_2}\left|Y_k - Y_{k-1}\right|,\ S^2_{t_1,t_2} = \frac{1}{t_2 - t_1}\sum_{k=t_1}^{t_2}(Y_k - \overline{Y}_{t_1,t_2})^2 \text{, and}$$

$$\overline{Y}_{t_1,t_2} = \frac{1}{t_2 - t_1 + 1}\sum_{k=t_1}^{t_2} Y_k \text{ for } t_2 \succ t_1 \geq 1.$$

A third way of making different forecasting models comparable across different series or periods is by working with relative errors from a benchmark $RE_t = e_t / e_t^{benchmark}$, relative absolute errors from a benchmark $RAE_t = |e_t| / \left|e_t^{benchmark}\right|$, or relative squared errors from a benchmark $RSE_t = e_t^2 / \left(e_t^{benchmark}\right)^2$, where $e_t^{benchmark}$ denotes the forecasting error resulting from the use of a specific benchmark model or method. At this stage, we would like to attract the reader's attention to the fast that most studies use the means of the above-mentioned errors as measures of accuracy. If the distributions of these errors are skewed, means would not be appropriate unless errors are normalized (e.g., Box and Cox, 1964; Swanson et al., 2000) or trimmed. Another alternative would be to use medians of errors instead, but one should be aware of their limitations. Note that measures based on PE_ts, APE_ts or SPE_ts have limitations in that these errors approach infinity as one or several observations Y_t approach zero, and are undefined at $Y_t = 0$. Therefore, these measures should be avoided whenever the series contains observations close to zero. Note also that measures based on RE_ts, RAE_ts and RSE_ts suffer the same problems as those based on PE_ts, APE_ts and SPE_ts, because errors from benchmarks

$e_t^{benchmark}$ may be zero or close to zero. The fourth and last way of making different forecasting models comparable across different series or periods is by working with what is referred to as relative measures. These measures are divided into ratio measures and percent better. A ratio measure is defined as the ratio of a measure to another one, usually referred to as a benchmark measure. Ratio measures could be mean or median based; e.g., $ME/ME^{benchmark}$, $MdAE/MdAE^{benchmark}$, could use any type of errors; e.g., PE_t, APE_t, SPE_t, RE_t, RAE_t, RSE_t, their numerators and denominators could use the same measures or different ones; e.g., $MAE/MAE^{benchmark}$, $ME/MAE^{benchmark}$, and the measures used in numerator and denominator could be based on errors resulting from different implementations of the methods under consideration; e.g., in-sample implementation ($in-sample\ ME/in-sample\ MAE^{benchmark}$), out-of-sample implementations ($out-of-sample\ ME/in-sample\ MAE^{benchmark}$). On the other hand, percent better measures refer to the number of times the forecast error produced by a method is lower than the one produced by another method, expressed in percentage, where the forecast error may be of any type; e.g., PE_t , APE_t , SPE_t . The metrics mentioned above, whether to measure goodness-of-fit or biasedness, are continuous measures. Note however that one might as well use categorical variables to measure biasedness. In fact, several statistical tests; e.g., standard sign test and Wilcoxon signed-ranked test, could be used to test for biasedness, suggesting that 0-1 variables could be used to reflect whether a model produces biased forecasts or not. In addition, whenever more than one test is used, one could use the proportion of tests in favor of unbiasedness as a measure of biasedness.

As to the third sub-criterion of accuracy; namely, the *ability of a model or method to forecast the correct sign*, it may also be measured using either a continuous or a discrete scale. For example, on a continuous scale, one might measure the ability of a model to forecast the correct sign using the proportion of correct sign predictions computed as $\sum_{t=1}^{n} z_t \Big/ n$, where n denotes the number of observations and z_t is a binary variable that

takes on 1 if $Y_t \cdot \hat{Y}_t \succ 0$ and 0 otherwise – some studies use $1 - \sum_{t=1}^{n} z_t \Big/ n$ instead and refer to it as the confusion rate criterion. On the other hand, statistical tests such as the chi-square test of independence based on a contingency table, commonly referred to as the confusion matrix, could serve as a basis for devising a discrete measure in the form of a categorical variable.

The fourth sub-criterion reliability is made of is referred to as *informational efficiency*, which may be measured using either a continuous or a discrete scale. In fact, the proportion of statistical tests – typically regression based t-tests (e.g., Mincer and Zarnowitz, 1969) – that do not find enough statistical evidence to reject the null hypothesis of informational efficiency could be used as a continuous measure, whereas a categorical variable, with as many categories as statistical tests one wishes to use, could be used as a discrete measure.

The fifth and last sub-criterion of reliability is referred to as the *degree of uncertainty of a forecast*, which may be measured on a continuous scale by a metric such as the length of the prediction interval. Prediction intervals are usually determined under the assumption that residuals are normally distributed. When such an assumption does not hold, one may express the management team past experience with forecasts generated by a specific model by using a categorical variable with, for example, three levels of uncertainty, say low, medium, and high.

The second criterion that could be used to discriminate between competing forecasting models or methods is *cost*. For measurement purposes, cost may be measured on a continuous scale using an amount of currency, or on a discrete scale using a categorical variable with, for example, three categories to reflect, say, low, medium and high cost. The third criterion refers to the *benefits* that could result from the use of a specific forecasting model or method and could be measured either on a continuous scale using, for example, the amount of cost reduction due to improved decisions, or on a discrete scale using, for example, a categorical variable to represent different service levels where the score is the service level resulting from improved decisions due to forecasts.

The fourth criterion refers to the *complexity* of the forecasting model or method. Although such a criterion might be measured on a continuous scale by a metric such as a weighted average of scores on different answers to a set of questions related to complexity – as perceived by a typical group of users, in practice complexity is usually measured on a discrete scale using a categorical variable to express the users opinions as to the extent to which they perceive a model or method as complex. The fifth criterion referred to as *universality* could also be measured on either a continuous or a discrete scale; in fact, the information/data collected from a survey could serve both purposes. Finally, the sixth and last criterion refers to the *ability of the model or method to incorporate managerial judgment* and is best measured using a categorical variable with several categories to reflect the extent to which managerial judgment could be taken into account, for example, through the choice of the values of the model or method parameters.

In summary, the literature provides both academics and practitioners with a rich pool of criteria and their measures to choose from; however, to the best of our knowledge no multidimensional framework for assessing the relative performance of competing forecasting models or methods has been proposed. In the next section, we intend to propose such a framework.

3.3. A DEA Framework for Model Evaluation and Selection in Forecasting

Performance measurement and evaluation of most systems – often referred to as assessment units or decision making units (DMUs), may be dealt with in many different ways, where a system is defined by its inputs, its processes, and its outputs. In most application areas, performance indicators – also referred to as partial productivity measures, are among the most commonly used tools for measuring performance, although these single output to single input ratios are only suitable when a single performance indicator pertains. In an attempt to overcome such a limitation, total factor productivity measures have been proposed, which may simply be defined as ratios of a combination of several outputs' quantities to a combination of several inputs' quantities

where the weights assigned to different inputs and outputs are fixed across all the units to be assessed. One of the main critics towards these measures is that the weights are most often chosen in a subjective fashion and are fixed across all assessment units, which often leads to unfairness. Parametric methods such as classical regression and stochastic frontier methods (Aigner et al., 1977) have been proposed to address some of these critics; however, these methods require explicit knowledge of the production function, can only handle situations with one input and several outputs or several inputs and one output, the underlying assumptions are often too restrictive in practice, and the benchmark used within these analyses frameworks reflects the average behavior of DMUs rather than the best or most efficient behavior. Within the same line of thought, data envelopment analysis (DEA) may be viewed as another approach to determine a multidimensional performance indicator where the weights are determined in an optimal and fair fashion using mathematical programming. Recall that DEA aims at assessing the relative performance of a given set of, say n, DMUs in transforming the same set of, say m, inputs (i.e., relevant resources) into the same set of, say s, outputs (i.e., outcomes of a production process) under the assumption that DMUs have control over the process of converting inputs into outputs. Recall also that the basic optimization problem addressed by DEA consists of maximizing the performance of a specific DMU, say DMU_k, as measured by the ratio of a weighted combination of the quantities of outputs produced by DMU_k to a weighted combination of the quantities of inputs used by DMU_k under the constraints that, for each DMU including DMU_k, the ratio of a weighted combination of its outputs' quantities to a weighted combination of its inputs' quantities is less than or equal to one and all weights are non-negative – this problem is commonly referred to as the *input-oriented problem* as compared to an equivalent problem referred to as the *output-oriented problem* where the objective is to minimize the ratio of a weighted combination of the quantities of inputs used by DMU_k to a weighted combination of the quantities of outputs produced by DMU_k. The translation of the narrative forms of these optimization problems into algebraic forms result in fractional mathematical programs – see Table 1, where $x_{i,j}$ denotes the amount of input i used by

DMU_j, $y_{r,j}$ denotes the amount of output r produced by DMU_j, v_i (respectively u_r) denotes the weight of input i (respectively, output r).

Input-Oriented	Output-Oriented
$\text{Maximize } e_k^{input} = \dfrac{\sum_{r=1}^{s} u_r y_{r,k}}{\sum_{i=1}^{m} v_i x_{i,k}}$	$\text{Minimize } e_k^{output} = \dfrac{\sum_{i=1}^{m} v_i x_{i,k}}{\sum_{r=1}^{s} u_r y_{r,k}}$
$\text{s.t.}: \quad \dfrac{\sum_{r=1}^{s} u_r y_{r,j}}{\sum_{i=1}^{m} v_i x_{i,j}} \le 1; j = 1,\ldots,n$	$\text{s.t.}: \quad \dfrac{\sum_{r=1}^{s} u_r y_{r,j}}{\sum_{i=1}^{m} v_i x_{i,j}} \le 1; j = 1,\ldots,n$
$v_i \ge 0; i = 1,\ldots,m$	$v_i \ge 0; i = 1,\ldots,m$
$u_r \ge 0; r = 1,\ldots,s$	$u_r \ge 0; r = 1,\ldots,s$

Table 1: Basic DEA Fractional Programs

After a change of variables, these fractional programs are transformed into linear programs commonly referred to as *multiplier or dual problems* – see Table 2.

Input-Oriented	Output-Oriented
$\text{Maximize } e_k^{input} = \sum_{r=1}^{s} u_r y_{r,k}$	$\text{Minimize } e_k^{output} = \sum_{i=1}^{m} v_i x_{i,k}$
$\text{s.t.}: \quad \sum_{i=1}^{m} v_i x_{i,k} = 1$	$\text{s.t.}: \quad \sum_{r=1}^{s} u_r y_{r,k} = 1$
$\sum_{r=1}^{s} u_r y_{r,j} \le \sum_{i=1}^{m} v_i x_{i,j}; j = 1,\ldots,n$	$\sum_{r=1}^{s} u_r y_{r,j} \le \sum_{i=1}^{m} v_i x_{i,j}; j = 1,\ldots,n$
$v_i \ge 0; i = 1,\ldots,m$	$v_i \ge 0; i = 1,\ldots,m$
$u_r \ge 0; r = 1,\ldots,s$	$u_r \ge 0; r = 1,\ldots,s$

Table 2: Basic DEA Multiplier Models

In sum, input-oriented (respectively, output-oriented) analysis maximizes a weighted sum of output quantities (respectively, minimizes a weighted sum of input quantities) under the same set of constraints mentioned above along with an additional constraint whereby a weighted sum of input quantities (respectively, a weighted sum of output quantities) is fixed to 1. Note that the duals of multiplier problems are commonly

referred to as *envelopment or primal problems* – see Table 3, where θ_k and λ_j are the dual variables and s_i^- and s_i^+ denote slack variables.

Input-Oriented	**Output-Oriented**
Minimize $\theta_k - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)$	Maximize $\phi_k + \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)$
s.t.: $\sum_{j=1}^{n} \lambda_j x_{i,j} + s_i^- = \theta_k \cdot x_{i,k}; \forall i$	s.t.: $\sum_{j=1}^{n} \lambda_j x_{i,j} + s_i^- = x_{i,k}; \forall i$
$\sum_{j=1}^{n} \lambda_j y_{r,j} - s_r^+ = y_{r,k}; \forall r$	$\sum_{j=1}^{n} \lambda_j y_{r,j} - s_r^+ = \phi_k \cdot y_{r,k}; \forall r$
$\lambda_j \geq 0; \forall j; s_i^- \geq 0; \forall i; s_r^+ \geq 0; \forall r$	$\lambda_j \geq 0; \forall j; s_i^- \geq 0; \forall i; s_r^+ \geq 0; \forall r$
θ_k unrestricted	ϕ_k unrestricted

Table 3: Basic DEA Envelopment Models

Note that as compared to regression analysis and stochastic frontier analysis, DEA can at the same time handle many inputs and many outputs, does not require any specific knowledge of the production function, and its benchmark reflects the best behavior amongst DMUs. In addition, DEA identifies the best practice production frontier or empirical standard of excellence – commonly referred to as the efficient frontier, along with a reference set or peer group for each DMU as well as different types of targets to aim for. Furthermore, DEA differentiates between different types of efficiencies and can identify different types of inefficiencies. DEA has known wide acceptance in the Management Science/Operations Research community and has been used as a performance evaluation framework in many application areas such as *health care* (Kleinsorge and Karney, 1992; Chang, 1998; Birman et al., 2003; Chang et al., 2004; Chilingerian and Sherman, 2004; Retzlaff-Roberts et al., 2004), *education* (Johnes, 2006), *automobile insurance* (Retzlaff-Roberts and Puelz, 1996); *justice* (Lewin et al., 1982; *regional planning and economic performance of cities* (Charnes et al., 1989), *economic performance of countries* (Ramanathan, 2006), *information systems* (Sowlati et al., 2005), *software development* (Banker and Kemerer, 1989), *Bank branches performance* (Sherman and Gold, 1985; Charnes et al., 1990; Cook et al., 2004), *Bank failure prediction* (Barr and Siems, 1997; Ravikumar and Ravi, 2008), *credit scoring*

(Cheng et al., 2007; Emel et al., 2003), *hedge fund performance appraisal* (Gregoriou et al., 2005), *highway maintenance efficiency* (Cook et al., 1990), *logistics systems* (Kleinsorge et al., 1989), *purchasing bids* (Zhu, 2004), *R&D* (Oral et al., 1990), *sports* (Anderson and Sharp, 1997), *Olympic rankings* (Wu et al., 2009), *textile industry performance* (Zhu, 1996a), *steel industry productivity* (Ray et al., 1998), *energy and environmental studies* (Zhou et al., 2008), etc. The reader is referred to Charnes et al. (1994), Seiford (1996) and Cooper et al. (2004) for further surveys on application areas. In this paper, we propose DEA as a multidimensional framework for relative performance evaluation of competing forecasting models. DEA is a generic framework and as such its implementation for a specific relative performance evaluation exercise requires a number of decisions to be made. In sum, we propose DEA as a multidimensional framework for relative performance evaluation of competing forecasting models and propose to adapt such a framework by making the following decisions. First, the units to be assessed or DMUs are forecasting models. Second, inputs and outputs are relevant performance measures. Actually, once the relevant performance criteria to be used for assessing a given set of forecasting models are chosen along with their measures – see next section, the remaining major decision to be made is concerned with choosing the appropriate optimization problem to address and the corresponding mathematical program to solve.

Before providing any guidelines as to how one could make these choices, we will first provide a list of the main DEA problems and models and then highlight the ones that could potentially be used for assessing the relative performance of competing forecasting models – for a recent survey of DEA models, the reader is referred to Cook and Seiford (2009). The main DEA models may be summarized as follows: (1) *radial DEA models*, which only allow for proportional changes in inputs and outputs (Charnes et al., 1978, 1979, 1981; Banker et al., 1984); (2) *non-radial DEA models*, which allow for non-proportional changes in inputs and outputs (Färe and Lovell, 1978; Charnes et al., 1985, Green et al., 1997; Cooper et al., 1999a; Tone, 2001); (3) *DEA models with restricted multipliers*; e.g., DEA models with absolute multiplier restrictions, which impose lower and/or upper bounds on input and/or output multipliers (Roll et al., 1991),

DEA models with assurance regions, which impose constraints on the relative magnitude of weights of some pairs of inputs or outputs (Thompson et al., 1986, 1990; Allen et al., 1997; Thanassoulis et al., 1998; Cook and Zhu, 2008), and DEA models with cone ratio restrictions, which impose a set of linear restrictions on multipliers that define a convex cone (Charnes et al., 1990; Thompson et al., 1995); (4) *DEA models with special variables*; e.g., DEA models with non-discretionary variables, which take account of the non-discretionary nature of some inputs and outputs (Banker and Morey, 1986a, 1986b; Ruggiero, 1996, 1998, 2007), DEA models with non-controllable variables, which take account of the uncontrollable nature of some inputs and outputs (Cooper et al., 2000, 2007; Cook and Seiford, 2009), DEA models for undesirable inputs and outputs, which allow one to reflect the "true" production process instead of treating undesirable outputs (resp. inputs) as inputs (resp. outputs) to minimize (resp. maximize) them (Scheel, 2001; Seiford and Zhu, 2002; Färe and Grosskopf, 2004; Hua and Bin, 2007), DEA models with categorical variables, which allow for "fair" comparison across relevant categories only (Banker and Morey, 1986b; Kamakura, 1988; Rousseau and Semple, 1993), DEA models with ordinal variables, which take account of qualitative inputs and outputs (Cook et al., 1993, 1996; Cooper et al. 1999a, 1999b; Zhu, 2003a; Cook and Zhu, 2006), and DEA models with flexible measures, which allow for a flexible role or status of performance measures as inputs or outputs (Cook et al., 2006; Cook and Zhu, 2007); (5) *weak disposability DEA models* – also referred to as *congestion models*, which are designed to reveal congestion; i.e., situations where decreases in one or more inputs generate or could be associated with increases in one or more outputs without worsening other input or output (Färe and Grosskopf, 1983; Färe et al.,1994; Brockett et al., 1998; Cooper et al., 1996, 2000); (6) *context-dependent DEA models*, which allow relative performance evaluation with respect to several contexts or alternatives that reflect local and global targets and several levels of best-practice frontiers to allow for progressive and feasible improvements and are used for ranking DMUs (Seiford and Zhu, 2003; Cook and Zhu, 2008); (7) *DEA benchmarking models*; e.g., variable-benchmark envelopment models which fix the efficient frontier but allow for each DMU to have a different benchmark, fixed-benchmark multiplier models used to

provide an upper efficiency bound under the assumption that all DMUs have the same benchmark, minimum efficiency multiplier models used to provide a lower efficiency bound under the assumption that all DMUs have the same benchmark, ideal-benchmark multiplier models used to provide an upper efficiency bound under the assumption that the DMU under evaluation is compared to an ideal benchmark, and ideal-benchmark minimum efficiency models used to provide a lower efficiency bound under the assumption that the DMU under evaluation is compared to an ideal benchmark (Zhu, 2003b; Cook et al., 2004); (8) *DEA super efficiency models*, which do not include the DMU under evaluation in the reference set and are used for ranking efficient DMUs , detecting influential DMUs , studying the sensitivity of an efficiency classification and determining efficiency stability regions (Andersen and Petersen, 1993; Thrall, 1996; Zhu 1996b; Dula and Hickman, 1997; Seiford and Zhu 1998a, 1998b; Mehrabian et al., 1999; Seiford and Zhu 1999a, 1999b; Zhu, 2001; Lovell and Rouse, 2003; Chen, 2004, 2005; Cook et al., 2008); (9) *DEA models with preference structures*, which take account of preferences on proportions by which levels of inputs and outputs could be changed (Zhu, 1996b; Zhu, 2003b, 2009); and (10) *DEA allocation models*, which take account of information on prices and revenues (Färe et al., 1985; Copper et al., 1999) – for a detailed presentation of these models, the reader is referred to Zhu (2003b, 2009), Cooper et al. (2004, 2007), Zhu and Cook (2007) and Cook and Zhu (2008).

Although all of the above mentioned models could be used to classify competing forecasting models into efficient and non-efficient models, only few of them could also provide a ranking; namely, context-dependent models, benchmarking models, and super efficiency models. Note however that the rankings resulting from most benchmarking models and super efficiency models are benchmark-dependent and are not robust enough for our application. Therefore, in this paper we propose context-dependent models to assess the relative performance of competing forecasting models and to rank them. Recall that context-dependent data envelopment analysis (CDEA) is concerned with partitioning a given set of DMUs into several levels of best-practice frontiers so that the first-level efficient frontier DMUs have a higher performance than the second-level

efficient frontier DMUs , the second-level efficient frontier DMUs have a higher performance than the third-level efficient frontier DMUs and so on, which enables performance benchmarking between different DMU groups – these efficient frontiers are referred to as evaluation contexts. Such a partitioning of the set of DMUs under consideration is obtained by the following iterative procedure – see Figure 3, where one has to choose the DEA model to use for determining efficient frontiers at different levels of performance such as the models of Table 3.

Initialization Step

Initialize the level counter ℓ to 1 and the set of DMUs to evaluate at level ℓ, J_ℓ, to $\{DMU_j, j = 1, \ldots, n\}$. Use the relevant DEA model to evaluate J_ℓ and set the ℓth-level best-practice frontier E_ℓ. Exclude the current 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 ℓth-level best-practice frontier E_ℓ set $J_{\ell+1} = J_\ell - E_\ell$, and increment ℓ by 1;

}

Figure 3: Partitioning Algorithm

Once DMUs have been partitioned into, say L, efficient frontiers with different levels of performance, one could obviously rank order them from best to worst starting with the first-level efficient frontier DMUs as best and ending with the Lth-level efficient frontier DMUs as worst. Note however that ties exist between all DMUs on the same efficient frontier. In order to get rid of the ties, one could use measures of relative attractiveness or relative progress to rank DMUs on the same frontier, where relative attractiveness is obtained when DMUs having worse performance are chosen as the evaluation context,

and relative progress is obtained when DMUs having better performance are chosen as the evaluation context. When the input-oriented (respectively, output-oriented) model of Table 3 is used to compute the efficient frontiers, attractiveness measures $\theta_k(d)$ (respectively, $\phi_k(d)^{-1}$) could be computed by solving the relevant model of Table 4 for values of d ranging from 1 to $L-\ell$, where $j \in F(E_\ell)$ means that $DMU_j \in E_\ell$ – $\theta_k(d)$ and $\phi_k(d)^{-1}$ are referred to as d-degree attractiveness. Note that the larger the value of $\theta_k(d)$ (respectively, $\phi_k(d)^{-1}$) the more attractive DMU_k. On the other hand, progress measures $\theta_k(d)^{-1}$ or $\phi_k(d)$ could be computed by solving the relevant model of Table 5 for values of d ranging from 1 to $\ell-1$ - these measures are referred to as d-degree progress. Note that the larger the value of $\theta_k(d)^{-1}$ (respectively, $\phi_k(d)$) the more attractive DMU_k. Note also that the difference between the models used for computing attractiveness and progress measures and those of Table 3 lies in the fact that the DMU_k under consideration is compared to a specific subset of DMUs in models of Table 4 and Table 5 instead of being compared to all DMUs in models of Table 3.

Input-Oriented	**Output-Oriented**
Minimize $\theta_k(d) - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)$	Maximize $\phi_k(d) + \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)$
s.t. : $\sum_{j \in F(E_{\ell+d})} \lambda_j x_{i,j} + s_i^- = \theta_k(d) \cdot x_{i,k}; \forall i$	s.t. : $\sum_{j \in F(E_{\ell+d})} \lambda_j x_{i,j} + s_i^- = x_{i,k}; \forall i$
$\sum_{j \in F(E_{\ell+d})} \lambda_j y_{r,j} - s_r^+ = y_{r,k}; \forall r$	$\sum_{j \in F(E_{\ell+d})} \lambda_j y_{r,j} - s_r^+ = \phi_k(d) \cdot y_{r,k}; \forall r$
$\lambda_j \geq 0; \forall j \in F(E_{\ell+d})$	$\lambda_j \geq 0; \forall j \in F(E_{\ell+d})$
$s_i^- \geq 0; \forall i; s_r^+ \geq 0; \forall r$	$s_i^- \geq 0; \forall i; s_r^+ \geq 0; \forall r$
$\theta_k(d)$ unrestricted	$\phi_k(d)$ unrestricted

Table 4: Basic DEA Models to Compute Attractiveness Measures

Input-Oriented	Output-Oriented
Minimize $\theta_k(d) - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)$	Maximize $\phi_k(d) + \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)$
s.t. : $\sum_{j \in F(E_{\ell-d})} \lambda_j x_{i,j} + s_i^- = \theta_k(d) \cdot x_{i,k}; \forall i$	s.t. : $\sum_{j \in F(E_{\ell-d})} \lambda_j x_{i,j} + s_i^- = x_{i,k}; \forall i$
$\sum_{j \in F(E_{\ell-d})} \lambda_j y_{r,j} - s_r^+ = y_{r,k}; \forall r$	$\sum_{j \in F(E_{\ell-d})} \lambda_j y_{r,j} - s_r^+ = \phi_k(d) \cdot y_{r,k}; \forall r$
$\lambda_j \geq 0; \forall j \in F(E_{\ell-d})$	$\lambda_j \geq 0; \forall j \in F(E_{\ell-d})$
$s_i^- \geq 0; \forall i; s_r^+ \geq 0; \forall r$	$s_i^- \geq 0; \forall i; s_r^+ \geq 0; \forall r$
$\theta_k(d)$ unrestricted	$\phi_k(d)$ unrestricted

Table 5: Basic DEA Models to Compute Progress Measures

To conclude this section, we would like to discuss one last decision; namely, how attractiveness and progress measures are used to rank competing forecasting models on the same efficient frontier. In sum, to overcome potential infeasibility problems under variable returns to scale (VRS) conditions, we compute global progress measures and use them to rank order the forecasting models of the same efficient frontier. To be more specific, for each performance level ℓ ranging from 2 to L, we compute progress measures using the relevant model, where d is chosen so that $\ell - d = 1$ or equivalently the evaluation context is the first-level efficient frontier, and use such measures to rank order the forecasting models of the ℓth-level efficient frontier. The choice of the first-level efficient frontier as the evaluation context for all efficient frontiers is made to obtain a robust and global ranking. As to the DMUs of the first-level efficient frontier, one could rank them using attractiveness measures under constant returns to scale conditions or accept their equal performance under VRS conditions. In the next section, we use CDEA to evaluate the performance of competing forecasting models of crude oil prices.

3.4. Illustrative Application of CDEA Framework for Model Evaluation and Selection in Forecasting

In this section, we use forecasting models of crude oil prices to illustrate the use of the proposed multidimensional performance evaluation framework. As a strategic commodity, crude oil has attracted the attention of investors, analysts, and academic

researchers. The relevant literature could be divided into two main categories; namely, data modeling and forecasting. Although our main interest is in the forecasting area, we would like to draw the reader's attention to the fact that research concerned with data modeling provides valuable input to model building in forecasting (Asche et al., 2003; Bekiros and Diks, 2008; Bacon, 1991; Cuñado and Perez de Gracia, 2003; Hamilton, 1983, 1985, 1996, 2009; Kilian, 2008, 2009). As to research on forecasting crude oil, it addresses several crude oil related variables such as prices, returns, supply, and demand. As far as prices and returns are concerned, quantitative forecasting models could be divided into three main categories; namely, non-artificial intelligence models, artificial intelligence models, and hybrid models. Non-artificial intelligence models include time series models; e.g., *random walk (RW) models* (Zeng and Swanson, 1998; Kaboudan, 2001; Lalonde et al., 2003; Abosedra, 2005; Knetsch, 2007; Coppola, 2008) and *autoregressive integrated moving average* (*ARIMA) models* (Sequeira and McAleer, 2000; Lalonde et al., 2003; Fernandez, 2006; Xie et al., 2006; Moshiri and Foroutan, 2006), and explanatory models; e.g., *linear regression models* (Bopp and Lady, 1991; Ye et al., 2002, 2005, 2006a, 2006b; Sequeira and McAleer, 2000; Abosedra and Baghestani, 2004; Knetsch, 2007), *vector autoregressive (VAR) models* (Zeng and Swanson, 1998), *error correction (EC) and vector error correction (VEC) models* (Zeng and Swanson, 1998; Sequeira and McAleer, 2000; Longo et al., 2007; Coppola, 2008; Murat and Tokat, 2009), *wavelet transform-based models* (Yousefi et al., 2005), and *state space models* (Pindyck, 1999; Bernard et al., 2004). Note that most of the variables used to build explanatory models are suggested by data modeling studies. As to artificial intelligence models, they include artificial neural networks (e.g., Kaboudan, 2001; Mirmirani and Li, 2004; Fernandez, 2006; Suriya, 2006; Xie et al., 2006; Yu et al., 2007; Fan et al., 2008; Yu et al., 2008; Ghaffari and Zare, 2009), genetic programming-based models (e.g., Kaboudan, 2001; Mirmirani and Li, 2004; Matilla-Garcia, 2007; Fan et al., 2008), pattern recognition-based models (e.g., Fernandez, 2006; Xie et al., 2006; Fan et al., 2008), and belief network-based methods (e.g., Abramson and Finizza, 1991; Abramson, 1994). Finally, the integration of non-artificial intelligence and artificial

intelligence models has let to what is referred to as hybrid models (e.g., Abramson and Finizza, 1995; Wang et al., 2004, 2005; Fan et al., 2008; Yu et al., 2008).

In this paper, the set of competing forecasting models or DMUs to consider for evaluation is chosen as a subset of the non-artificial intelligence models proposed in the literature; namely, *RW models* (Zeng and Swanson, 1998; Kaboudan, 2001; Lalonde et al., 2003; Abosedra, 2005; Knetsch, 2007; Coppola, 2008), *ARIMA models* (Sequeira and McAleer, 2000; Lalonde et al., 2003; Fernandez, 2006; Xie et al., 2006; Moshiri and Foroutan, 2006), *linear regression models* (Bopp and Lady, 1991; Ye et al., 2002, 2005, 2006a, 2006b; Sequeira and McAleer, 2000; Abosedra and Baghestani, 2004; Knetsch, 2007), *VAR models* (Zeng and Swanson, 1998), and *EC and VEC models* (Zeng and Swanson, 1998; Sequeira and McAleer, 2000; Longo et al., 2007; Coppola, 2008). In addition, among these models only valid ones are considered, where the validity of a model refers to the fact that the underlying assumptions hold over the chosen period of study; namely, January 1994 to August 2007. Note that there are several reasons for choosing this specific period; namely, change in International Energy Agency data collection methodology in 1990, non-availability of data on all relevant variables before January 1994, and recent credit crunch. Note also that some models have been discarded because they are dominated by one or several models on all criteria under consideration or because of the unavailability of data on the explanatory variables they use.

In sum, Table 6 summarizes the remaining ten forecasting models that are valid for at least one implementation method including the Holt-Winter's model with multiplicative seasonality, which was not reported in the literature on forecasting crude oil prices, as another benchmark along with random walk adjusted for trend – for details on these models and definitions of the explanatory variables used in linear regression models, the reader is referred to the original papers as referenced in the table. Notice that, the fourth implementation method; that is, out-of-sample implementation with rolling origin and fixed window, results in all models being valid and is the one that we use for measuring the performance of these models on each criterion under consideration – for a review on implementation methods, the reader is referred to Tashman (2000). With respect to the

performance criteria and their measures that will be used for evaluation of the models of Table 6, we deliberately restricted ourselves to only consider the reliability criterion and its sub-criteria, because of the lack of data on the remaining criteria or their irrelevance in the academic context. Note that, in our application, theoretical relevance will not be taken into account to avoid penalizing models that are not necessarily theoretically relevant, but do a good job in forecasting crude oil prices. In addition, as most statistical and econometric software packages at our disposal do not provide prediction intervals, the degree of uncertainty of forecasts will not be considered. Furthermore, our empirical results revealed that all the models attempted in this study – including the valid ones, are not informationally efficient; therefore, as informational efficiency does not discriminate between models, it will be discarded.

Valid Forecasting Models	**Implementation Methods**				
	1	2	3	4	5
1. RW with Trend (Zeng and Swanson, 1998)	X	√	√	√	X
2. Holt-Winter Exponential Smoothing with Multiplicative Seasonality (HWESMS)	X	X	√	√	√
3. ARIMA (1,1,1)	√	√	√	√	√
4. ARIMA (1,1,1) (1,0,1)	√	√	√	√	√
5. REG1: $WTI_t = \beta_0 + \beta_1 WTI_{t-1} + \sum_{i=0}^{3} \beta_{i+2} RIN_{t-i} + \sum_{j=0}^{5} \beta_{j+6} D_j\,911 + \beta_{12} APR99 + \varepsilon_t$ (Ye et al., 2005)	X	X	X	√	X
6. REG2: $WTI_t = \beta_0 + \beta_1 WTI_{t-1} + \beta_2 OECD\_Stocks_{t-1} + \beta_3 ANN_{t-1} + \beta_4 T + \sum_{j=0}^{5} \beta_{j+5} D_j\,911 + \beta_{12} LAPR99 + \varepsilon_t$ (Ye et al., 2005)	X	X	X	√	X
7. REG3: $WTI_t = \beta_0 + \beta_1 AR(1) + \beta_2 AR(12) + \sum_{j=0}^{5} \beta_{j+3} D_j\,911 + \beta_9 LAPR99 + \varepsilon_t$ (Ye et al., 2005)	X	X	X	√	√
8. REG4: $WTI_t = \beta_0 + \beta_1 WTI_{t-1} + \sum_{i=0}^{3} \beta_{i+2} RIN_{t-i} + \sum_{i=0}^{3} (\gamma_i LIN_{t-i} + \gamma_i^* LIN2_{t-i}) + \sum_{i=0}^{3} (\gamma_{i+4} HIN_{t-i} + \gamma_{i+4}^* HIN2_{t-i}) + \sum_{j=1}^{6} \beta_{j+9} D_j\,911 + \beta_{12} LAPR99 + \varepsilon_t$ (Ye et al., 2006a)	X	√	√	√	X
9. REG5: $WTI_t = \beta_0 + \beta_1 WTI_{t-1} + \sum_{i=0}^{1} \beta_{i+2} \Delta RIN_{t-i} + \beta_4 \log(Excess\_cap\_OPEC) + \sum_{j=1}^{6} \beta_{j+4} D_j\,911 + \beta_{11} LAPR99 + + \varepsilon_t$ (Ye et al., 2006b)	X	X	X	√	√
10. REG6: $WTI_t = \beta_0 + \beta_1 WTI_{t-1} + \sum_{i=0}^{4} \beta_{i+2} \Delta RIN_{t-i} + \beta_7 \Delta RIN_{t-5} + \sum_{j=0}^{4} \beta_{j+8} \Delta LIN_{t-i} + \beta_{13} LIN_{t-5} + \beta_{14} AIN_t + \varepsilon_t$ (Ye et al., 2002)	X	X	X	√	√

[1]In-Sample Implementation; [2]Out-of-sample with fixed origin & static forecast; [3]Out-of-sample with fixed origin & dynamic forecast; [4]Out-of-sample with rolling origin & fixed window; [5]Out-of-sample with rolling origin & variable window; [X]Invalid Model; [√]Valid Model

Table 6: Valid Forecasting Models for At Least One Implementation Method

The performance measures of the models summarized in Table 6 are reported in Table 7, where PSTSU denotes the proportion of statistical tests supporting unbiasedness. Note that the comparison of forecasting models with respect to biasedness as measured by ME,

MPE, MAdjPE, MTrdScE, or MVolScE depends on the decision making situation; therefore, we have chosen to use an aggregate measure of the results of statistical tests of biasedness; namely, the proportion of tests supporting unbiasedness, PSTSU, instead of the above mentioned measures.

	DMU No.	1	2	3	4	5	6	7	8	9	10
	Measures & Tests	RW With Trend	HWESMS	ARIMA (111)	ARIMA (111)(101)	REG1	REG2	REG3	REG4	REG5	REG6
Biasedness	ME	-0.722	0.249	0.470	0.451	0.359	0.464	0.649	0.403	0.202	0.247
	MPE	-1.318	0.209	0.828	0.779	0.917	0.749	1.403	0.997	0.510	0.460
	MTrdScE	-1.000	0.345	0.652	0.625	0.497	0.643	0.900	0.558	0.279	0.342
	MVolScE	-0.003	0.001	0.002	0.002	0.001	0.002	0.003	0.002	0.001	0.001
	MAdjPE	-1.046	0.447	1.103	1.055	1.158	1.042	1.697	1.248	0.740	0.706
Goodness-of-Fit	MSE	12.900	11.433	13.052	13.273	11.545	13.810	13.920	11.744	11.140	11.821
	MSPE	54.025	48.389	55.238	55.566	48.145	59.226	58.755	49.777	46.164	49.758
	MMSEU	12.915	11.462	13.088	13.303	11.582	13.850	13.965	11.771	11.186	11.865
	MMSEO	12.942	11.477	13.084	13.299	11.580	13.841	13.937	11.781	11.170	11.853
	MTrdScSE	17.878	15.845	18.089	18.394	16.000	19.138	19.292	16.276	15.439	16.383
	MVolScSE	0.053	0.047	0.053	0.054	0.047	0.056	0.057	0.048	0.046	0.048
	MSAdjPE	55.055	47.048	54.749	54.967	48.349	58.107	58.776	50.768	45.899	48.937
	MAE	2.876	2.603	2.921	2.928	2.809	2.968	3.060	2.870	2.726	2.755
	MAPE	6.088	5.509	6.186	6.189	5.987	6.332	6.514	6.126	5.748	5.845
	MMAEU	7.022	6.847	8.328	8.381	7.422	8.745	9.002	7.897	6.875	7.234
	MMAEO	8.811	7.262	7.711	7.877	7.004	8.105	8.039	6.781	7.066	7.419
	MTrdScAE	3.986	3.607	4.048	4.058	3.893	4.114	4.241	3.978	3.778	3.818
	MVolScAE	0.012	0.011	0.012	0.012	0.011	0.012	0.013	0.012	0.011	0.011
	MAAdjPE	6.081	5.484	6.198	6.197	6.016	6.331	6.554	6.178	5.755	5.838
Correct Sign	PCSP	1.000	0.984	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	PCDCP	1.000	0.609	0.594	0.531	0.609	0.516	0.406	0.641	0.672	0.625
Biasedness	Sign test - Exact Binomial	0	1	1	1	1	1	0	1	1	1
	Sign Test - Normal Approximation	0	1	1	1	1	1	0	1	1	1
	Regression Based Biasedness Test	0	1	1	1	0	1	1	0	1	1
	PSTSU	0	1	1	1	0.666	1	0.333	0.666	1	1
Informational Efficiency	Informational Efficiency Test (Beta)	0	0	0	0	0	0	0	0	0	0

Table 7: Performance Measures Corresponding to Out-of-sample Implementation with Rolling Origin and Fixed Window

As to the choice of inputs and outputs to be used within CDEA, any measure of goodness-of-fit could be used as an input, as it needs to be minimized, and PSTSU as well as PCDCP – as measures of unbiasedness and correct sign prediction, respectively, will be used as outputs, as they need to be maximized. Our returns-to-scale analysis revealed that the constant returns-to-scale assumption is invalid for our application data; therefore, we performed CDEA under variable returns-to-scale (VRS). The efficient

frontiers with different performance levels identified by CDEA-VRS with an input orientation are reported in Table 8-13 for different inputs.

Level	Efficient frontier (Efficient Forecasting Models)
1	RW With Trend and REG5
2	HWESMS, REG4 and REG6
3	ARIMA111000 and REG1
4	ARIMA111101
5	REG2
6	REG3

Table 8: Efficient Frontiers with Different Performance Levels Based on One of The Following Inputs: MSE, MMSEU, MMSEO, MTrdScSE, MVolScSE, MSAdjPE

Level	Efficient frontier (Efficient Forecasting Models)
1	RW With Trend and REG5
2	HWESMS, REG1, REG4 and REG6
3	ARIMA111000
4	ARIMA111101
5	REG2 and REG3

Table 9: Efficient Frontiers with Different Performance Levels Based on Input: MSPE

Level	Efficient frontier (Efficient Forecasting Models)
1	RW With Trend, HWESMS and REG5
2	REG4 and REG6
3	ARIMA111000 and REG1
4	ARIMA111101
5	REG2
6	REG3

Table 10: Efficient Frontiers with Different Performance Levels Based on One of The Following Inputs: MAE, MAPE, MMAEU, MtrdScAE

Level	Efficient frontier (Efficient Forecasting Models)
1	RW With Trend, REG4 and REG5
2	HWESMS, REG1 and REG6
3	ARIMA111000
4	ARIMA111101
5	REG2 and REG3

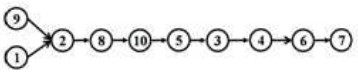
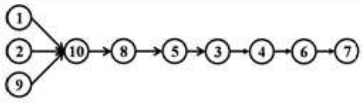
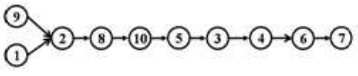
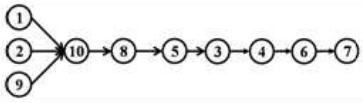
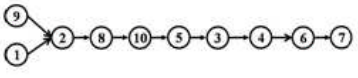
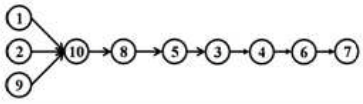
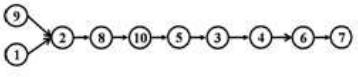
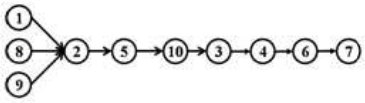
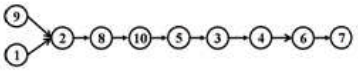
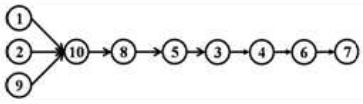
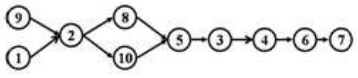
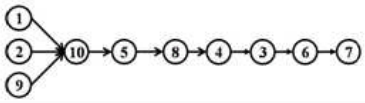
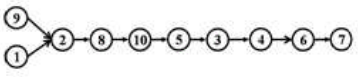
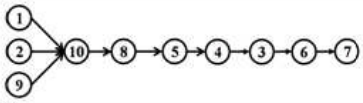
Table 11: Efficient Frontiers with Different Performance Levels Based on Input: MMAEO

Level	Efficient frontier (Efficient Forecasting Models)
1	RW With Trend, HWESMS and REG5
2	REG1, REG4 and REG6
3	ARIMA111000 and ARIMA111101
4	REG2
5	REG3

Table 12: Efficient Frontiers with Different Performance Levels Based on Input: MVolScAE

Level	Efficient frontier (Efficient Forecasting Models)
1	RW With Trend, HWESMS and REG5
2	REG4 and REG6
3	ARIMA111000, ARIMA111101 and REG1
4	REG2
5	REG3

Table 13: Efficient Frontiers with Different Performance Levels Based on Input: MAAdjPE

Performance Measures	Rank in Decreasing Order	Performance Measures	Rank in Decreasing Order
Input: MSE **Outputs**: PCDCP & PSTSU		**Input**: MAE **Outputs**: PCDCP & PSTSU	
Input: MSPE **Outputs**: PCDCP & PSTSU		**Input**: MAPE **Outputs**: PCDCP & PSTSU	
Input: MMSEU **Outputs**: PCDCP & PSTSU		**Input**: MMAEU **Outputs**: PCDCP & PSTSU	
Input: MMSEO **Outputs**: PCDCP & PSTSU		**Input**: MMAEO **Outputs**: PCDCP & PSTSU	
Input: MTrdScSE **Outputs**: PCDCP & PSTSU		**Input**: MTrdScAE **Outputs**: PCDCP & PSTSU	
Input: MVolScSE **Outputs**: PCDCP & PSTSU		**Input**: MVolScAE **Outputs**: PCDCP & PSTSU	
Input: MSAdjPE **Outputs**: PCDCP & PSTSU		**Input**: MAAdjPE **Outputs**: PCDCP & PSTSU	

"RW with Trend; "HWESMS; "ARIMA (111); "ARIMA (111) (101); "REG1; "REG2; "REG3; "REG4; "REG5; "REG6

Table 14: CDEA Rankings of Competing Forecasting Models

	Level	DMU/Model Name	Progress Score		Level	DMU/Model Name	Progress Score
MSE	2	HWESMS	1.026	**MAE**	2	REG4	1.0932
		REG4	1.054			REG6	1.0458
		REG6	1.061		3	ARIMA(111)	1.1222
	3	ARIMA(111)	1.172			REG1	1.0791
		REG1	1.036		4	ARIMA(111)(101)	1.1249
	4	ARIMA(111)(101)	1.191		5	REG2	1.1402
	5	REG2	1.240		6	REG3	1.1756
	6	REG3	1.250				
MSPE	2	HWESMS	1.048	**MAPE**	2	REG4	1.103
		REG1	1.043			REG6	1.049
		REG4	1.078		3	ARIMA(111)	1.123
		REG6	1.078			REG1	1.087
	3	ARIMA(111)	1.197		4	ARIMA(111)(101)	1.123
	4	ARIMA(111)(101)	1.204		5	REG2	1.149
	5	REG2	1.283		6	REG3	1.183
		REG3	1.273				
MMSEU	2	HWESMS	1.025	**MMAEU**	2	REG4	1.151
		REG4	1.052			REG6	1.055
		REG6	1.061		3	ARIMA(111)	1.216
	3	ARIMA(111)	1.170			REG1	1.084
		REG1	1.035		4	ARIMA(111)(101)	1.224
	4	ARIMA(111)(101)	1.189		5	REG2	1.277
	5	REG2	1.238		6	REG3	1.315
	6	REG3	1.248				
MMSEO	2	HWESMS	1.028	**MMAEU**	2	HWESMS	1.028
		REG4	1.055			REG1	1.033
		REG6	1.061			REG6	1.050
	3	ARIMA(111)	1.171		3	ARIMA(111)	1.091
		REG1	1.037		4	ARIMA(111)(101)	1.115
	4	ARIMA(111)(101)	1.191		5	REG2	1.147
	5	REG2	1.239			REG3	1.186
	6	REG3	1.248				
MTrdScSE	2	HWESMS	1.026	**MTrdScAE**	2	REG4	1.093
		REG4	1.054			REG6	1.046
		REG6	1.061		3	ARIMA(111)	1.122
	3	ARIMA(111)	1.172			REG1	1.079
		REG1	1.036		4	ARIMA(111)(101)	1.125
	4	ARIMA(111)(101)	1.191		5	REG2	1.140
	5	REG2	1.240		6	REG3	1.176
	6	REG3	1.250				
MVolScSE	2	HWESMS	1.016	**MVolScAE**	2	REG1	1.034
		REG4	1.043			REG4	1.118
		REG6	1.043			REG6	1.025
	3	ARIMA(111)	1.152		3	ARIMA(111)	1.128
		REG1	1.022			ARIMA(111)(101)	1.125
	4	ARIMA(111)(101)	1.179		4	REG2	1.128
	5	REG2	1.217		5	REG3	1.222
	6	REG3	1.239				

MSAdjPE	2	HWESMS	1.025	MAAdjPE	2	REG4	1.117
		REG4	1.106			REG6	1.051
		REG6	1.066		3	ARIMA(111)	1.130
	3	ARIMA(111)	1.193			ARIMA(111)(101)	1.130
		REG1	1.053			REG1	1.097
	4	ARIMA(111)(101)	1.198		4	REG2	1.154
	5	REG2	1.266		5	REG3	1.195
	6	REG3	1.281				

Table 15: CDEA Progress Scores

	Ranking in Descending Order of A Specific Performance Measure										
Goodness-of-Fit	MSE	REG5	HWESMS	REG1	REG4	REG6	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MSPE	REG5	REG1	HWESMS	REG6	REG4	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG3	REG2
	MMSEU	REG5	HWESMS	REG1	REG4	REG6	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MMSEO	REG5	HWESMS	REG1	REG4	REG6	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MTrdScSE	REG5	HWESMS	REG1	REG4, REG6		RW With Trend	A ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MVolScSE	REG5	HWESMS	REG1	REG6	REG4	ARIMA (111)	ARIMA (111)(101)	RW With Trend	REG2	REG3
	MSAdjPE	REG5	HWESMS	REG1	REG4	REG6	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MAE	HWESMS	REG5	REG6	REG1	REG4	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MAPE	HWESMS	REG5	REG6	REG1	RW With Trend	REG4	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MMAEU	HWESMS	REG5	RW With Trend	REG6	REG1	REG4	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MMAEO	REG4	REG1	REG5	HWESMS	REG6	ARIMA (111)	ARIMA (111)(101)	REG3	REG2	RW With Trend
	MTrdScAE	HWESMS	REG1, REG5, REG6			RW With Trend	ARIMA (111)(101)	ARIMA (111)	REG2 REG4		REG3
	MVolScAE	HWESMS	REG5	REG6	REG1	RW With Trend	REG4	ARIMA (111)(101)	ARIMA (111)	REG2	REG3
	MAAdjPE	HWESMS	REG5	REG6	REG1	REG4	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
Correct Sign	PCDCP	RW With Trend	REG5	REG4	REG6	HWESMS, REG1		ARIMA (111)	ARIMA (111)(101)	REG2	REG3
Biasedness	PSTSU	HWESMS, ARIMA(111), ARIMA(111)(101), REG2, REG5, REG6						REG1, REG4		REG3	RW With Trend

Table 16: Unidimensional Rankings of Competing Forecasting Models

Note that some inputs have the same set of efficient frontiers; namely, MSE, MMSEU, MMSEO, MTrdScSE, MVolScSE and MSAdjPE for measures based on squared errors, and MAE, MAPE, MMAEU and MTrdScAE for measures based on absolute errors. These results reveal that, regardless of whether performance measures are based on squared errors or absolute errors, RW with Trend and REG5 are consistently the best

whereas REG2 and REG3 are consistently the worst; however, HWESMS tend to be privileged by all measures based on absolute errors except MMAEU.

The rankings obtained by CDEA-VRS with an input orientation are summarized in Table 14; see Table 15 for detailed information on progress scores. For comparison purposes, unidimensional rankings are reported in Table 16. Notice that CDEA-VRS rankings corresponding to the measures of input or goodness-of-fit MSE, MSPE, MMSEU, MMSEO, MTrdScSE and MSAdjPE are identical, but for different reasons. In fact, MSE, MMSEU, MMSEO, MTrdScSE and MSAdjPE lead to the same partition of the set of competing forecasting models into efficient frontiers of different performance levels, on one hand, and have similar progress measures, on the other hand. As to MSPE, although it leads to a different partitioning of the set of competing forecasting models under consideration, the values of progress measures are such that REG1 is ranked last amongst the models of the second-level efficient frontier and ranked first amongst the third-level efficient frontier – see Table 9. Furthermore, the values of progress measures are such that REG2 outperforms REG3 – see Table 9 and Table 15. With respect to MVolScSE however the tie between REG4 and REG6 is due to identical values of progress measures. Notice also that CDEA-VRS rankings corresponding to the measures of goodness-of-fit MAE, MAPE, MMAEU and MTrdScAE are also identical, because these measures lead to the same partition of the set of competing forecasting models into efficient frontiers of different performance levels, on one hand, and have similar progress measures, on the other hand – see Table 10. As to MMAEO, MVolScAE and MAAdjPE, the corresponding rankings are different, because they lead different partitions of the set of competing forecasting models as well as different values of progress measures – see Tables 11-13. Last, but not least, we would like to draw the reader's attention to the fact that although the unidimensional rankings of models under several measures are different – see Table 16, their multidimensional rankings could be identical – see Table 14.

3.5. Conclusion

The lack of a multidimensional framework for performance evaluation of competing forecasting models has motivated this research in which we proposed a framework based on context-dependent data envelopment analysis under variable returns-to-scale (CDEA-VRS). We also surveyed and classified the literature on performance criteria and their measures – including statistical tests – commonly used in evaluating and selecting forecasting models or methods. In order to illustrate the use of the proposed CDEA-VRS framework, we have chosen ten forecasting models of crude oil prices. 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, linear regression models such as REG2, REG3 and REG5, Holt-Winter Exponential Smoothing with Multiplicative Seasonality and Random Walk adjusted for Trend tend to have ranks that are not sensitive to performance measures, which suggest that the rankings of these models are reliable. Furthermore, REG5, Holt-Winter Exponential Smoothing with Multiplicative Seasonality and Random Walk adjusted for Trend are superior to the remaining models. Finally, in practice, we recommend that the forecasts produced by models with similar performance such as these should be combined and compared to all forecasts produced by individual models before deciding on the forecasting model or method to implement.

Reference

Abosedra S, Baghestani H. On the predictive accuracy of crude oil future prices. Energy Policy 2004; 32; 1389–1393.

Abramson B. The design of belief network-based systems for price forecasting. Computers and Electrical Engineering 1994; 20; 163–180.

Abramson B, Finizza A. Using belief networks to forecast oil prices. International Journal of Forecasting 1991; 7; 299–315.

Abramson B, Finizza A. Probabilistic forecasts from probabilistic models: a case study in the oil market. International Journal of Forecasting 1995; 11; 63–72.

Aigner D.J., Lovell C.A.K., Schmidt P. Formulation and estimation of stochastic frontier production functions. Journal of Econometrics 1977; 6; 21-37.

Allen R, Athanasopoulos A, Dyson R.G., Thanassoulis E. Weights restrictions and value judgments in data envelopment analysis: Evolution, development and future directions. Annals of Operations Research 1997; 73; 13–34.

Andersen P, Petersen N.C. A procedure for ranking efficient units in data envelopment analysis. Management Science 1993; 39; 1261–1294.

Anderson T.R., Sharp G.P. A new measure of baseball batters using DEA, Annals of Operations Research 1997; 73; 141-155.

Armstrong J.S. Long-range forecasting: From Crystal Ball to computer. John Wiley: New York; 1985.

Armstrong J.S. Evaluating forecasting methods. In: Armstrong JS (Eds), Principles of forecasting: A handbook for researchers and practitioners; Kluwer Academic Publishers: Boston; 2001a. p. 443 – 472.

Armstrong J.S. Standards and practices for forecasting. In: Armstrong JS (Eds), Principles of forecasting: A handbook for researchers and practitioners; Kluwer Academic Publishers: Boston; 2001b. p. 679 – 732.

Armstrong J.S., Adya M, Collopy F. Standards and practices for forecasting. In: Armstrong JS (Eds), Principles of forecasting: A handbook for researchers and practitioners; Kluwer Academic Publishers: Boston; 2001. p. 285–300.

Asche F, Gjolberg O, Volker T. Price relationships in the petroleum market: an analysis of crude oil and refined product prices. Energy Economics 2003; 25; 289-301.

Bacon R. Modelling the price of oil. Oxford Review of Economic Policy 1991; 7; 17–34.

Banker R.D., Charnes A, Cooper W.W. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science; 1984; 30; 1078–1092.

Banker R.D., Kemerer C.F. Scale economies in new software development. IEEE Transactions on Software Engineering 1989; 15; 1199- 1205.

Banker R.D., Morey R.C. Efficiency analysis for exogenously fixed inputs and outputs. Operations Research 1986a; 34; 513–521.

Banker R.D., Morey R.C. The use of categorical variables in data envelopment analysis. Management Science 1986b; 32; 1613–1627.

Barr R.S. Siems T.F. Bank failure prediction using DEA to measure management quality. In Barr RS, Helgason RV, Kennington JL (Eds) Advances in Metaheuristics, Optimization, and Stochastic Modeling Techniques. Kluwer Academic Publishers: Boston; 1997. p. 341-365.

Bekiros S, Diksa C. The relationship between crude oil spot and futures prices: Cointegration, linear and nonlinear causality. Energy Economics 2008; 30; 2673-2685.

Bernard J.T., Khalaf L, Kichian M. Structural change and forecasting long-run energy prices. Bank of Canada 2004; Working Papers 04-5.

Birman S.V., Pirondi P.E., Rodin E.Y. Application of DEA to medical clinics. Mathematical and computer modelling 2003; 37; 923-936.

Bopp A.E., Lady G.M. A comparison of petroleum futures versus spot prices as predictors of prices in the future. Energy Economics 1991; 13; 274-282.

Box G.E.P., Cox D.R. An Analysis of Transformations. Journal of the Royal Statistical Society 1964; 26; 211-252.

Brockett P.L., Cooper W.W., Shin H.C., Wang Y. Inefficiency and congestion in Chinese production before and after the 1978 economic reforms. Socio-Economic Planning Sciences 1998; 32; 1-20.

Carbone R. Armstrong J.S. Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners. Journal of Forecasting 1982; 1; 215-217.

Chang H. Determinants of hospital efficiency: The case of central government-owned hospitals in Taiwan. Omega 1998; 26; 307–317.

Chang H, Cheng M, Das S. Hospital ownership and operating efficiency: Evidence from Taiwan. European Journal of Operational Research 2004; 159; 513–527.

Charnes A, Cooper W.W., Golany B, Seiford L.M., Stutz J. Foundations of data envelopment analysis and Pareto–Koopmans empirical production functions, Journal of Econometrics 1985; 30; 91–107.

Charnes A, Cooper W.W., Huang Z.M., Sun D.B. Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks. Journal of Econometrics 1990; 46; 73–91.

Charnes A, Cooper W.W., Li S. Using data envelopment analysis to evaluate the efficiency of economic performance by Chinese cities. Socio-Economic Planning Science 1989; 23; 325–344.

Charnes A, Cooper W.W., Lewin A.Y. Seiford L.M. Data envelopment analysis: theory, methodology and applications. Kluwer Academic Publishers: Boston; 1994.

Charnes A, Cooper W.W., Rhodes E. Measuring the efficiency of decision making units. European Journal of Operational Research 1978; 2; 429–444.

Charnes A, Cooper W.W., Rhodes E. Short Communication: Measuring the Efficiency of Decision Making Units. European Journal of Operational Research 1979; 3; 339.

Charnes A, Cooper W.W., Rhodes E. Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through. Management Science 1981; 27; 668-697.

Cheng E.W.L., Chiang Y.H., Tang B.S. Alternative approach to credit scoring by DEA: Evaluating borrowers with respect to PFI projects. Building and Environment 2007; 42; 1752-1760.

Chen Y. Ranking efficient units in DEA. Omega 2004; 32; 213-219

Chen Y. Measuring super-efficiency in DEA in the presence of infeasibility. European Journal of Operational Research 2005; 161; 545–551.

Chilingerian J.A., Sherman H.D. Health care applications: from hospitals to physicians, from productive efficiency to quality frontiers. In Cooper WW, Seiford LM, Zhu J (Eds). Handbook on Data Envelopment Analysis. Kluwer Academic Publishers: Boston. 2004. p.481-538.

Cook W.D., Kress M, Seiford L.M. On the use of ordinal data in data envelopment analysis. Journal of the Operational Research Society 1993; 44; 133–140.

Cook W.D., Kress M, Seiford L.M. Data envelopment analysis in the presence of both quantitative and qualitative factors. Journal of the Operational Research Society 1996; 47; 945–953.

Cook W.D., Green R, Zhu J. Dual role factors in DEA. IIE Transactions 2006; 38; 1-11.

Cook W.D., Liang L, Zha Y, Zhu J. A Modified Super-efficiency DEA Model for Infeasibility. Journal of Operational Research Society 2008; 69; 276-281.

Cook W.D., Roll Y, Kazakov A. A DEA model for measuring the relative efficiency of highway maintenance patrols. INFOR 1990; 28; 113–124.

Cook W.D., Seiford L.M. Data envelopment analysis (DEA) – thirty years on. European Journal of Operational Research 2009; 192; 1–17.

Cook W.D., Seiford L.M., Zhu J. Models for performance benchmarking: Measuring the effect of e-commerce activities on banking performance. Omega 2004; 32; 313-322.

Cook W.D., Zhu J. Rank order data in DEA: A general framework. European Journal of Operational Research 2006; 174; 1021–1038.

Cook W.D., Zhu J. Classifying inputs and outputs in data envelopment analysis. European Journal of Operational Research 2007; 180; 692-699.

Cook W.D., Zhu J. CAR-DEA: Context dependent assurance regions in DEA. Operations Research 2008; 56; 69-78.

Cooper W.W., Huang Z, Li S. Satisficing DEA models under change constraints. Annuals of Operations Research 1996; 66; 279 – 295.

Cooper W.W., Park K.S., Pastor J.T. RAM: Range adjusted measure of inefficiency for use with additive models and relations to other models and measures in DEA. Journal of Productivity Analysis 1999a; 11; 5–42.

Cooper W.W., Park K.S., Yu G. IDEA and AR-IDEA: Models for dealing with imprecise data in DEA. Management Science 1999b; 45; 597 – 607.

Cooper W.W., Seiford L.M., Tone K. Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software. Kluwer Academic Publishers: Boston; 2000.

Cooper W.W., Seiford L.M., Tone K. Introduction to Data Envelopment Analysis and its Uses: With DEA-solver Software and References. Springer Science and Business Media Inc: New York; 2007.

Cooper W.W., Seiford L.M., Zhu J. Handbook on Data Envelopment Analysis, Kluwer Academic Publishers, Boston; 2004.

Coppola A. Forecasting oil price movements: exploiting the information in the futures market. Journal of Futures Market 2008; 28; 34–56.

Crowder W. J., Hamid A. A cointegration test for oil futures market efficiency. Journal of Futures Markets 1993; 13; 933–941.

Cuñado J, Perez de Gracia F. Do Oil Price Shocks Matter? Evidence from Some European Countries. Energy Economics 2003; 25; 137-154

Dalrymple D.J. Sales forecasting methods and accuracy. Business Horizons 1975; 18; 69– 73.

Dalrymple D.J. Sales forecasting practices. International Journal of Forecasting 1987; 3; 379-391.

Dula J.H., Hickman B.L. Effects of excluding the column being scored from the DEA envelopment LP technology matrix. Journal of the Operational Research Society 1997; 48; 1001–1012.

Emel A.B., Oral M, Reisman A, Yolalan R. A credit scoring approach for the commercial banking sector. Socio-Economic Planning Sciences 2003; 37; 103–123.

Fan Y, Liang Q, Wei Y.M. A generalized pattern matching approach for multi-step prediction of crude oil price. Energy Economics 2006; 30; 889–904.

Färe R, Grosskopf S. Measuring congestion in production. Journal of Economics 1983; 43; 257-271.

Färe R.S., Grosskopf S. Modelling undesirable factors in efficiency evaluation: Comment. European Journal of Operational Research 2004; 157; 242-245.

Färe R.S., Grosskopf S, Lovell C.A.K. The measurement of efficiency of production. Kluwer-Nijhoff Publishers: Boston; 1985.

Färe R.S., Grosskopf S, Lovell C.A.K. Production Frontiers. Cambridge University Press: Cambridge; 1994.

Färe R.S., Lovell C.A.K. Measuring the technical efficiency of production. Journal of Economic Theory; 1978; 19; 150-162.

Fernandez V. Forecasting crude oil and natural gas spot prices by classification methods. Documentos de Trabajo from Centro de Economía Aplicada, Universidad de Chile, 2006; No 229.

Ghaffari A, Zare S. A novel algorithm for prediction of crude oil price variation based on soft computing. Energy Economics 2009; 31; 531-536

Goodwin P, Lawton R. On the asymmetry of the symmetric MAPE. International Journal of Forecasting 1999; 14; 405-408.

Green R.H., Cook W.D., Doyle J. A note on the additive data envelopment analysis model. Journal of the Operational Research Society 1997; 48; 446–448.

Gregoriou G.N., Sedzro K, Zhu J. Hedge fund performance appraisal using data envelopment analysis. European Journal of Operational Research 2005; 164; 555-571.

Hamilton J.D. Oil and the macroeconomy since World War II. Journal of Political Economy 1983; 91; 228–248.

Hamilton J.D. Historical causes of postwar oil shocks and recessions. Energy Journal 1985; 6; 97–116.

Hamilton J.D. This is what happened to the oil price–macroeconomy relationship. Journal of Monetary Economics 1996; 38; 215–220.

Hamilton J.D. Understanding Crude Oil Prices. Energy Journal 2009; 30; 179-206.

Hua Z, Bin Y. DEA with undesirable factors. In Zhu J, Cook WD (Eds), Modelling Data Irregularities and Structural Complexities in Data Envelopment Analysis. Springer Science: New York; 2007. p. 103-122.

Hyndman R.J., Koehler A.B. Another look at measures of forecast accuracy. International Journal of Forecasting 2006; 22; 679-688.

Jones J. Data envelopment analysis and its application to the measurement of efficiency in higher education. Economics of Education Review 2006; 25; 273-288.

Kaboudan M.A. Compumetric forecasting of crude oil prices. Proceedings of the 2001 Congress on Evolutionary Computation 2001; 1, 283-287.

Kamakura W.A. A note on the use of categorical variables in data envelopment analysis. Management Science 1988; 34; 1273–1276.

Kilian L. Exogenous oil supply shocks: how big are they and how much do they matter for the U.S. economy? Review of Economics and Statistics 2008; 90; 216–240.

Kilian L. Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market. American Economic Review 2009; 99; 1053-1069.

Kleinsorge I.K., Schary P.B., Tanner R.D. Evaluating logistics decisions, International Journal of Physical Distribution and Materials Management 1989; 19; 200-219.

Koehler A.B. The asymmetry of the sAPE measure and other comments on the M3-competition. International Journal of Forecasting 2001; 17; 537-584.

Lalonde R, Zhu Z, Demers F. Forecasting and Analyzing World Commodity Prices. Bank of Canada 2003; No. 24.

Lewin A.Y., Morey R.C., Cook T.J. Evaluating the administrative efficiency of courts. Omega 1982; 10; 401-11.

Longo C, Manera M, Markandya A, Scarpa E. Evaluating the empirical performance of alternative econometric models for oil price forecasting. FEEM Working Paper 2007; No 4.

Loungani P. Oil Price Shocks and the Dispersion Hypothesis. Review of Economics and Statistics 1986; 68; 536–539.

Lovell C.A.K., Rouse A.P.B. Equivalent standard DEA models to provide super-efficiency scores. Journal of the Operational Research Society 2003; 54; 101–108.

Mahmoud E. Accuracy in forecasting: A survey. Journal of Forecasting 1984; 3; 139–159.

Mahmoud E, Rice G. Malhotra N. Emerging issues in sales forecasting and decision support systems. Journal of Academy of Marketing Science 1986; 16; 47-61.

McCarthy T.M., Davis D.F., Golicic S.L., Mentzer J.T. The Evolution of sales forecasting management: A 20-year longitudinal study of forecasting practices. Journal of Forecasting 2006; 25; 303-324.

Mehrabian S, Alirezaee M.R., Jahanshahloo G.R. A complete efficiency ranking of decision making units in data envelopment analysis. Computational Optimization and Applications 1999; 14; 261–266.

Mentzer J.T., Cox J. A model of the determinants of achieved forecast accuracy. Journal of Business Logistics. Journal of Business Logistics 1984a; 5; 143–155.

Mentzer J.T., Cox J. Familiarity, application, and performance of sales forecasting techniques. Journal of Forecasting 1984b; 3; 27-36.

Mentzer J.T., Kahn K.B. Forecasting technique familiarity, satisfaction, usage, and application. Journal of Forecasting 1995; 14; 465–476.

Mincer J. Zarnowitz V. The evaluation of economic forecasts. In: Mincer J (Eds), Economic forecasts and expectations: Analyses of forecasting behavior and performance. Columbia University Press: New York; 1969.

Mirmirani S, Li H.C. A comparison of VAR and neural networks with genetic algorithm in forecasting price of oil. Advances in Econometrics 2004; 19; 203–223.

Oral M. Kettani O, Lang P. A methodology for collective evaluation and selection of industrial R&D projects. Management Science 1991; 7; 871–883.

Ramanathan R. A multi-factor efficiency perspective to the relationships among world GDP, energy consumption and carbon dioxide emissions. Technological Forecasting and Social Change 2006; 73; 483–494.

Ravikumar P, Ravi V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques: A review. European Journal of Operational Research 2007; 180; 1–28.

Ray S, Seiford L.M., Zhu J. Market entity behavior of Chinese State-Owned Enterprises. OMEGA 1998; 26; 263-278.

Retzlaff-Roberts D, Puelz R. Classification in automobile insurance using a DEA and discriminant hybrid. Journal of Productivity Analysis 1996; 17; 417–427.

Retzlaff-Roberts D, Chang C.F., Rubin R.M. Technical efficiency in the use of health care resources: a comparison of OECD countries. Health Policy 2004; 69; 55-72.

Pindyck R. S. The long-run evolution of energy prices. The Energy Journal 1999; 20; 1-27.

Roll Y, Cook W.D., Golany B. Controlling factor weights in data envelopment analysis. IIE Transactions 1991; 23; 2–9.

Rousseau J.J., Semple J.H. Categorical outputs in data envelopment analysis. Management Science 1993; 39; 384–386.

Ruggiero J. On the measurement of technical efficiency in the public sector. European Journal of Operational Research 1996; 90; 553-565.

Ruggiero J. Non-discretionary inputs in data envelopment analysis. European Journal of Operational Research 1998; 111; 461-469.

Ruggiero J. Non-discretionary inputs. In Zhu J, Cook WD (Eds), Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis. Springer Science: New York 2007. p.85-102.

Sanders N.R., Manrodt K.S. Forecasting practices in U.S. corporations. Interfaces 1994; 24; 92-100.

Sequeira J.M., McAleer M. A market-augmented model for SIMEX Brent crude oil futures contracts. Applied Financial Economics 2000; 10; 543–552.

Seiford L.M. Data envelopment analysis: the evolution of the state of the art (1978-1995). The Journal of Productivity Analysis 1996; 7; 99–137.

Seiford L.M. A bibliography for Data Envelopment Analysis (1978–1996). Annals of Operations Research 1997; 73; 393–438.

Seiford L.M., Zhu J. Sensitivity analysis of DEA models for simultaneous changes in all of the data. Journal of the Operational Research Society 1998a; 49; 1060–1071.

Seiford L.M., Zhu J. An acceptance system decision rule with data envelopment analysis. Computers and Operations Research 1998b; 25; 329–332.

Seiford L.M., Zhu J. An investigation of returns to scale in data envelopment analysis. Omega 1999a; 27; 1–11.

Seiford L.M., Zhu J. Infeasibility of super-efficiency data envelopment analysis models. INFOR 1999b; 37; 174–187.

Seiford L.M. Zhu J. Modeling undesirable factors in efficiency evaluation. European Journal of Operational Research 2002; 142; 16-20.

Seiford L.M. Zhu J. Context-dependent data envelopment analysis: measuring attractiveness and progress. Omega 2003; 31; 397-480.

Serletis A. A cointegration analysis of petroleum futures prices. Energy Economics 1994; 16; 93–97.

Shambora W.E., Rossitera R. Are there exploitable inefficiencies in the futures market for oil? Energy Economics 2007; 29; 18-27.

Scheel H. Undesirable outputs in efficiency valuations. European Journal of Operational Research 2001; 35; 109-126.

Sherman H.D., Gold F. Bank branch operating efficiency: Evaluation with data envelopment analysis. Journal of Banking and Finance 1985; 9; 297–316.

Sowlati T, Paradi J.C., Suld C. Information Systems Project Prioritization Using Data Envelopment Analysis. Mathematical and Computer Modelling 2005; 41; 1279-1298.

Suriya K. Forecasting Crude Oil Price Using Neural Networks. CMU Journal 2006; 5; 377-386.

Swanson D.A., Tayman J, Barr C. F. A Note on the Measurement of Accuracy for Subnational Demographic Estimates. Demography 2000; 37; 193-201.

Tashman L.J. Out of sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting 2000; 16; 437-450.

Thanassoulis E, Allen R. Simulating weights restrictions in data envelopment analysis by means of unobserved DMUs. Management Science 1998; 44; 586–594.

Thompson R.G., Dharmapala S. Thrall R.M. Linked-cone DEA profit ratios and technical inefficiencies with applications to Illinois coal mines. International Journal of Production Economics 1995; 39; 99–115.

Thompson R.G., Langemeir L.N., Lee C, Lee E, Thrall R.M. The role of multiplier bounds in efficiency analysis with application to Kansas farming. Journal of Econometrics 1990; 46; 93–108.

Thompson R.G., Singleton F.D. Thrall Jr. R.M., Smith B.A. Comparative site evaluations for locating a high-energy physics lab in Texas. Interfaces 1986; 16; 35–49.

Thrall R.M. Duality, classification and slacks in data envelopment analysis. The Annals of Operations Research 1996; 66; 109–138.

Tone K. A slacks-based measure of efficiency in data envelopment analysis. European Journal of Operational Research 2001; 130; 498–509.

Wang S.Y., Yu L, Lai K.K. A novel hybrid AI system framework for crude oil price forecasting. Lecture Notes in Computer Science 2004; 3327; 233–242.

Wang S.Y., Yu L, Lai K.K. Crude oil price forecasting with Tei@I methodology, Journal of Systems Science and Complexity 2005; 18; 145–166.

Wu J, Liang L, Chen Y. DEA game cross-efficiency approach to Olympic rankings. Omega 2009; 37; 909-918.

Xie W, Yu L, Xu S.Y., Wang S.Y. A new method for crude oil price forecasting based on support vector machines. Lecture Notes in Computer Science 2006; 3994; 441–451.

Ye M, Zyren J. Shore J. Forecasting crude oil spot price using OECD petroleum inventory levels. International Advances in Economic Research 2002; 8; 324–334.

Ye M, Zyren J. Shore J. A monthly crude oil spot price forecasting model using relative inventories. International Journal of Forecasting 2005; 21; 491–501.

Ye M, Zyren J. Shore J. Forecasting short-run crude oil price using high and low-inventory variables. Energy Policy 2006a; 34; 2736–2743.

Ye M, Zyren J. Shore J. Short-run crude oil price and surplus production capacity. International Advances in Economic Research 2006b; 12; 390–394.

Yokum J, Armstrong J. Beyond Accuracy: Comparison of Criteria Used to Select Forecasting Methods. International Journal of Forecasting 1995; 11; 591-97.

Yousefi S, Weinreich I. Reinarz D. Wavelet-based prediction of oil prices. Chaos, Solitons and Fractals 2005; 25; 265–275.

Yu L, Lai K.K., Wang S, He K. Oil price forecasting with an EMD-based multiscale neural network learning paradigm. Lecture Notes in Computer Science 2007; 4489; 925–932.

Yu L, Wang S, Lai K.K. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Economics 2008; 30; 2623–2635.

Zeng T, Swanson N.R. Predictive evaluation of econometric forecasting models in commodity futures markets. Studies in Nonlinear Dynamics and Econometrics 1998; 2; 1037-1037.

Zhou P, Ang B.W., Poh K.L. Measuring environmental performance under different environmental DEA technologies. Energy Economics 2008; 30; 1-14.

Zhu J. DEA/AR analysis of the 1988-1989 performance of the Nanjing Textile Cooperation, Annals of Operations Research 1996a; 66; 311-335.

Zhu J. Data envelopment analysis with preference structure. European Journal of Operational Research 1996b; 47; 136-150.

Zhu J. Super-efficiency and DEA Sensitivity Analysis. European Journal of Operational Research 2001; 129; 443-455.

Zhu J. Imprecise data envelopment analysis (IDEA): A review and improvement with an application. European Journal of Operational Research 2003a; 144; 513–529.

Zhu J. Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets. Kluwer Academic Publishers: Boston; 2003b.

Zhu J. Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets. Springer Science and Business Media Inc: New York; 2009.

Zhu J. A buyer-seller game model for selection and negotiation of purchasing bids: Extensions and New Models. European Journal of Operational Research 2004; 154; 150-156.

Zhu J, Cook W.D. Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis. Springer: Boston; 2007.

Chapter 4:
Performance Evaluation of Competing Forecasting Models:
A Multidimensional Framework based on MCDA

Performance Evaluation of Competing Forecasting Models:
A Multidimensional Framework based on MCDA

Abstract

The purpose of this paper is to fill a major gap in the field of forecasting; namely, the lack of a multidimensional framework for performance evaluation of competing forecasting models. Our framework is based on Multi-Criteria Decision Analysis (MCDA) methodology. In order to present and discuss how one might adapt such MCDA framework to address the problem of relative performance evaluation of competing forecasting models, we revisit MCDA methodology, on one hand, and survey the literature on performance criteria and their measures that are commonly used in evaluating and selecting forecasting models and propose a classification that will serve as a basis for the operationalisation of the MCDA framework, on the other hand. In sum, we propose a revised MCDA methodological framework that consists of a sequential decision making process with feedback adjustment mechanisms and provide guidelines as to how to operationalise it. Then, we adapt such a methodological framework to address the problem of performance evaluation of competing forecasting models. For illustration purposes, we have chosen the forecasting of crude oil prices as an application area.

Keywords: Forecasting, Performance Measurement, Performance Evaluation, Multi-Criteria Decision Analysis (MCDA), Crude Oil

4.1. Introduction

Nowadays, forecasts play a crucial role in our lives as individuals, organizations and societies; in fact, regardless of whether forecasting is performed implicitly or explicitly and whether the approach to forecasting is qualitative, quantitative, or hybrid, forecasts do drive our decisions and shape our future plans in a wide rage of 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 such as the forecasting object (e.g., time series, event outcome, event timing), the time dimensions of the forecasts (e.g., periodicity or frequency, forecasting horizon of interest, one-step vs. multi-step ahead forecasts, multi-step ahead forecast vs. multi-step ahead extrapolation forecasts) and the suitable way to state forecasts (e.g., point forecasts, interval forecasts, density forecasts, probability forecasts). However, regardless of how one defines the forecasting problem, a common issue faced by both academics and practitioners is related to the performance evaluation of competing forecasting models; to be more specific, although the performance evaluation exercise requires one to take account of several criteria at the same time, to the best of our knowledge, in the field of forecasting there is no published multidimensional framework designed for this purpose. Consequently, conflicting results about the performance of specific forecasting models are often reported in that some models perform better than others with respect to a specific criterion but worse with respect to other criteria; thus, leading to a situation where one cannot make an informed decision as to which model performs best overall; i.e., taking all criteria into account. It is the aim of this paper to fill this gap by proposing such a multidimensional framework; to be more specific, we design such multidimensional framework based on Multi-Criteria Decision Analysis (MCDA) methodology.

The remainder of this paper is organized as follows. In section 4.2, we revisit MCDA methodology in that we propose a revised methodological framework that consists of a sequential decision making process with feedback adjustment mechanisms and validation sub-processes and provide general guidelines as to how to operationalise it

regardless of the application area. In section 4.3, we propose a classification of performance criteria that are commonly used in evaluating and selecting forecasting models, which will serve as a basis for the operationalisation of the proposed MCDA methodological framework, and discuss their measures as well as methodological problems with respect to the performance evaluation of competing forecasting models. In section 4.4, we adapt the proposed MCDA methodological framework to address the problem of performance evaluation of competing forecasting models under several criteria and illustrate its use, on one hand, and test its performance, on the other hand, using the problem of forecasting crude oil prices as an application. Finally, section 4.5 concludes the paper.

4.2. MCDA – A Methodological Framework

Multi-Criteria Decision Analysis (MCDA) is a Management Science discipline concerned with decision making and analysis in the presence of multiple and often conflicting criteria. Although a substantial number of textbooks and papers have been published on MCDA, most of these publications focused on MCDA methods rather than its methodology; however, there are exceptions. For example, Roy (1985), Guitouni and Martel (1998), Guitouni et al. (2000), Belton and Stewart (2001), Bouyssou et al. (2006) and Tsoukias (2008) present and discuss the main methodological steps, but within the so-called decision aiding process; that is, a sequential decision making process but without any formal feedback mechanisms. Such a methodological framework not only suffers from the lack of formal feedback but also from the lack of relevant validation processes. Furthermore, most papers do not provide useful and much needed guidelines as to how to operationalise such a decision aid process. In this section, we make an attempt to overcome these limitations. In sum, we propose an adaptation of the operations research methodology proposed by Landry et al. (1983) to MCDA along with guidelines as to how to operationalise it in the form of questions and answers as well as classifications of MCDA approaches, methods and tools from which one could choose to address a specific application. To be more specific, the proposed methodological framework, as depicted in Figure 1, is a sequential decision making process with

feedback adjustment mechanisms along with validation sub-processes, where the managerial situation refers to a set of events related to the internal and/or the external environments of the organization that attracts stakeholders and Management attention to setup an agenda for investigation and analysis. As Oral and Kettani (1993) stated, a managerial situation could be a problem situation involving a gap in performance that need to be removed, or an assessment to position oneself vis-à-vis others, or a prediction to foresee likely opportunities and threats ahead, or an analysis to better understand the factors governing a system and its environment, etc. Thus, a managerial situation is more general than a problem situation.

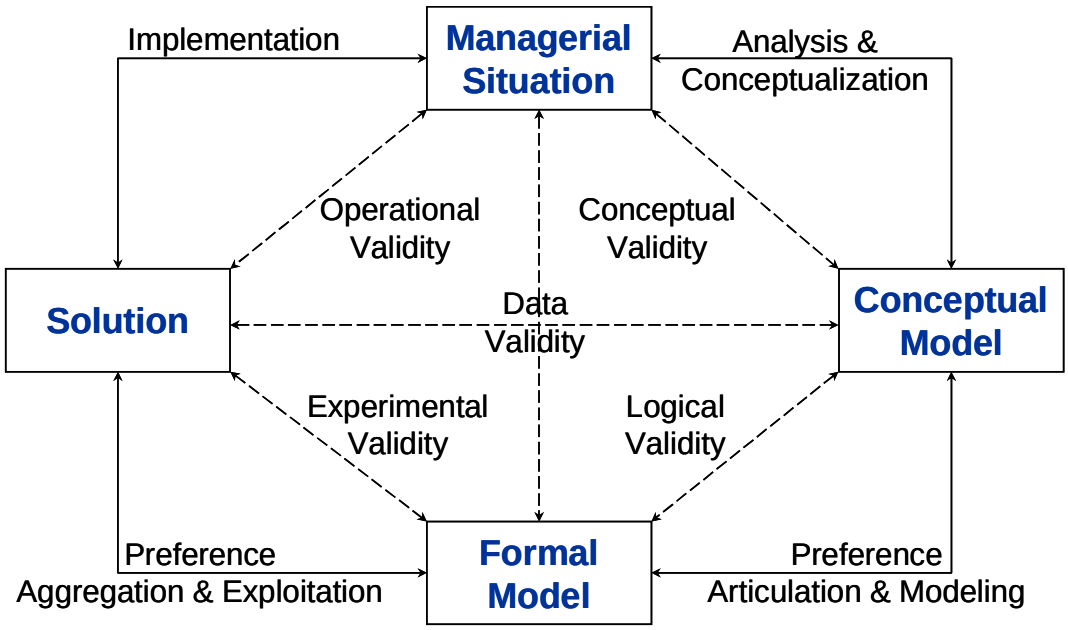


Figure 1: MCDA Methodological Framework

Once a managerial situation is acknowledged, the next stage involves its *conceptualization and analysis*, which would lead to a conceptual model; that is, a mental image of the managerial situation, which once translated in narrative terms would result in a problem definition. As such the conceptualization of the managerial situation is a sub-process that evolves gradually through time, as relevant information on the managerial situation is gathered and appropriate analysis is performed, and ends when

an acceptable level of conceptual validity is reached. Typical questions that one would need to address during the process of building a conceptual model are:

- Who are the main actors; i.e., stakeholders and decision makers? Note that a managerial situation or equivalently the conceptual model is shaped by the perceptions and behaviors of these actors.
- What criteria are used in assessing the managerial situation? What is their degree of conflict? Do they have a hierarchical relationship? What is their relative importance?
- What elements of the managerial situation and the organization environments to include in the analysis and those to exclude and their level of aggregation? What types of relationships exist between them?
- What caused the events characterizing the managerial situation?
- What are the goals and the targets decision makers want to achieve and their relative importance?
- What are the existing and the potential constraints one has or might have to deal with?
- What are the potential alternative courses of actions to address the managerial situation? How to generate alternatives; i.e., explicitly or implicitly? Is the set of alternatives static or dynamic?
- What MCDA problematic or decision analysis is appropriate to address the managerial situation; i.e., choice problematic, ranking problematic, sorting problematic, classification problematic, clustering problematic, or description problematic? where the *choice problematic* – also referred to as the selection or design problematic, is concerned with identifying the best alternative or selecting a subset of best alternatives, the *ranking problematic* is concerned with constructing a rank ordering of alternatives from best to worst, the *sorting problematic* is concerned with classifying or sorting alternatives into predefined and ordered homogenous groups, the *classification problematic* is concerned with classifying or sorting alternatives into predefined, but not ordered homogenous groups, the *clustering problematic* is concerned with classifying or

sorting alternatives into not predefined and not ordered homogenous groups, and the *description problematic* is concerned with identifying the major distinguishing features of alternatives and performing their description based on these features. For more details on these problematics, the reader is referred to Roy (1985, 1996) and Belton and Stewart (2001).

At this stage of conceptualization and analysis, several approaches and tools could be used to assists with addressing these questions, among others (see von Winterfeldt and Edwards, 2007; von Winterfeldt and Fasolo, 2009). For example, one could use *network analysis* and related analyses (e.g., *centrality analysis*, *link analysis*) to identify important individuals within the social network of the organization and to understand and/or discover relationships in such a network that would assist in building a problem definition team – for more details on network analysis, the reader is referred to Wasserman and Faust (1994), Carrington et al. (2005) and Knoke and Yang (2008). On the other hand, one could use an informal and unstructured approach such as *brainstorming* to generate ideas that would lead to a problem definition. Alternatively, one could use an informal and structured approach to brainstorming such as the *nominal group technique* (NGT) or one of its variants (e.g., improved NGT), and the *Delphi technique* or one of its variants (e.g., Argument Delphi) – for details on brainstorming, NGT and Delphi technique, the reader could refer to Proctor (2005), Mukherjee (2006), and Rocha (2007). Teams such as the problem definition team could use, within their brainstorming sessions, *cognition models*; e.g., *mind maps*, *concept maps* and *cognitive maps* (Eden, 1988, 1994; Eden et al., 1983, Buzan and Buzan, 1993; Kitchin and Freundschuh, 2000; Eden and Ackermann, 2004; Dalkir, 2005; Tague, 2005), *affinity diagrams* (Mizuno, 1988; Tague, 2005; Bauer et al., 2006), *storyboards* (Stamatis, 1997), *strategic choice* (Friend and Hickling, 1987; Friend and Jessop, 1969), *soft systems methodology* (Checkland, 1981), the *fact-net method* (Ramakrishna and Brightman, 1986) and the *purpose expansion method* (Nadler, 1981; Nadler and Hibino, 1998) to organize ideas and concepts and to assist with structuring the managerial situation and defining the problem, and tools such as *valued focused thinking* (Keeney, 1992) to structure criteria, attributes and objectives. In addition, the problem definition team

could use *tree structures* (e.g., work breakdown structure and its variants, bill of material, product breakdown structure, feature breakdown structure) and *tables or matrices* (e.g., responsibility assignment matrix, responsible-accountable-consulted-informed matrix and its variants) to understand or improve their understanding of structures (Harrison and Lock, 2004; Christensen et al., 2007; Hillson and Simon, 2007; Lewis, 2007). They could also use *flow charts* – also known as *process maps*, *activity diagrams*, *process decision program charts* (Ishikawa, 1985; Gitlow et al., 1995; Mitra, 1998; Alwan, 2000; Foster, 2001; Gryna and Juran, 2001; Evans and Lindsay, 1999; Tague, 2005), *SIPOC diagrams* (Ott et al., 2000; Harmon, 2007), *value stream maps*, *Gantt charts*, and *event chain diagrams* to understand or improve their understanding of processes and related activities and events (Christensen et al., 2007; Virine and Trumper, 2008). Tools such as cause-and-effect diagrams could be used to identify potential causes of a problem or what looks like a problem (Ishikawa, 1985; Gitlow et al., 1995; Mitra, 1998; Alwan, 2000; Foster, 2001; Gryna and Juran, 2001; Evans and Lindsay, 1999; Tague, 2005). Furthermore, one could use tools such as *gap analysis* to identify and analyze gaps (Brue, 2002). Once gaps are identified, one could then identify and analyze their causes and their consequences using formal analyses such as *root cause analysis* (Ishikawa, 1985; Tague, 2005) and *impact analysis* (Stamatis, 1997; Tague, 2005; Mukherjee, 2006). Other analyses such as *value analysis* could also be used to find out about sources of waste and identify improvement opportunities (Dhillon, 2002; Mukhopadhyaya, 2003; Younker, 2003). Finally, one could perform one or several other analyses to assist, for example, with making decisions as to whether to fix a problem or not. These analyses include, but are not restricted to, *strengths, weaknesses, opportunities, and threats (SWOT) analysis*, *cost-benefit analysis*, *benefit-effectiveness analysis*, *cost-effectiveness analysis*, *cost-volume-profit analysis*, *return on investment analysis*, *opportunity cost analysis*, and *risk analysis* (Simon et al., 1997; Williamson, 2004; Hillson and Simon, 2007). Finally, whatever quantitative data is available, its analysis using both informal tools (e.g., graphical representations) and formal tools (e.g. statistical tests) would help improving our understanding of the situation at hand and eventually suggests starting points for empirical investigation – for details on graphical

tools and statistical analyses, the reader is referred to Massey (1951), Anderson and Darling (1952, 1954), Siegel and Tukey (1960), Lewis (1961), Shapiro and Wilk (1965), Lilliefors (1967), Durbin and knott (1972), Dent (1974), Durbin et al. (1975), Mckellar (1982), D'Agostino and Stephens (1986), Dallal and Wilkinson (1986), Stephens (1986), Davis and Stephens (1989) and Csörgö and Faraway (1996).

Once the relevant actors have reached an agreement on the conceptual model, or equivalently the problem definition, the next stage starts and involves *preference articulation and modeling*, which would lead to a formal model; that is, a translation of the conceptual model using MCDA tools. Typical questions that one would need to address during the process of building a formal model are:

- What type(s) of scales to use for measuring criteria; i.e., nominal, ordinal, interval, ratio?
- What nature(s) of scales to use for measuring criteria; i.e., local scale or global scale? where a local (respectively, global) scale refers to one whose reference points are dependent on (respectively, independent of) the set of alternatives under consideration
- When preference articulation takes place; i.e., a priori, progressively, or a posteriori?
- What degree of compensation between criteria is acceptable; i.e., total, partial, or none? – See Fishburn (1976).
- What preference elucidation mode(s) the decision maker is comfortable with; i.e., absolute or direct rating, relative rating involving pairwise comparisons, tradeoffs, or lotteries? where tradeoffs and lotteries are indicated when alternatives have certain and uncertain outcomes, respectively.
- What type(s) of discriminating powers to use; i.e., true, quasi-, pseudo-, or interval criteria? and Whenever appropriate, should threshold(s) be constant or vary along the scale(s)? where a *true criterion* provides an absolute discriminating power; i.e., any difference matters regardless of its magnitude, a

quasi- or semi-criterion provides a non-absolute or nuanced discriminating power in that differences within a small range are not meaningful, which involves the use of a single threshold, and a *pseudo-criterion* provides a non-absolute discriminating power in that differences below a first threshold imply indifference, above a second threshold imply strict preference, and in between imply hesitation, and an interval criterion. Thresholds used within quasi- and pseudo-criteria are constant along the scale(s); when such thresholds are allowed to vary along the scale(s), *interval criteria* are obtained. For more details on types of criteria, the reader is referred to Roy (1985) and Vincke (1992).

- What type(s) of preference structure(s) to use; i.e., {P, I}, {P, Q, I} or {P, Q, I, R}? Is it necessary to work with valued preference relations; i.e., relations that reflect preference intensity? and if, yes, how to define them? where elementary preferences; namely, indifference (I), strong preference (P), weak preference (Q), and incomparability (R) are defined as follows: *indifference* corresponds to the existence of good reasons that justify equivalence between two actions, *strong preference* corresponds to the existence of good reasons that justify significant preference in favor of one of two actions, *weak preference* corresponds to the existence of good reasons that justify significant preference in favor of one of two actions, but that are insufficient to deduce either strict preference in favor of the other action or indifference between the two actions, thereby not allowing either of the two preceding situations to be distinguished as appropriate, and *incomparability* corresponds to an absence of good reasons that justify any of the three above mentioned relations. For more details on preference structures, the reader is referred to Roy (1977, 1985), Perny and Roy, (1992), Vincke (1992, 2001), Tsoukias and Vincke (1992, 1997), and Fishburn (1999), Ozturk et al. (2005).
- What type of order is required? Note that the choice of a specific preference structure will affect the choice of the type(s) of order(s) to choose from. In fact, when incomparability is not allowed, one might choose from the following

orders: complete order, weak order, semi-order, pseudo-order, or interval order, where a *complete order* – also referred as a total order, a simple order, or a linear order – is a strict simple order or ranking without ties, a *weak order* – also referred to as a total preorder or a complete preorder – is nothing but a ranking with possible ties, a *semi- or quasi-order* is a ranking that allows for ties and differences in performance less than a pre-specified threshold are not meaningful, a *pseudo-order* is a ranking that allows for ties and differences in performance less than a first threshold are not meaningful, above a second threshold are meaningful, and in between imply hesitation, and an *interval order* is an order similar to semi- and pseudo-orders except that the relevant thresholds do vary along the scale(s). These orders counterparts allowing for incomparability are referred to as partial orders. For more details on types of orders, the reader is referred to Roy and Vincke (1984, 1987), Fishburn (1985, 1997) and Tsoukias and Vincke (2003).

- What MCDA paradigm to adopt as a modeling framework for preferences; i.e., social choice vs. conjoint measurement? and Within the chosen paradigm, which models or methods to use? Recall that the *social choice paradigm* or approach to modeling preferences consists of transforming single-dimensional information on alternatives into a global preference using an aggregation procedure, commonly referred to as a multi-criteria aggregation procedure (MCAP), where the input to the MCAP is a profile or a preference table and its output is an order (e.g., a weak order) – for more details on the social choice paradigm, the reader is referred to Arrow (1963), Black (1958), Arrow and Raynaud (1986), Fishburn (1973) and Bouyssou et al. (2006). Note that one of the main differences between different MCAPs is their underlying aggregation principle(s). The aggregation principles most commonly used are: the majority principle proposed by Condorcet (1785) – also referred to as the Condorcet principle, the penalty principle proposed by de Borda (1784) – also referred to as the Borda principle, the central tendency principle, where global preference depends on a central tendency measure such as the mean (e.g., Kolmogoroff, 1930; Fodor and

Roubeans, 1995), the extreme value principle, where global preference depends on an extreme value such as a minimum or a maximum (e.g., Roberts, 1980; Bouyssou 1991, 1995; Pirlot, 1995; Fortemps and Pirlot, 2004) and the lexicographic ordering principle (e.g., Fishburn, 1974; 1980). On the other hand, the *conjoint measurement paradigm* or approach to modeling preferences consists of decomposing a global preference relation into elements related to the description of the alternatives on various dimensions, where the input to the MCAP is a global preference relation along with a set of dimensions and its output is a description of alternatives on various dimensions (e.g., marginal preferences or value functions) – for more details on the conjoint measurement paradigm, the reader is referred to Raiffa (1969, 1970), Edwards (1971), Keeney and Raiffa (1976).

At this stage of preference articulation and modeling, several approaches and tools could be used to assists with addressing these questions. Hereafter, we will present a classification of MCAPs for each paradigm so as to reflect the nature of their input and output as well as the underlying aggregation principle. Note that these classifications should be viewed as tools to operationalise our methodological framework in that they would assist decision makers in their choice of MCAPs.

Classification of Social Choice Paradigm-based Methods – The social choice approach-based MCAPs could be divided into two main categories; namely, procedures that aggregate several preference relations into one relation and procedures that aggregate a performance table into a single preference relation. The first category could be further divided into two sub-categories depending on whether the input (i.e., preference relations) is crisp or fuzzy. The main methods belonging to the first sub-category of MCAPs that aggregate several crisp preference relations into a single one are based on the majority principle (e.g., Condorcet or simple majority method – see Condorcet, 1785; May, 1952; Fishburn, 1977; weighted Condorcet method where weights are assigned to criteria – see Marchant, 2003; Bouyssou et al., 2006; qualified and absolute majority methods where threshold are used to define majority – see Fishburn, 1973), the

penalty principle (e.g., Borda method – see de Borda, 1784; Chamberlin and Courant 1983; Debord, 1992; Dummett, 1998; McLean and Urken 1995; Marchant, 1996, 1998, 2000, 2001; Nitzan and Rubinstein, 1981; Regenwetter and Grofman, 1998; and Van Newenhizen, 1992), the lexicographic ordering principle (e.g., lexicographic method and semi-lexicographic method; that is, a threshold-based lexicographic method – see Fishburn, 1974, 1980), or a hybrid principle (e.g., a two-stage method where the first stage is based on the majority principle and the second stage is based on the penalty principle referred to as a Borda Voting method or Adjusted Borda method – see Luce and Raiffa, 1957; Black, 1958). Note that the nature of the output of these methods; that is, a preference relation, is crisp. The main methods belonging to the second sub-category of MCAPs that aggregate several fuzzy preference relations into a single one could be further divided according to whether such output is crisp or fuzzy. The main methods leading to a crisp output are based on either the majority principle (e.g., generalized Condorcet method and other majorities – see Bouyssou et al., 2006) or the penalty principle (e.g., generalized Borda method where alternatives are ranked based on their sum of single-criterion scores, which represent the difference between the intensities of their out-performances and their under-performances – see Marchant, 1996). As to the methods leading to a fuzzy output, they typically are based on pairwise aggregation by means of an aggregation operator such as the arithmetic mean, the weighted arithmetic mean, the geometric mean, the median, the ordered weighted average, the min, the max or the leximin – for a detailed presentation of these methods and underlying operators, the reader is referred to and Fodor and Roubeans (1995), Fodor et al. (1995), Garcia-Lapresta and Llamazeres (2000), and Fortemps and Pirlot (2004). On the other hand, the second category of MCAPs; that is, the category of procedures that aggregate a performance table into a single preference relation, could be further divided into sub-categories depending on whether inter-criteria comparisons make sense or not. When inter-criteria comparisons make sense, one might refine the classification depending on whether criteria are measured on the same scale or not. When inter-criteria comparisons make sense and the criteria are measured on the same scale, the main methods are based on the extreme value principle (e.g., maximin method

and minimax method that rank alternatives in decreasing and increasing order of their minimum and maximum performance on all criteria, respectively – see Roberts, 1980; Bouyssou, 1991, 1995; Pirlot, 1995; Fortemps and Pirlot, 2004), the central tendency principle (e.g., weighted sum method where the ranking of alternatives is done in decreasing order of a weighted sum of performance on each criterion – see Kolmogoroff, 1930; Fodor and Roubeans 1995) and the lexicographic ordering principle (e.g., leximin and leximax methods where the lexicographic method is applied to a performance table with rows ordered in ascending and descending order, respectively – see Roberts, 1980; Fortemps and Pirlot, 2004) with the exception of ELECTRE I, which is an outranking procedure (see Pirlot, 1997). When inter-criteria comparisons make sense and the criteria are measured on different scales, the main methods are based on either the majority principle (e.g., ELECTRE Iv, ELECTRE IS, ELECTRE II, ELECTRE III, ELECTRE IV, ELECTRE TRI – see Roy, 1968, 1969, 1978, 1985, 1990, 1991, 1996; Roy and Bertier, 1973; Roy and Skalka, 1984; Roy et al., 1986; Hugonnard and Roy, 1982; Yu, 1992a, 1992b; Vincke, 1992; Figueira et al., 2005) or the penalty principle (e.g., PROMETHEE I, PROMETHEE II, PROMETHEE III, PROMETHEE IV, PROMETHEE V, PROMETHEE VI – see Brans, 1982, 1996; Brans et al., 1984, 1985, 1986; Brans and Vincke, 1985; Brans and Mareschal, 1992, 1994, 1995, 2005; Behzadian et al., 2010). Finally, when inter-criteria comparisons do not make sense, the main methods are adaptations of the ones belonging to the first category of MCAPs that aggregate several crisp preference relations into a single one.

Classification of Conjoint Measurement Paradigm-based Methods – The conjoint measurement approach-based MCAPs could be divided into two main categories; namely, value function methods and utility function methods, where value functions refer to preference representation functions under certainty and utility functions refer to preference representation functions under uncertainty. Value function methods could be divided into three sub-categories. The first sub-category consists of methods that first aggregate marginal preferences into a global preference, which is then used to compare alternatives, and includes additive value function models (see Raiffa, 1969; Edwards, 1971; Keeney and Raiffa, 1976), marginal preferences-based models (e.g., strict

decomposable models and non-strict decomposable models – see Krantz et al., 1971) and marginal traces-based models (e.g., generalized decomposable models and their variants – see Blackorby et al., 1978). The second sub-category consists of methods that first compare alternatives on each dimension and then aggregate, and consists of marginal traces on differences-based models (e.g., additive difference models, weak additive difference models, non-transitive additive conjoint measurement models, generalized additive conjoint measurement models – see Bouyssou, 1986; Fishburn, 1990a, 1990b, 1991a, 1992; Vind, 1991; Bouyssou and Pirlot, 2002b). The third and last sub-category consists of hybrid methods and consists of marginal traces- and traces on differences-based models (Bouyssou and Pirlot, 2004b). On the other hand, utility function methods are variants of Keeney and Raiffa's general multi-attribute utility function model (see Fishburn, 1970; Raiffa, 1970; Keeney and Raiffa, 1976). Note however that, with the exception of additive value function models, the remaining models are rather difficult to implement because of the lack of appropriate elicitation methods.

Once preference articulation and modeling related decisions have been made and the relevant actors have reached an agreement on the resulting formal model, the next stage starts and involves *preference aggregation and exploitation*; to be more specific, one would need to estimate the inputs to the chosen MCAP(s) and use them to devise a solution to or recommendations on how to address the managerial situation. Note that, as pointed out by Edwards and Barron (1994), the choice of preference modeling approaches or models and the choice of elicitation methods for such models involves a tradeoff between errors due to modeling choices and those due to elicitation choices. Typical questions that one would need to address during the process of devising a solution are:

- How to estimate the inputs of social choice paradigm-based methods? or, equivalently, What methods to use for eliciting preference relations – also referred to as scores – and parameters such as weights and thresholds?

- How to estimate the inputs of conjoint measurement paradigm-based methods? or, equivalently, What methods to use for eliciting preference functions – also referred to as scores – and parameters such as weights?

Several methods have been proposed to elicit scores and estimate parameters. With respect to social choice methods, one would typically use a questioning procedure to elicit preference relations (Bana e Costa and Vansnick, 1999). As to the methods for eliciting weights – also referred to as relative importance coefficients (e.g., Stewart, 1992; Weber and Borcherding, 1993), they are divided into five categories; namely, *rating methods* (e.g., absolute rating methods such as direct rating – see von Winterfeldt and Edwards, 1986; and relative rating methods such as Max100, Min10, and Point Allocation – see Doyle et al., 1997; Bottomley et al., 2000; Bottomley and Doyle, 2001), *ranking methods* (e.g., simple ranking where weights are positions – see Pöyhönen and Hämäläinen, 1998, 2001; and weighted ranking methods such as the pack of cards technique, also referred to as the Simos' procedure, and the revised Simos' procedure – see Simos, 1990; Figueira and Roy, 2002), *Mousseau System* (e.g., Mousseau, 1995), *resistance to change grid* (e.g., Rogers and Bruen, 1998a; 1998b), and *examples-based inference methods* (e.g., Mousseau and Slowinski, 1998; Mousseau, Figueira and Naux, 2001). Finally, methods for eliciting thresholds include *mathematical equations-based models* (e.g., Roy, Present and Silhol, 1986; Bouyssou, 1990) and *examples-based inference methods* (e.g., Mousseau and Slowinski, 1998; Ngo The and Mousseau, 2002; Dias and Mousseau, 2006). On the other hand, within a conjoint measurement paradigm, elicitation methods of preference functions are divided into two categories depending on whether the preference function is a value function or a utility function. Elicitation methods of value functions could be divided into two categories; namely, *direct assessment methods* of partial or marginal value functions (e.g., direct rating methods and ratio estimation methods – see von Winterfeldt and Edwards, 1986; category estimation methods, which are concerned with constructing a qualitative value scale – see Torgerson, 1958), and *indirect assessment methods* of marginal value functions (e.g., bisection methods – see von Winterfeldt and Edwards, 1986; difference methods such as the standard sequence method – see von Winterfeldt and Edwards, 1986; regression-

based methods and their extensions such as UTA – see Jacquet-Lagreze and Siskos, 1982, 2001; UTADIS – see Jacquet-Lagreze, 1995; Doumpos and Zopounidis, 2002; and MHDIS – see Zopounidis and Doumpos, 2000). As to the elicitation of utility functions, several methods have been proposed including the variable probability method – also referred to as the basic reference lottery ticket (Raiffa, 1968; von Winterfeldt and Edwards, 1986) and the variable certainty equivalent method (Keeney and Raiffa, 1976; von Winterfeldt and Edwards, 1986). Finally, for eliciting weights – also referred to as inter-criterion information, several methods could be used. These methods could be divided into five categories; namely, *rating methods* (e.g., absolute rating methods such as direct rating – see von Winterfeldt and Edwards, 1986; relative rating methods, often referred to as ratio estimation methods such as Max100, Min10, point allocation, swing weighting – see von Winterfeldt and Edwards, 1986; Doyle et al., 1997; Bottomley et al., 2000; Bottomley and Doyle, 2001), *ranking methods* (e.g., simple ranking, where weights are positions – see Pöyhönen and Hämäläinen, 1998, 2001; and weighted ranking methods such as the extended pack of cards technique – see Pictet and Bollinger, 2008), *compensation methods* (e.g., tradeoff weighting – see Keeney and Raiffa, 1976), and *hybrid methods* (e.g., rank order centroid weights method, which is a two-stage ranking and rating method – see Solymosi and Dombi, 1986; Olson and Dorai, 1992; Edwards and Barron, 1994).

Once preference aggregation and exploitation related decisions have been made and related tasks have been performed at the satisfaction of the relevant actors; that is, relevant actors have reached an agreement on the resulting solution or recommendations, the next stage is concerned with the implementation of such solution or recommendations.

Finally, we would like to draw the reader's attention to the fact that the above described methodological framework not only includes formal feedback adjustment mechanisms, but is equipped with validation sub-processes to enhance the likelihood that the final solution would represent a consensus or, equivalently, be valid from all parties involved perspectives. For reasons of space, we refer the reader to Landry et al. (1983) for a

discussion of validation sub-processes. In this paper, we adapt the proposed MCDA methodological framework to address the problem of performance evaluation of competing forecasting models – see section 4.4; however, before presenting such adaptation, we first discuss the problem of performance evaluation of forecasting models in the next section.

4.3. Performance Evaluation of Forecasting Models

In this section, we propose a classification of performance criteria and discuss their measures along with the current practice in terms of performance evaluation of competing forecasting models and related methodological issues. Our literature survey reveals that performance evaluation of forecasting models or methods could be based on six criteria; namely, reliability, costs, benefits, complexity, universality, and ability to incorporate managerial judgment, where reliability is broken down into five sub-criteria; that is, theoretical relevance, validity, accuracy, informational efficiency, and degree of uncertainty of forecasts, and accuracy is in tern broken down into three sub-criteria; that is, goodness-of-fit, biasedness, and correct sign – see Table 1 for definitions.

Obviously, the relative importance of these criteria and their sub-criteria depends on the application context and the individuals involved such as decision makers and analysts or researchers. In business environments, all these criteria are in general taken into account; however, in academia reliability seems to be privileged and accuracy seems to be a predominant sub-criterion as reflected by the number of papers reporting on these criteria in their evaluation of forecasts. Most of these criteria could be measured on both continuous and discrete scales. For suggestions on to how to measure the above mentioned criteria, the reader is referred to Table 2. For additional information on performance criteria and their relative importance in practice, the reader could refer to Dalrymple (1975; 1987), Carbone and Armstrong (1982), Mentzer and Cox (1984a, 1984b), Mahmoud (1984), Armstrong (1985), Mahmoud et al. (1986), Sanders and Manrodt (1994), Mentzer and Kahn (1995), Yokum and Armstron (1995), Armstrong (2001a, 2001b), Armstrong et al. (2001) and McCarthy et al. (2006).

Criteria and Sub-Criteria	Definition
Reliability	Multidimensional construct, where its dimensions reflect the extent to which the model is theoretically relevant and valid, produces accurate forecasts, incorporates all information within data, and is highly likely to produce results that are not too dispersed around a central tendency measure of forecasts
Theoretical Relevance	This criterion is concerned with the built-in design features of a forecasting model and includes elements such as extent to which the model is by design able to predict patterns and turning points, if any; able to adapt to new conditions; e.g., structural change, if any; able to account for interventions, if any; and suitable for the specific forecasting horizon under consideration
Validity	This criterion is concerned with whether the assumptions underlying the model hold or not, or the extent to which they hold
Accuracy	Multidimensional construct, where its dimensions reflect the extent or degree to which the model has a good goodness-of-fit, is unbiased, and is able to predict the correct sign
Goodness-of-fit	This sub-criterion is concerned with the extent or degree to which the model produces forecasts that are close to the actual observations
Biasedness	This sub-criterion is concerned with the extent to which the model produces forecasts that are unbiased; that is, the model does not tend to systematically over- or under-estimate the forecasts
Correct Sign	This sub-criterion is concerned with the extent to which the model is able to produce forecasts that are consistent with actuals in that forecasts reveal increase (resp. decrease) in value when actuals increase (resp. decrease) in value
Informational Efficiency	This criterion is concerned with whether the model is able to capture all elements of information within the data or not
Degree of Uncertainty of Output/Forecasts	This criterion is concerned with the extent to which the model is highly likely to produce results that are not too dispersed around a central tendency measure of forecasts
Costs	This criterion is concerned with relevant costs; e.g., cost of developing or purchasing and maintaining a forecasting system, cost of purchasing or collecting data and its storage and pre-processing, costs related to the time required to obtain forecasts, costs of repeated applications or use of a model to produce forecasts, etc.
Benefits	This criterion is concerned with improvements resulting from improved decisions such as direct and indirect cost savings and higher service levels
Complexity	This criterion is concerned with the extent to which the model or method is easy to understand by users/managers; extent to which it is easy to interpret results; level of conceptual and technical knowledge or expertise required for an effective use of the model, etc.
Universality	This criterion is concerned with the extent to which the model is widely used in practice or relevant users are familiar with; extent to which the model is available within popular software packages, etc.
Ability to Incorporate Managerial Judgement	This criterion is concerned with the extent to which the model or method provides for means to incorporate subjective judgement

Table 1: Classification and Definitions of Criteria and Sub-Criteria used in Forecasting

Criteria and Sub-Criteria	Measures
Reliability	Reliability could be measured in many different ways depending on the choice of the measurement scale and the choice of a relevant performance evaluation framework. In fact, reliability could be measured on a discrete scale by a rank determined by a multi-criteria decision analysis method such as an outranking method. Alternatively, reliability could be measured on a continuous scale by an index determined by a relevant data envelopment analysis model.
Theoretical Relevance	Theoretical relevance is a criterion that could be measured on both continuous and discrete scales; to be more specific, one could use a checklist or multiple choice-multiple response questionnaire to gather information on this criterion where the checklist would provide the researcher with desirable features of a forecasting model measured, for example, by categorical variables each either reflecting whether the model possesses a feature or not and coded by 1 and 0, or reflecting the magnitude of an element of theoretical relevance, say low, medium and high, and coded by, say 1, 2 and 3; e.g., ability to predict patterns such as trend, seasonality and cycles, ability to predict turning points, ability to adapt to new conditions such as structural change, ability to account for interventions, suitability for the specific forecasting horizon under consideration, etc. The theoretical relevance criterion would then be measured, on a continuous scale, by a weighted combination of scores on individual elements of theoretical relevance. On a discrete scale, this criterion could be measured by a score that represent the number of elements of theoretical relevance that the model possesses. Note that one could choose to measure this criterion in an aggregate fashion and use a single categorical variable to measure, on a discrete scale, the overall theoretical relevance of a model as perceived by managers or users.
Validity	As models are simplifications of reality, they are based on simplifying assumptions. In order for a model to be valid and be used with confidence, one would need to check whether the underlying assumptions hold or not and, sometimes, to what extent. One could use a checklist to gather information on this criterion where the checklist would provide the researcher with the relevant assumptions underlying the model. For each relevant assumption, such a checklist could also provide a list of tools that could be used to check whether each assumption holds or not. Thus, a potential measure of this criterion is a score that represent the number of assumptions that hold expressed in percentage. When a list of tools; e.g., statistical tests, to check whether an assumption holds or not is provided, we suggest to compute the proportion of tests that support each assumption and use a weighted combination of these proportions as a measure of validity, where weights are chosen so as to normalize such proportions.
Accuracy	Accuracy could be measured in the same way reliability is, as it is a multidimensional construct
Goodness-of-fit	For continuous variables; e.g., price of a strategic commodity, the goodness-of-fit of a forecasting model is generally measured by an *absolute measure* such as a central tendency statistic; e.g., mean or median, of absolute residuals or errors (AEs), squared errors (SEs), mixed errors (MEs), or trimmed errors (TEs), where MEs are defined in such a way that either positive or negative errors are more heavily penalized and TEs are obtained by replacing values above and/or below given

	threshold(s) by such threshold(s); such measures would be data scale-dependent and would only be appropriate to use in comparing models using the same data set or different data sets with similar characteristics. Alternatively, one could compute such statistics for absolute percentage residuals or errors (APEs), squared percentage errors (SPEs), modified percentage errors (MPEs), scaled AEs and APEs (ScAEs and ScAPEs), scaled SEs and SPEs (ScSEs and ScSPEs), relative AEs (RAEs), relative SEs (RSEs), trimmed AEs (TAEs), and trimmed SEs (TSEs), where scaled errors refer to errors adjusted for the level or the volatility of the series (TrdSc and VolSc) and relative errors refer to ratios of errors resulting from the use of the specific model under consideration and those resulting from the use of a benchmark model. Finally, one could use relative measures such as ratios of the above mentioned absolute measures or percent better (PB) of these errors to measure the goodness-of-fit of a model, where PB refers to the proportion of times a model produces better forecasts as compared to a competing model. As for discrete or categorical variables, the hit ratio; i.e., the percentage of cases correctly classified, is generally used to measure the goodness-of-fit of a forecasting model of categorical variables; e.g., logit.
Biasedness	On a continuous scale, central tendency statistics; e.g., mean or median, of residuals or errors produced by a forecasting model are typically used to measure its degree and its type of biasedness, where the degree of biasedness is reflected by the magnitude of the statistic and the type of biasedness is reflected by the sign of the statistic; i.e., a negative (respectively, positive) sign reveals that the model tends to systematically over- (respectively, under-) estimate the forecasts. On a discrete scale, one could use a categorical variable to measure this criterion, where the categories could be defined so as to reflect whether the model produces biased forecasts or not – coded by 1 and 0 respectively – using, for example, statistical tests such as the standard sign test, the Wilcoxon signed-ranks test, and a regression-based test where the actual observations of the variable of interest are regressed on their forecasts and the models includes an intercept coefficient that should not be statistically significantly different from zero if the forecasts are to be unbiased.
Correct Sign	On a continuous scale, the correct sign criterion could be measured by the percentage of the correct sign predictions (PCSP) or the percentage of the correct direction change predictions (PCDCP) or both, where both PCSP and PCDCP are equal to $\sum_{t=1}^{n} z_t \Big/ n$, but $z_t = \begin{cases} 1 & \text{If } Y_t \cdot \hat{Y}_t \succ 0 \\ 0 & \text{Otherwise} \end{cases}$ for PCSP and $z_t = \begin{cases} 1 & \text{If } (Y_t - Y_{t-1}) \cdot (\hat{Y}_t - Y_{t-1}) \succ 0 \\ 0 & \text{Otherwise} \end{cases}$ for PCDCP. On a discrete scale, one could use a categorical variable to measure this criterion, where the categories could be defined so as to reflect whether the model produces forecasts that are consistent with actual observations sign-wise using, for example, statistical tests such as the chi-square test of independence based on a confusion matrix of correct sign predictions.
Informational Efficiency	Typically, informational efficiency is tested for using a statistical test rather than measured. However, one could measure it on a discrete scale using a categorical variable where, for example, one could specify two categories; namely, model produces informationally efficient forecasts, and model does not produce informationally efficient forecasts, and coded

	by 1 and 0 respectively. One could also use more than two categories if several tests are used and one is interested in taking account of the number of tests that support the null hypothesis of informational efficiency. In general, efficient forecasts should be uncorrelated with the information available at the time the forecasts are made – otherwise one would eventually be able to exploit such correlations to improve the forecasts; therefore, one could either estimate the regression model $Y_t = \beta_0 + \beta_1 \hat{Y}_t + \varepsilon_t$ and conclude that the forecasting model or method is informationally efficient if $\beta_1 = 1$, or test whether forecast errors e_ts are uncorrelated with actual observations of the variable of interest as well as with each of the explanatory variables, if any.
Degree of Uncertainty of Output/Forecasts	This criterion could be measured on both continuous and discrete scales. In fact, one could use the length of the prediction interval; that is, the confidence interval of the forecast, to measure this criterion on a continuous scale, or use a categorical variable to measure this criterion on a discrete scale where the categories could, for example, be low, medium and high degrees of uncertainty and are specified by the relevant decision makers.
Costs	On a continuous scale, one could use the amount of money expressed in an appropriate currency as a measure of this criterion. Alternatively, one could use a categorical variable to measure this criterion on a discrete scale, where the categories could, for example, be low, medium and high cost, and are specified by the relevant decision makers.
Benefits	This criterion could be measured on both continuous and discrete scales. For example, when realistic estimates expressed in currency could be obtained for each benefit, one could use the total amount of savings as a continuous measure of this criterion. However, in practice, one might be uncomfortable with assigning a figure to a benefit, in which case a categorical variable could be used to measure this criterion where the categories could, for example, be low, medium, and high.
Complexity	This criterion could be measured on both continuous and discrete scales in the same way the theoretical relevance criterion is, where the checklist would include element such as extent to which the model or method is easy to understand by users/managers; extent to which it is easy to interpret results; level of conceptual and technical knowledge or expertise required for an effective use of the model, etc.
Universality	This criterion could be measured on both continuous and discrete scales in the same way the theoretical relevance criterion is, where the checklist would include element such as extent to which the model is widely used in practice; extent to which relevant users are familiar with the model; extent to which the model is available within popular or available software packages, etc.
Ability to Incorporate Managerial Judgement	This criterion could be measured on both continuous and discrete scales in the same way the theoretical relevance criterion is, where the check list would include elements such as the number of parameters of the model that could be used to incorporate managerial judgement, the flexibility with which the values of such model parameters could be chosen or adjusted to incorporate managerial judgement, etc.

Table 2: Measures of Criteria and Sub-Criteria used in Forecasting

Regardless of how one defines a forecasting problem, a common issue faced by both academics and practitioners is related to the performance evaluation of competing forecasting models; to be more specific, although the performance evaluation exercise requires one to take account of several criteria at the same time, to the best of our knowledge, in the field of forecasting there is no published multidimensional framework designed for this purpose. Consequently, conflicting results about the performance of specific forecasting models are often reported in that some models perform better than others with respect to a specific criterion but worse with respect to other criteria; thus, leading to a situation where one cannot make an informed decision as to which model performs best overall; i.e., taking all criteria into account. In fact, although several performance criteria and measures are used in most papers, the assessment exercise of competing forecasting models is generally restricted to their ranking by measure; thus, the current methodology is unidimensional in nature. Consequently, one may obtain different rankings of models for different measures leading to inconsistent and often confusing results both within and across studies – see Table 5. In this paper, we intend to contribute, from a methodological perspective, to the filed of forecasting by proposing a multidimensional framework for the performance evaluation of competing forecasting models. We design such a multidimensional framework based on the Multi-Criteria Decision Analysis (MCDA) methodology presented in the previous section. In sum, the adaptation of such MCDA methodology to the performance evaluation of competing forecasting models is presented in the next section.

4.4. Adaptation of MCDA Methodology to Performance Evaluation of Competing Forecasting Models and Its Application

In this section, we propose an adaptation of the MCDA methodological framework proposed in section 4.2 to address the problem of performance evaluation of competing forecasting models under several criteria. For illustration purposes, we have chosen as an application the problem of evaluating the performance of competing models for forecasting crude oil prices.

In this paper, the managerial situation is not much of a concern as it is well understood and structured in the form of a conceptual model that is translated in narrative terms in a problem definition, where the problem is concerned with the design of a multidimensional framework for performance evaluation of competing forecasting models. Thus, with reference to the modeling-validation tetrahedron proposed by Oral and Kettani (1993), this application is concerned with its theoretical facet. However, from an operational perspective, some questions need to be addressed and some decisions need to be made to obtain a problem definition that is refined enough to work with. With reference to the MCDA methodological framework presented in section 4.2, the relevant decisions to make or questions to address along with their answers could be summarized as follows. The main actors are the authors of this paper; that is, academics, and as such we deliberately restrict ourselves to only consider the reliability criterion and its sub-criteria, because of the lack of data on the remaining criteria or their irrelevance in the academic context (see Table 1). Note that, in our application, theoretical relevance will not be taken into account to avoid penalizing models that are not necessarily theoretically relevant, but do a good job in forecasting crude oil prices. In addition, as most statistical and econometric software packages at our disposal do not provide prediction intervals, the degree of uncertainty of forecasts will not be considered. Furthermore, our empirical results revealed that all the models attempted in this study – including the valid ones, are not informationally efficient; therefore, as informational efficiency does not discriminate between models, it will be discarded. As to the set of potential alternative courses of action, in our forecasting application such alternatives are competing models for forecasting crude oil prices. As a strategic commodity, crude oil has attracted the attention of many individuals including investors, analysts, and academic researchers. The literature relevant to this paper could be divided into two main categories; namely, data modeling and forecasting. Although our main interest is in the forecasting area, we would like to draw the reader's attention to the fact that research concerned with data modeling provides valuable input to model building in forecasting. In sum, research in the area of data modeling is generally concerned with one of the following three questions:

Question 1: Are crude oil markets consistent with the market efficiency hypothesis? Most academic contributions address this question in a direct fashion using statistical analyses such as regression analysis (e.g., Green and Mork, 1991; Moosa and AI-Loughani, 1994) and cointegration analysis (e.g., Crowder and Hamid, 1993; Schwartz and Szakmary, 1994; Gulen, 1999; Bekiros and Diks, 2008), and econophysics analysis such as fractal analysis (e.g., Alvarez-Ramirez et al., 2002). Other contributions however address the question indirectly using, for example, artificial intelligence methods such as artificial neural networks (e.g., Shambora and Rossiter, 2007). The results of these studies tend to be mixed in that some support the market efficiency hypothesis whereas others shed doubt on it; but in general conclusions seem to depend on the frequency of the data used.

Question 2: Are there any relationships between oil prices and economic variables? Most academic contributions address this question using statistical analyses such as factor analysis (e.g., Faff and Brailsford, 1999; Sadorsky, 2001; El-Sharif et al., 2005), cointegration analysis (e.g., Amano and Norden, 1998; Ciner, 2002; Rautava, 2004; Lardic and Mignon, 2006; Panagiotidis and Rutledge, 2007), vector autoregressive analysis (e.g., Darby, 1982; Hamilton, 1983, 1985; Gisser and Goodwin, 1986; Huang et al., 1996; Ferderer, 1996; Sadorsky, 1999; Papapetrou, 2001; Rautava, 2004; Kilian, 2008a; 2008b; 2009; Kilian et al., 2009; Kilian and Park, forthcoming), and threshold analysis (e.g., Loungani, 1986; Mork, 1989; Lee et al.,1995; Hamilton, 1996; Cunado and Perez de Gracia, 2003). The results of studies addressing this question also tend to be mixed, but in general conclusions seem to depend on the period of study.

Question 3: Are there any relationships between oil prices and prices of petroleum products? Most academic contributions address this question using statistical analyses such as cointegration analysis (e.g., Serletis, 1994; Hansen and Johansen, 1999; Siliverstovs et al., 2005; Bachmeier and Griffin, 2006; Asche et al., 2006), vector autoregressive analysis (e.g., Adrangi et al., 2001; Asche et al., 2003), and threshold analysis (e.g., Bacon, 1991; Godby et al., 2000; Galeotti et al., 2003; Kaufmann and Laskowski, 2005; Radchenko, 2005; Al-Gudhea et al., 2007). In general, most studies

seem to suggest that there is enough empirical evidence to support the existence of long-term relationships between crude oil prices and prices of petroleum products.

On the other hand, research on forecasting crude oil addresses several crude oil related variables such as prices, returns, supply, and demand. As far as prices and returns are concerned, quantitative forecasting models could be divided into three main categories; namely, non-artificial intelligence models, artificial intelligence models, and hybrid models. Non-artificial intelligence models include time series models; e.g., *random walk (RW) models* (Zeng and Swanson, 1998; Kaboudan, 2001; Lalonde et al., 2003; Abosedra, 2005; Knetsch, 2007; Coppola, 2008) and *autoregressive integrated moving average* (*ARIMA) models* (Sequeira and McAleer, 2000; Lalonde et al., 2003; Fernandez, 2006; Xie et al., 2006; Moshiri and Foroutan, 2006), and explanatory models; e.g., *linear regression models* (Bopp and Lady, 1991; Ye et al., 2002, 2005, 2006a, 2006b; Sequeira and McAleer, 2000; Abosedra and Baghestani, 2004; Knetsch, 2007), *vector autoregressive (VAR) models* (Zeng and Swanson, 1998), *error correction (EC) and vector error correction (VEC) models* (Zeng and Swanson, 1998; Sequeira and McAleer, 2000; Longo et al., 2007; Coppola, 2008), *wavelet transform-based models* (Yousefi et al., 2005), and *state space models* (Pindyck, 1999; Bernard et al., 2004). Note that most of the variables used to build explanatory models are suggested by data modeling studies. As to artificial intelligence models, they include artificial neural networks (e.g., Kaboudan, 2001; Mirmirani and Li, 2004; Fernandez, 2006; Suriya, 2006; Xie et al., 2006; Yu et al., 2007; Fan et al., 2008; Yu et al., 2008; Ghaffari and Zare, 2009), genetic programming-based models (e.g., Kaboudan, 2001; Mirmirani and Li, 2004; Matilla-Garcia, 2007; Fan et al., 2008), pattern recognition-based models (e.g., Fernandez, 2006; Xie et al., 2006; Fan et al., 2008), and belief network-based methods (e.g., Abramson and Finizza, 1991; Abramson, 1994). Finally, the integration of non-artificial intelligence and artificial intelligence models has let to what is referred to as hybrid models (e.g., Abramson and Finizza, 1995; Wang et al., 2004, 2005; Fan et al., 2008; Yu et al., 2008).

In this paper, the set of potential alternative courses of action; that is, the set of competing forecasting models, is chosen as a subset of the non-artificial intelligence models proposed in the literature; namely, *RW models* (Zeng and Swanson, 1998; Kaboudan, 2001; Lalonde et al., 2003; Abosedra, 2005; Knetsch, 2007; Coppola, 2008), *ARIMA models* (Sequeira and McAleer, 2000; Lalonde et al., 2003; Fernandez, 2006; Xie et al., 2006; Moshiri and Foroutan, 2006), *linear regression models* (Bopp and Lady, 1991; Ye et al., 2002, 2005, 2006a, 2006b; Sequeira and McAleer, 2000; Abosedra and Baghestani, 2004; Knetsch, 2007), *VAR models* (Zeng and Swanson, 1998), and *EC and VEC models* (Zeng and Swanson, 1998; Sequeira and McAleer, 2000; Longo et al., 2007; Coppola, 2008). In addition, among these models only valid ones are considered, where the validity of a model refers to the fact that the underlying assumptions hold over the chosen period of study; namely, January 1994 to August 2007. Note that there are several reasons for choosing this specific period; namely, change in International Energy Agency data collection methodology in 1990, non-availability of data on all relevant variables before January 1994, and recent credit crunch. Note also that some models have been discarded because they are dominated by one or several models on all criteria under consideration or because of the unavailability of data on the explanatory variables they use.

In sum, Table 3 summarizes the remaining ten forecasting models that are valid for at least one implementation method including the Holt-Winter's model with multiplicative seasonality, which was not reported in the literature on forecasting crude oil prices, as another benchmark along with random walk adjusted for trend – for details on these models and definitions of the explanatory variables used in linear regression models, the reader is referred to the original papers as referenced in the table. Notice that, the fourth implementation method; that is, out-of-sample implementation with rolling origin and fixed window, results in all models being valid and is the one that we use for measuring the performance on these models on each criterion under consideration – for a review on implementation methods, the reader is referred to Tashman (2000). The performance measures of the models summarized in Table 3 are reported in Table 4, where PSTSU denotes the proportion of statistical tests supporting unbiasedness.

Valid Forecasting Models	Implementation Methods				
	1	2	3	4	5
1. RW with Trend (Zeng and Swanson, 1998)	X	√	√	√	X
2. Holt-Winter Exponential Smoothing with Multiplicative Seasonality (HWESMS)	X	X	√	√	√
3. ARIMA (1,1,1)	√	√	√	√	√
4. ARIMA (1,1,1) (1,0,1)	√	√	√	√	√
5. REG1: $WTI_t = \beta_0 + \beta_1 WTI_{t-1} + \sum_{i=0}^{3} \beta_{i+2} RIN_{t-i} + \sum_{j=0}^{5} \beta_{j+6} D_j 911 + \beta_{12} APR99 + \varepsilon_t$ (Ye et al., 2005)	X	X	X	√	X
6. REG2: $WTI_t = \beta_0 + \beta_1 WTI_{t-1} + \beta_2 OECD\_Stocks_{t-1} + \beta_3 ANN_{t-1} + \beta_4 T + \sum_{j=0}^{5} \beta_{j+5} D_j 911 + \beta_{12} LAPR99 + \varepsilon_t$ (Ye et al., 2005)	X	X	X	√	X
7. REG3: $WTI_t = \beta_0 + \beta_1 AR(1) + \beta_2 AR(12) + \sum_{j=0}^{5} \beta_{j+3} D_j 911 + \beta_9 LAPR99 + \varepsilon_t$ (Ye et al., 2005)	X	X	X	√	√
8. REG4: $WTI_t = \beta_0 + \beta_1 WTI_{t-1} + \sum_{i=0}^{3} \beta_{i+2} RIN_{t-i} + \sum_{i=0}^{3} (\gamma_i LIN_{t-i} + \gamma_i^* LIN2_{t-i}) + \sum_{i=0}^{3} (\gamma_{i+4} HIN_{t-i} + \gamma_{i+4}^* HIN2_{t-i}) + \sum_{j=1}^{6} \beta_{j+9} D_j 911 + \beta_{16} LAPR99 + \varepsilon_t$ (Ye et al., 2006a)	X	√	√	√	X
9. REG5: $WTI_t = \beta_0 + \beta_1 WTI_{t-1} + \sum_{i=0}^{1} \beta_{i+2} \Delta RIN_{t-i} + \beta_4 \log(Excess\_cap\_OPEC) + \sum_{j=1}^{6} \beta_{j+4} D_j 911 + \beta_{11} LAPR99 + + \varepsilon_t$ (Ye et al., 2006b)	X	X	X	√	√
10. REG6: $WTI_t = \beta_0 + \beta_1 WTI_{t-1} + \sum_{i=0}^{4} \beta_{i+2} \Delta RIN_{t-i} + \beta_7 \Delta RIN_{t-5} + \sum_{j=0}^{4} \beta_{j+8} \Delta LIN_{t-i} + \beta_{13} LIN_{t-5} + \beta_{14} AIN_t + \varepsilon_t$ (Ye et al., 2002)	X	X	X	√	√

[1]In-Sample Implementation; [2]Out-of-sample with fixed origin & static forecast; [3]Out-of-sample with fixed origin & dynamic forecast; [4]Out-of-sample with rolling origin & fixed window; [5]Out-of-sample with rolling origin & variable window; [X]Invalid Model; [√]Valid Model

Table 3: Valid Forecasting Models for At Least One Implementation Method

Note that the comparison of forecasting models with respect to biasedness as measured by ME, MPE, MAdjPE, MTrdScE, or MVolScE depends on the decision making situation; therefore, we have chosen to use an aggregate measure of the results of statistical tests of biasedness; namely, the proportion of tests supporting unbiasedness, PSTSU, instead of the above mentioned measures. Last, but not least, the MCDA problematics most relevant to our application are the choice problematic and the ranking problematic. In this paper, we focus on the ranking problematic.

	DMU No.	1	2	3	4	5	6	7	8	9	10
	Measures & Tests	RW With Trend	HWESMS	ARIMA (111)	ARIMA (111)(101)	REG1	REG2	REG3	REG4	REG5	REG6
Biasedness	ME	-0.722	0.249	0.470	0.451	0.359	0.464	0.649	0.403	0.202	0.247
	MPE	-1.318	0.209	0.828	0.779	0.917	0.749	1.403	0.997	0.510	0.460
	MTrdScE	-1.000	0.345	0.652	0.625	0.497	0.643	0.900	0.558	0.279	0.342
	MVolScE	-0.003	0.001	0.002	0.002	0.001	0.002	0.003	0.002	0.001	0.001
	MAdjPE	-1.046	0.447	1.103	1.055	1.158	1.042	1.697	1.248	0.740	0.706
Goodness-of-Fit	MSE	12.900	11.433	13.052	13.273	11.545	13.810	13.920	11.744	11.140	11.821
	MSPE	54.025	48.389	55.238	55.566	48.145	59.226	58.755	49.777	46.164	49.758
	MMSEU	12.915	11.462	13.088	13.303	11.582	13.850	13.965	11.771	11.186	11.865
	MMSEO	12.942	11.477	13.084	13.299	11.580	13.841	13.937	11.781	11.170	11.853
	MTrdScSE	17.878	15.845	18.089	18.394	16.000	19.138	19.292	16.276	15.439	16.383
	MVolScSE	0.053	0.047	0.053	0.054	0.047	0.056	0.057	0.048	0.046	0.048
	MSAdjPE	55.055	47.048	54.749	54.967	48.349	58.107	58.776	50.768	45.899	48.937
	MAE	2.876	2.603	2.921	2.928	2.809	2.968	3.060	2.870	2.726	2.755
	MAPE	6.088	5.509	6.186	6.189	5.987	6.332	6.514	6.126	5.748	5.845
	MMAEU	7.022	6.847	8.328	8.381	7.422	8.745	9.002	7.897	6.875	7.234
	MMAEO	8.811	7.262	7.711	7.877	7.004	8.105	8.039	6.781	7.066	7.419
	MTrdScAE	3.986	3.607	4.048	4.058	3.893	4.114	4.241	3.978	3.778	3.818
	MVolScAE	0.012	0.011	0.012	0.012	0.011	0.012	0.013	0.012	0.011	0.011
	MAAdjPE	6.081	5.484	6.198	6.197	6.016	6.331	6.554	6.178	5.755	5.838
Correct Sign	PCSP	1.000	0.984	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	PCDCP	1.000	0.609	0.594	0.531	0.609	0.516	0.406	0.641	0.672	0.625
Biasedness	Sign test - Exact Binomial	0	1	1	1	1	1	0	1	1	1
	Sign Test - Normal Approximation	0	1	1	1	1	1	0	1	1	1
	Regression Based Biasedness Test	0	1	1	1	0	1	1	0	1	1
	PSTSU	0	1	1	1	0.666	1	0.333	0.666	1	1
Informational Efficiency	Informational Efficiency Test (Beta)	0	0	0	0	0	0	0	0	0	0

Table 4: Performance Measures Corresponding to Out-of-sample Implementation with Rolling Origin and Fixed Window

As to preference articulation and modeling, the relevant decisions to make or questions to address along with their answers could be summarized as follows. The type of scale to use for measuring criteria is the ratio scale and such a choice is a consequence of the choices of criteria and their measures made above. In addition, the scales used are global in that they do not depend on the forecasting models under consideration. With respect to when preference articulation takes place, the nature of this research exercise as well as other decisions to be discussed later suggest that preference articulation should take place a priori. Again the nature of this research exercise suggests that an acceptable degree of compensation between criteria would be a partial one, as in practice one would in general tradeoff a criterion for another to a certain extent, but not completely. Regarding the choice of the preference elucidation mode, we do not privilege any

specific one although the choice of MCDA methods made later will imply specific modes of preference elucidation. Once again, the nature of this research exercise suggests that an appropriate type of discriminating power to use would be pseudo-criteria – the choice of the relevant thresholds will be discussed after the choice of the methods. With respect to the choice of preference structure(s), we opt for the most general preference structure; namely, {P, Q, I, R}, as it takes account of our choice of a non-absolute discriminating power; namely, pseudo-criteria, on one hand, and it allows for the possibility of incomparability between competing forecasting models, on the other hand. In addition, we do not discard the option of using valued preference relations. With respect to the choice of the type of order, a partial order would be allowed; however, we would prefer a complete order or weak order, which will be reflected in our choice of MCDA methods. Once again, the nature of this research exercise suggests that an appropriate choice of the MCDA paradigm to adopt as a modeling framework for preferences would be the social choice paradigm, because currently the conjoint measurement paradigm-based methods that could realistically be implemented do not allow for the possibility of incomparability between alternatives. Finally, with respect to the choice of the social choice paradigm-based methods, we opt for ELECTRE and PROMETHEE methods appropriate for the ranking problematic, on one hand, and compatible with our choices made so far, on the other hand; namely, ELETRE III, PROMETHEE I and PROMETHEE II.

As to preference aggregation and exploitation, the relevant decisions to make or questions to address along with their answers could be summarized as follows. Preference aggregation decisions are intimately related to our choices of MCDA methods; therefore, we are concerned with aggregating a performance table into a single outranking relation, which involves the estimation of inputs to ELETRE III, PROMETHEE I and PROMETHEE II methods and the exploitation of such outranking relations to devise a solution; that is, a ranking of competing forecasting models. Three types of inputs are required by ELETRE III, PROMETHEE I and PROMETHEE II methods; namely, a performance table, weights reflecting the relative importance of criteria, and relevant thresholds or preference function (i.e., indifference and preference

thresholds required by both ELECTRE III and PROMETHEE I & II, and veto threshold required by ELECTRE III). In this forecasting application, the performance table reflects the scores of the competing forecasting models with respect to the chosen criteria; namely, goodness-of-fit, biasedness, and correct sign, as measured by different possible measures – see Table 4. As to the weights, also referred to as the relative importance coefficients, we opted for a relative rating method; namely, point allocation. Our choice of this method is motivated by its simplicity from a user perspective in that criteria are rated relative to each other by distributing 100 points between them so as to reflect their relative importance. Given the nature of our application and the empirical studies on the relative importance of criteria (Carbone and Armstrong, 1982; Yokum and Armstron, 1995; Armstrong, 2001a; Winklhofer and Diamantopoulos, 2002; McCarthy et al., 2006), the weights of goodness-of-fit, biasedness, and correct sign are set to 0.50, 0.30, and 0.20, respectively; thus, reflecting that goodness-of-fit is by far the most important criterion followed by correct sign and then by biasedness. Obviously, one could argue that biasedness is more important than correct sign; however, in a crude oil application where price is the main concern, an investor is likely to prefer a method that provides good predictions of direction than a method that does not systematically under- or over-estimate the forecasts. Finally, for both goodness-of-fit and correct sign criteria, we define the above mentioned thresholds as functions of the range of values of the corresponding measure as follows:

$$\text{Indifference Threshold: } \iota_j = \alpha_j \times \left(\max\limits_{a \in A} imum \left\{ g_j(a) \right\} - \min\limits_{a \in A} imum \left\{ g_j(a) \right\} \right); \alpha_j \in (0,1)$$

$$\text{Preference Threshold: } \pi_j = \beta_j \times \left(\max\limits_{a \in A} imum \left\{ g_j(a) \right\} - \min\limits_{a \in A} imum \left\{ g_j(a) \right\} \right); \beta_j \in (0,1)$$

$$\text{Veto Threshold: } \nu_j = \gamma_j \times \left(\max\limits_{a \in A} imum \left\{ g_j(a) \right\} - \min\limits_{a \in A} imum \left\{ g_j(a) \right\} \right); \gamma_j \in (0,1)$$

where A denotes the set of alternative; that is, the competing forecasting models under consideration – see Table 3. Note that α_j, β_j and γ_j reflect percentages of the range of values taken by criterion j that would lead to indifference, preference, and veto situations, respectively. Given the nature of our application and the range of the values

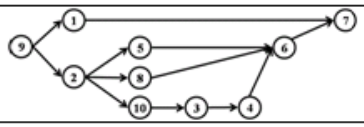
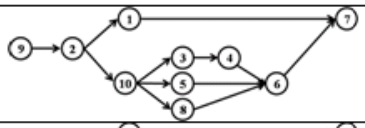
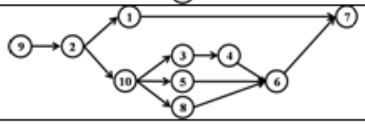
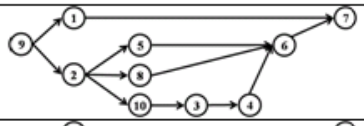
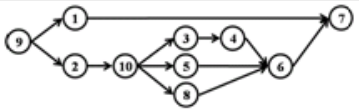
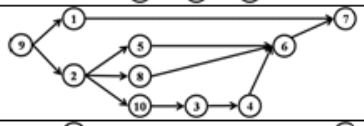
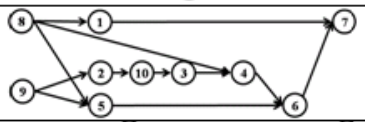
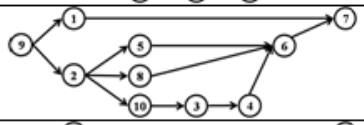
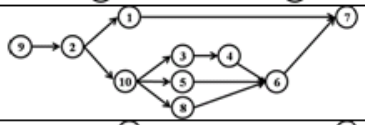
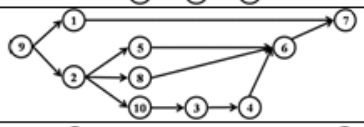
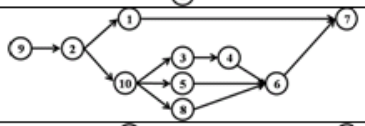
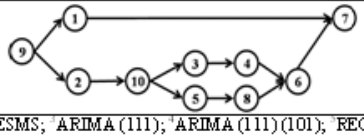
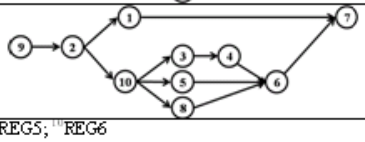
taken by the criteria under consideration, the values of α_j, β_j and γ_j are set to 1%, 5% and 10%, respectively. On the other hand, for biasedness criterion, ι_j, π_j and ν_j are set to 0.20, 0.33 and 0.67, respectively. Furthermore, the preference function $P_j(.)$ used within PROMETHEE I and PROMETHEE II is chosen as follows, where d_j denotes the difference in performance with respect to criterion j between a pair of alternatives or competing forecasting models:

$$P_j(d_j) = \begin{cases} 0 & \text{IF } d_j \le \iota_j \\ \dfrac{d_j - \iota_j}{\pi_j - \iota_j} & \text{IF } \iota_j \prec d \le \pi_j \\ 1 & \text{IF } d \succ \pi_j \end{cases}$$

	Ranking in Descending Order of A Specific Performance Measure										
Goodness-of-Fit	MSE	REG5	HWES MS	REG1	REG4	REG6	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MSPE	REG5	REG1	HWESMS	REG6	REG4	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG3	REG2
	MMSEU	REG5	HWES MS	REG1	REG4	REG6	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MMSEO	REG5	HWES MS	REG1	REG4	REG6	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MTrdScSE	REG5	HWES MS	REG1	REG4, REG6		RW With Trend	A ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MVolScSE	REG5	HWES MS	REG1	REG6	REG4	ARIMA (111)	ARIMA (111)(101)	RW With Trend	REG2	REG3
	MSAdjPE	REG5	HWES MS	REG1	REG4	REG6	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MAE	HWESMS	REG5	REG6	REG1	REG4	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MAPE	HWESMS	REG5	REG6	REG1	RW With Trend	REG4	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MMAEU	HWESMS	REG5	RW With Trend	REG6	REG1	REG4	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
	MMAEO	REG4	REG1	REG5	HWESMS	REG6	ARIMA (111)	ARIMA (111)(101)	REG3	REG2	RW With Trend
	MTrdScAE	HWESMS	REG1, REG5, REG6			RW With Trend	ARIMA (111)(101)	ARIMA (111)	REG2 REG4		REG3
	MVolScAE	HWESMS	REG5	REG6	REG1	RW With Trend	REG4	ARIMA (111)(101)	ARIMA (111)	REG2	REG3
	MAAdjPE	HWESMS	REG5	REG6	REG1	REG4	RW With Trend	ARIMA (111)	ARIMA (111)(101)	REG2	REG3
Correct Sign	PCDCP	RW With Trend	REG5	REG4	REG6	HWESMS, REG1		ARIMA (111)	ARIMA (111)(101)	REG2	REG3
Biasedness	PSTSU	HWESMS, ARIMA(111), ARIMA(111)(101), REG2, REG5, REG6						REG1, REG4		REG3	RW With Trend

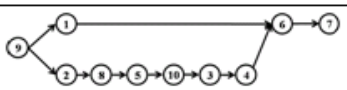
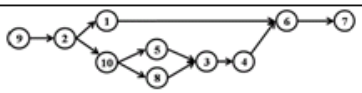
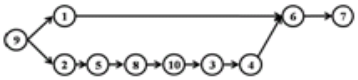
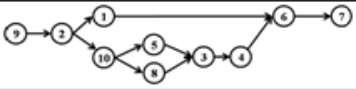
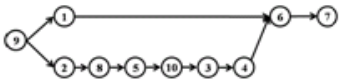
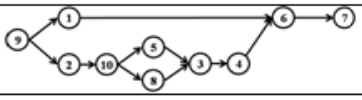
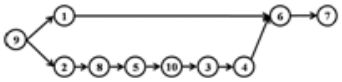
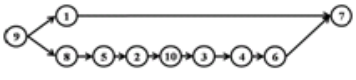
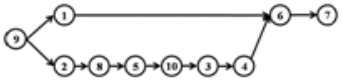
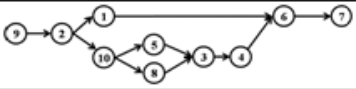
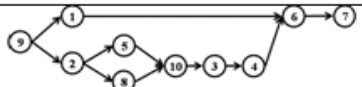
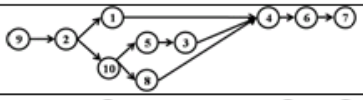
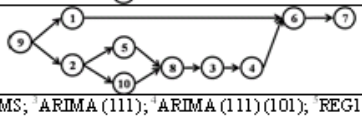
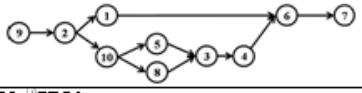
Table 5: Unidimensional Rankings of Competing Forecasting Models

The results or rankings obtained by ELETRE III are summarized in Tables 6 and 7, the rankings obtained by PROMETHEE I are summarized in Tables 8 and 9, and the rankings obtained by PROMETHEE II are summarized in Tables 10 and 11.

Performance Measures	Rank in Decreasing Order	Performance Measures	Rank in Decreasing Order
MSE, PCDCP, PSTSU		MAE, PCDCP, PSTSU	
MSPE, PCDCP, PSTSU		MAPE, PCDCP, PSTSU	
MMSEU, PCDCP, PSTSU		MMAEU, PCDCP, PSTSU	
MMSEO, PCDCP, PSTSU		MMAEO, PCDCP, PSTSU	
MTrdScSE, PCDCP, PSTSU		MTrdScAE, PCDCP, PSTSU	
MVolScSE, PCDCP, PSTSU		MVolScAE, PCDCP, PSTSU	
MSAdjPE, PCDCP, PSTSU		MAAdjPE, PCDCP, PSTSU	

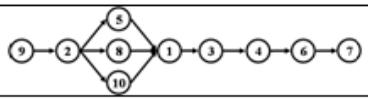
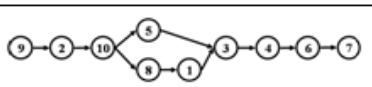
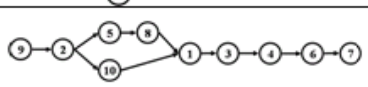
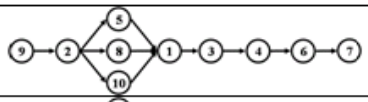
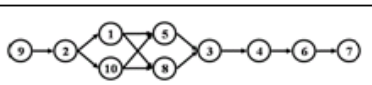
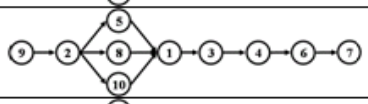
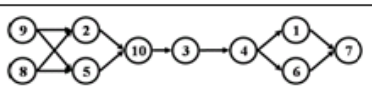
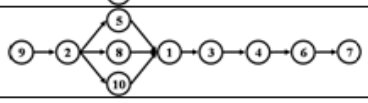
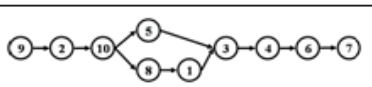
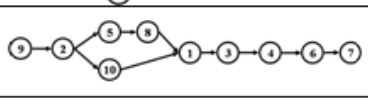
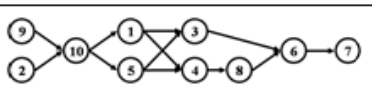
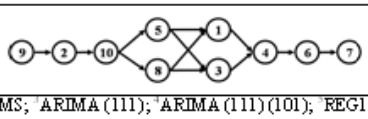
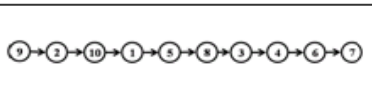
[1]RW with Trend; [2]HWESMS; [3]ARIMA (111); [4]ARIMA (111)(101); [5]REG1; [6]REG2; [7]REG3; [8]REG4; [9]REG5; [10]REG6

Table 6: ELECTRE III Ranking of Competing Forecasting Models with Weight Vector (50, 30, 20)

Performance Measures	Rank in Decreasing Order	Performance Measures	Rank in Decreasing Order
MSE, PCDCP, PSTSU		MAE, PCDCP, PSTSU	
MSPE, PCDCP, PSTSU		MAPE, PCDCP, PSTSU	
MMSEU, PCDCP, PSTSU		MMAEU, PCDCP, PSTSU	
MMSEO, PCDCP, PSTSU		MMAEO, PCDCP, PSTSU	
MTrdScSE, PCDCP, PSTSU		MTrdScAE, PCDCP, PSTSU	
MVolScSE, PCDCP, PSTSU		MVolScAE, PCDCP, PSTSU	
MSAdjPE, PCDCP, PSTSU		MAAdjPE, PCDCP, PSTSU	

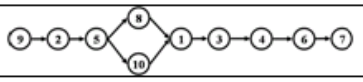
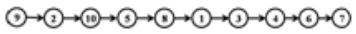
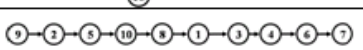
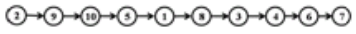
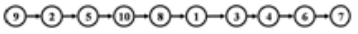
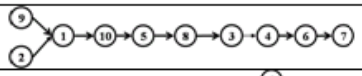
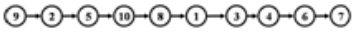
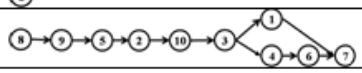
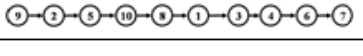
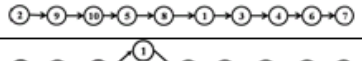
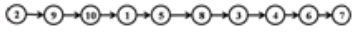
[1]RW with Trend; [2]HWESMS; [3]ARIMA (111); [4]ARIMA (111) (101); [5]REG1; [6]REG2; [7]REG3; [8]REG4; [9]REG5; [10]REG6

Table 7: ELECTRE III Ranking of Competing Forecasting Models with Weight Vector (70, 20, 10)

Performance Measures	Rank in Decreasing Order	Performance Measures	Rank in Decreasing Order
MSE, PCDCP, PSTSU		MAE, PCDCP, PSTSU	
MSPE, PCDCP, PSTSU		MAPE, PCDCP, PSTSU	
MMSEU, PCDCP, PSTSU		MMAEU, PCDCP, PSTSU	
MMSEO, PCDCP, PSTSU		MMAEO, PCDCP, PSTSU	
MTrdScSE, PCDCP, PSTSU		MTrdScAE, PCDCP, PSTSU	
MVolScSE, PCDCP, PSTSU		MVolScAE, PCDCP, PSTSU	
MSAdjPE, PCDCP, PSTSU		MAAdjPE, PCDCP, PSTSU	

[1]RW with Trend; [2]HWESMS; [3]ARIMA (111); [4]ARIMA (111) (101); [5]REG1; [6]REG2; [7]REG3; [8]REG4; [9]REG5; [10]REG6

Table 8: PROMETHEE I Ranking of Competing Forecasting Models With Weight Vector (50, 30, 20)

Performance Measures	Rank in Decreasing Order	Performance Measures	Rank in Decreasing Order
MSE, PCDCP, PSTSU		MAE, PCDCP, PSTSU	
MSPE, PCDCP, PSTSU		MAPE, PCDCP, PSTSU	
MMSEU, PCDCP, PSTSU		MMAEU, PCDCP, PSTSU	
MMSEO, PCDCP, PSTSU		MMAEO, PCDCP, PSTSU	
MTrdScSE, PCDCP, PSTSU		MTrdScAE, PCDCP, PSTSU	
MVolScSE, PCDCP, PSTSU		MVolScAE, PCDCP, PSTSU	
MSAdjPE, PCDCP, PSTSU		MAAdjPE, PCDCP, PSTSU	

[1]RW with Trend; [2]HWESMS; [3]ARIMA (111); [4]ARIMA (111) (101); [5]REG1; [6]REG2; [7]REG3; [8]REG4; [9]REG5; [10]REG6

Table 9: PROMETHEE I Ranking of Competing Forecasting Models With Weight Vector (70, 20, 10)

Performance Measures	Rank in Decreasing Order	Performance Measures	Rank in Decreasing Order
MSE, PCDCP, PSTSU	9→2→10→5→8→1→3→4→6→7	MAE, PCDCP, PSTSU	9→2→10→5→1→8→3→4→6→7
MSPE, PCDCP, PSTSU	9→2→5→10→8→1→3→4→6→7	MAPE, PCDCP, PSTSU	9→2→10→8→5→1→3→4→6→7
MMSEU, PCDCP, PSTSU	9→2→10→5→8→1→3→4→6→7	MMAEU, PCDCP, PSTSU	9→2→1→10→8→5→3→4→6→7
MMSEO, PCDCP, PSTSU	9→2→10→5→8→1→3→4→6→7	MMAEO, PCDCP, PSTSU	9→8→5→2→10→3→4→1→6→7
MTrdScSE, PCDCP, PSTSU	9→2→10→5→8→1→3→4→6→7	MTrdScAE, PCDCP, PSTSU	9→2→10→5→1→8→3→4→6→7
MVolScSE, PCDCP, PSTSU	9→2→10→5→8→1→3→4→6→7	MVolScAE, PCDCP, PSTSU	9→2→10→5→1→4→8→3→6→7
MSAdjPE, PCDCP, PSTSU	9→2→10→5→8→3→1→4→6→7	MAAdjPE, PCDCP, PSTSU	9→2→10→5→1→4→8→3→6→7

[1]RW with Trend; [2]HWESMS; [3]ARIMA (111); [4]ARIMA (111) (101); [5]REG1; [6]REG2; [7]REG3; [8]REG4; [9]REG5; [10]REG6

Table 10: PROMETHEE II Ranking of Competing Forecasting Models With Weight Vector (50, 30, 20)

Performance Measures	Rank in Decreasing Order	Performance Measures	Rank in Decreasing Order
MSE, PCDCP, PSTSU	9→2→5→8→10→1→3→4→6→7	MAE, PCDCP, PSTSU	2→9→10→5→8→1→3→4→6→7
MSPE, PCDCP, PSTSU	9→2→5→10→8→1→3→4→6→7	MAPE, PCDCP, PSTSU	2→9→10→5→1→8→3→4→6→7
MMSEU, PCDCP, PSTSU	9→2→5→8→10→1→3→4→6→7	MMAEU, PCDCP, PSTSU	9→2→1→10→5→8→3→4→6→7
MMSEO, PCDCP, PSTSU	9→2→5→8→10→1→3→4→6→7	MMAEO, PCDCP, PSTSU	8→9→5→2→10→3→4→1→6→7
MTrdScSE, PCDCP, PSTSU	9→2→5→8→10→1→3→4→6→7	MTrdScAE, PCDCP, PSTSU	2→9→10→5→8→1→3→4→6→7
MVolScSE, PCDCP, PSTSU	9→2→5→8→10→1→3→4→6→7	MVolScAE, PCDCP, PSTSU	2→9→10→5→1→4→8→3→6→7
MSAdjPE, PCDCP, PSTSU	9→2→10→5→8→3→1→4→6→7	MAAdjPE, PCDCP, PSTSU	2→9→10→1→5→8→3→4→6→7

[1]RW with Trend; [2]HWESMS; [3]ARIMA (111); [4]ARIMA (111) (101); [5]REG1; [6]REG2; [7]REG3; [8]REG4; [9]REG5; [10]REG6

Table 11: PROMETHEE II Ranking of Competing Forecasting Models With Weight Vector (70, 20, 10)

Notice that the ELETRE III rankings corresponding to the MSE, MMSEU, MMSEO, MTrdScSE and MVolScSE – along with the other same measures; that is, PCDCP and PSTSU – are identical (see Table 6), but for different reasons. In fact, the ELETRE III rankings corresponding to MSE, MMSEU and MMSEO are identical, because MSE, MMSEU and MMSEO are approximately the same for each model under consideration (see Table 4), which result from a very small proportion of errors being in the interval [0, 1]. In addition, the unidimensional ranking of models under MSE is the same as the ones under MMSEU and MMSEO. On the other hand, although the unidimensional rankings of models under MSE, MTrdScSE and MVolScSE are different, their multidimensional rankings turned out to be identical, because the multidimensional or multi-criteria rankings make use of pseudo-criteria as compared to the true criteria used in unidimensional rankings. Notice also that the rankings corresponding to MSE, MSPE and MSAdjPE – along with the other same measures – are different, but for different reasons. In fact, the rankings corresponding to MSE and MSPE differ, because squaring a number between 0 and 1 (respectively, greater than 1) results in a smaller (respectively, larger) number and in this application it turned out that when the MSPE is used, the veto takes effect between REG1 and REG4 as compared to no veto when using MSE. On the other hand, the rankings corresponding to MSE and MSAdjPE differ, because the one corresponding to MSE involves strict preference between REG1 and REG6 whereas the one corresponding to MSAdjPE involves hesitation between REG1 and REG6. As to the ELETRE III rankings of models based on measures involving absolute errors, with the exception of measures that penalize under- and over-estimation of forecasts; namely, MMAEU and MMAEO, the remaining rankings are identical (see Table 6) – although their unidimensional counterparts are different, because the multidimensional rankings make use of pseudo-criteria as compared to the true criteria used in unidimensional rankings. Notice that the rankings corresponding to MAE, MMAEU and MMAEO – along with the other same measures – are different, but for different reasons. In fact, the rankings corresponding to MAE and MMAEU differ, because when the MMAEU is used, the veto takes effect between RW with Trend and HWESMS as compared to no veto when using MAE. On the other hand, the rankings corresponding to MAE and

MMAEO are different, because REG4 dominates (respectively, is dominated by) REG5 on MMAEO (respectively, MAE).

In terms of PROMETHEE I, MSE, MMSEU, MMSEO and MTrdScSE – along with the other same measures; that is, PCDCP and PSTSU – lead to identical rankings (see Table 8), for the same reasons they are identical under ELECTRE III; namely, MSE, MMSEU and MMSEO are approximately the same for each model under consideration and the unidimensional rankings of models under MSE, MMSEU and MMSEO are the same. The rankings corresponding to MSE, MSPE, MVolScSE and MSAdjPE – along with the other same measures – are however different, because of the effect of pseudo-criteria on their positive and negative outranking flows due to their differences with respect to the measures MSPE, MVolScSE and MSAdjPE. For example, under MSE, REG1 and REG4 are incomparable, whereas REG1 outranks REG4 under MSPE due to changes in the values of their positive and negative outranking flows – resulting from a change from one condition of the preference function to another. As to the PROMETHEE I rankings of models based on measures involving absolute errors, with the exception of measures on MAE and MTrdScAE, the remaining ranking are all different (See Table 8). These differences a direct consequence of the effect of pseudo-criteria on their positive and negative outranking flows due to differences with respect to MAPE, MMAEU, MMAEO, MVolScAE and MAAdjPE, which led to changes from one condition of the preference function to another.

Finally, with our numerical data and regardless of whether the performance measures are functions of squared errors or absolute errors, PROMETHEE II rankings are all complete orders; thus, there are no ties between models. Note however that the differences between some rankings are also a direct consequence of the effect of pseudo-criteria on their positive, negative, and net outranking flows due to differences with respect to performance measures, which led to changes from one condition of the preference function to another. In addition, as expected, relatively high importance weights assigned to a specific criterion tend to produce rankings that are similar to the ones obtained by a unidimensional ranking method.

To conclude this section, we would like to remind the reader that, by design, the outputs of ELECTRE III and PROMETHEE I are partial weak orders; thus, allowing for incomparability, whereas the output of PROMETHEE II is a weak order, which does not allow for incomparability. Empirical results or rankings provided by ELECTRE III and PROMETHEE I are however substantially different in that ELECTRE III rankings have more incomparabilities as compared to PROMETHEE I rankings, which could be attributed to the underlying principles and to the difference in their designs. At this stage, a couple of general remarks on ELECTRE III, PROMETHEE I and PROMETHEE II are worth mentioning. First, the higher the importance weight assigned to a specific criterion or subset of criteria, the more discriminating it becomes; thus, leading to less cases of incomparability (see Table 7 and Table 9). Second, the ranks of some models (e.g., HWESMS, REG2, REG3, and REG5) are robust to changes in the importance weights as well as ranking methods, as compared to others, which help in selecting models especially in situations involving heterogeneous groups of decision makers. Last, but not least, we recommend to use several outranking methods and to make the final model(s) selection decision only after a comparative analysis of their outputs, as each method has different features. Furthermore, although in practice decision makers tend to prefer a method that produces a complete or a weak order such as PROMETHEE II, methods such as ELECTRE III and PROMETHEE I allow for incomparability between models which could result in higher quality forecasts when the individual forecasts produced by each model separately are combined.

4.5. Conclusion

The lack of a multidimensional framework for performance evaluation of competing forecasting models has motivated this research in which we proposed a framework based on Multi-Criteria Decision Analysis (MCDA) methodology. In sum, we revisited MCDA methodology and proposed a classification of those performance criteria and their measures that are commonly used in evaluating and selecting forecasting models, which would assist with the operationalisation of the revised MCDA framework. Finally, we discussed how one might adapt such MCDA framework to address the problem of

relative performance evaluation of competing forecasting models of crude oil prices. Three outranking methods have been used in our empirical experiments; namely, ELECTRE III, PROMETHEE I and PROMETHEE II, and our main conclusions 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, when relatively high importance weights are assigned to a specific criterion, some multidimensional outranking methods tend to produce rankings that are similar to the one obtained by a unidimensional ranking method. Third, in practice, one should use several outranking methods, compare their outputs, and eventually combine the forecasts of some of the models under consideration, which are incomparable, before deciding on the forecasting model or method to implement. Finally, as far as the evaluation of the relative performance of the valid forecasting models considered in this study is concerned, models such as the linear regression model REG5 and Holt-Winter Exponential Smoothing with Multiplicative Seasonality tend to have ranks that are not sensitive to importance weights or outranking methods, which suggest that these models are superior.

References

Abosedra S, Baghestani H. On the predictive accuracy of crude oil future prices. Energy Policy 2004; 32; 1389–1393.

Abramson B. The design of belief network-based systems for price forecasting. Computers and Electrical Engineering 1994; 20; 163–180.

Abramson B, Finizza A. Using belief networks to forecast oil prices. International Journal of Forecasting 1991; 7; 299–315.

Abramson B, Finizza A. Probabilistic forecasts from probabilistic models: a case study in the oil market. International Journal of Forecasting 1995; 11; 63–72.

Adrangi B, Chatrath A, Dhanda K.K., Raffiee K. Chaos in Oil prices? Evidence from future markets. Energy Economics 2001; 23; 405-425.

Adrangi B, Chatrath A, Raffiee K, Rippe R.D. Alaska North Slope crude oil price and the behavior of diesel prices in California. Energy Economics 2001; 23; 29-42.

Al-Gudhea S, Kenc T, Dibooglu S. Do retail gasoline prices rise more readily than they fall? A threshold cointegration approach. Journal of Economics and Business 2007; 59; 560–574.

Alvarez-Ramirez J, Cisneros M, Ibarra-Valdez C, Soriano A. Multifractal Hurst Analysis of crude oil prices. Physica A 2002; 313; 651-670.

Allais M 1979. The so-called Allais paradox and rational decisions under uncertainty. In: Allais M, Hagen O (Eds), Expected utility hypotheses and the Allais paradox. D. Reidel, Dordrecht. p. 437-681.

Alwan L.C. Statistical Process Analysis. McGraw-Hill: Boston; 2000.

Amano R.A., van Norden S. Oil prices and the rise and fall of the US real exchange rate. Journal of International Money and Finance 1998; 17; 299-316.

Anderson T.W., Darling D.A. A Test of Goodness of Fit. Journal of the American Statistical Association 1954; 49; 765-769.

Anderson T.W., Darling D.A. Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes. Annals of Mathematical Statistics 1952; 23; 193-212.

Armstrong J.S. Long-range forecasting: From Crystal Ball to computer. John Wiley: New York; 1985.

Armstrong J.S. Evaluating forecasting methods. In: Armstrong JS (Eds), Principles of forecasting: A handbook for researchers and practitioners; Kluwer Academic Publishers: Boston; 2001a. p. 443 – 472.

Armstrong J.S. Standards and practices for forecasting. In: Armstrong JS (Eds), Principles of forecasting: A handbook for researchers and practitioners; Kluwer Academic Publishers: Boston; 2001b. p. 679 – 732.

Armstrong J.S., Adya M, Collopy F. Standards and practices for forecasting. In: Armstrong JS (Eds), Principles of forecasting: A handbook for researchers and practitioners; Kluwer Academic Publishers: Boston; 2001. p. 285–300.

Arrow K.J, Raynaud H. Social choice and multicriterion decision-making. MIT Press: Cambridge; 1986.

Arrow K.J. Social choice and individual values. John Wiley; 1963.

Asche F, Osmunddsen P, Sandssmark M. The UK market for natural gas, oil and electricity: are the prices decoupled? The Energy Journal 2006; 27; 27–40.

Asche F, Gjolberg O, Volker T. Price relationships in the petroleum market: an analysis of crude oil and refined product prices. Energy Economics 2003; 25; 289-301.

Bachmeier L.J., Griffin J.M. New evidence on asymmetric gasoline price responses. Review of Economics and Statistics 2003; 85; 772–776.

Bachmeier L.J., Griffin J.M. Testing for market integration: crude oil, coal, and natural gas. The Energy Journal 2006; 27; 55–71.

Bacon R. Modelling the price of oil. Oxford Review of Economic Policy 1991; 7; 17–34.

Bana e Costa, C.A, Vansnick. J.C. Reference Relations and MCDM. In: Multicriteria decision making: advances in MCDM models, algorithms, theory, and applications. Kluwer: Boston; 1999. p.4.1-4.23.

Barker S, Cole R. Brilliant Project Management: What the Best Project Managers Know, Say, and Do. Prentice-Hall: Hants; 2007.

Bauer J.E, Duffy G.L, Westcott R. The Quality Improvement Handbook. ASQ Quality Press: Milwaukee; 2006.

Behzadian M, Kazemzadeh R.B., Albadvi A, Aghdasi M. PROMETHEE: A comprehensive literature review on methodologies and applications. European Journal of Operational Research 2010; 200; 198-215.

Bekiros S, Diksa C. The relationship between crude oil spot and futures prices: Cointegration, linear and nonlinear causality. Energy Economics 2008; 30; 2673-2685.

Belton V. A comparison of the analytic hierarchy process and a simple multi-attribute value function. European Journal of Operational Research 1986; 26; 7–21.

Belton V, Stewart T. Multiple Criteria Decision Analysis: An Integrated Approach. Kluwer Academic Publishers: Dordrecht; 2001.

Bernard J.T., Khalaf L, Kichian M. Structural change and forecasting long-run energy prices. Bank of Canada 2004; Working Papers 04-5.

Black D. The theory of committees and elections. Cambridge University Press: London; 1958.

Blackorby C, Primont D, Russell R. Duality, separability, and functional structure: Theory and economic applications. North-Holland: New York; 1978.

Bopp A.E., Sitzer S. Are petroleum futures prices good predictors of cash value. The Journal of Futures Markets 1987; 7; 705–719.

Bopp A. E., Lady G.M. A comparison of petroleum futures versus spot prices as predictors of prices in the future. Energy Economics 1991; 13; 274-282.

Borcherding K, Eppel T, von Winterfeldt D. Comparison of weighting judgments in multiattribute utility measurement. Management Science 1991; 37; 1603-1619.

Bottomley P.A., Doyle J.R. A comparison of three weight elicitation methods: good, better, and best. Omega 2001; 29; 553-560.

Bottomley P.A., Doyle J.R., Green R.H. Testing the reliability of weight elicitation methods: direct rating vs. point allocation. Journal of Marketing Research 2000; 37; 508–513.

Bouyssou D. Some remarks on the notion of compensation in MCDM. European Journal of Operational Research 1986; 26; 150-160.

Bouyssou D. A note of the Lmin in favor' ranking method for valued preference relations. In: Cerny M, Gliikaufovti D, Loula D (Eds), Multicriteria decision making. Methods, algorithms, applications. Czechoslovak Academy of Sciences: Prague; 1991. p. 16-25.

Bouyssou D. A note on the 'min in favor' choice procedure for fuzzy preference relations. In: Pardalos PM, Siskos Y, Zopounidis C (Eds), Advances in multicriteria analysis. Kluwer: Dordrecht; 1995. p. 9-16.

Bouyssou D, Marchant T, Pirlot M, Tsoukias A, Vincke Ph. Evaluation and decision models with multiple criteria: Stepping stones for the analyst. International Series in Operations Research and Management Science: Boston; 2006.

Bouyssou D, Perny P. Ranking methods for valued preference relations: A characterization of a method based on entering and leaving flows. European Journal of Operational Research 1992; 61; 186-194.

Bouyssou D. Building criteria: A prerequisite for MCDA. In: Bana e Costa CA (Eds), Readings in Multiple Criteria Decision Aid. Springer-Verlag: Berlin; 1990. p. 58-80.

Bouyssou D, Perny P, Pirlot M, Tsoukihs A, Vincke Ph. A manifesto for the new MCDM era. Journal of Multi-Criteria Decision Analysis 1993; 2; 125-127.

Bouyssou D. Some remarks on the notion of compensation in MCDM. European Journal of Operational Research 1986; 26; 150-160.

Bouyssou D, Pirlot M. A characterization of strict concordance relations. In: Bouyssou D, Jacquet-LagrBze E, Perny P, Slowinski R, Vanderpooten D, Vincke Ph (Eds), Aiding decisions with multiple criteria: Essays in honour of Bernard Roy. Kluwer: Dordrecht. 2002a. p. 121-145.

Bouyssou D, Pirlot M. Nontransitive decomposable conjoint measurement. Journal of Mathematical Psychology 2002b; 46; 677-703.

Bouyssou, D, Pirlot M. Preferences for multi-attributed alternatives: Traces, dominance and numerical representations. Journal of Mathematical Psychology 2004; 48; 167-185.

Bouyssou D, Pirlot M. Conjoint measurement tools for MCDM - A brief introduction. In: Figueira J, Greco S, Ehrgott M (Eds), Multiple criteria decision analysis - State of the art surveys. Springer-Verlag: Berlin. 2005. p. 73-130.

Brans J.P. L'ingénièrie de la décision; Elaboration d'instruments d'aide à la décision. La méthode PROMETHEE. In: Nadeau R, Landry M (Eds), L'aide à la décision: Nature, Instruments et Perspectives d'Avenir. Québec: Presses de l'Université Laval; 1982. p. 183–213.

Brans J.P. The space of freedom of the decision maker modeling the human brain. European Journal of Operational Research 1996; 92; 593-602.

Brans J.P., Mareschal B. PROMETHEE V: MCDM problems with segmentation constraints. INFOR 1992; 30; 85–96.

Brans J.P., Mareschal B. The PROMCALC and GAIA decision support system for MCDA. Decision Support Systems 1994; 12; 297–310.

Brans J.P., Mareschal B. The PROMETHEE VI procedure. How to differentiate hard from soft multicriteria problems. Journal of Decision Systems 1995; 4; 213–223.

Brans J.P., Mareschal B. PROMETHEE methods. In: Figueira J, Greco S, Ehrgott M (Eds), Multiple criteria decision analysis - State of the art surveys. Springer-Verlag: Berlin. 2005. p. 163–195.

Brans J.P., Mareschal B, Vincke Ph. PROMETHEE: A new family of outranking methods in multicriteria analysis. In: Brans JP (Eds), Operational Research. North-Holland: Amsterdam; 1984; 477-490.

Brans J.P., Mareschal B, Vincke Ph. How to select and how to rank projects: The PROMETHEE method. European Journal of Operational Research 1986; 24; 228-238.

Brans J.P., Vincke Ph. A preference ranking organisation method. The PROMETHEE method for MCDM. Management Science 1985; 31; 647-656.

Brue G. Six Sigma for Managers. McGraw-Hill: New York; 2002.

Buzan T, Buzan B. The Mind Map Book: How to Use Radiant Thinking to Maximize Your Brain's Untapped Potential. Penguin Books: New York; 1993.

Carbone R. Armstrong J.S. Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners. Journal of Forecasting 1982; 1; 215-217.

Carrington P.J., Scott J, Wasserman S. Models and Methods in Social Network Analysis. Cambridge University Press: New York; 2005.

Ciner C. Energy shocks and financial markets: nonlinear linkages, Studies in Nonlinear Dynamics and Econometrics 2001; 5; 203–212.

Chamberlin J. R., Courant P.N. Representative deliberations and representation decisions: Proportional representation and the Borda rule. American Political Science Review 1983; 77; 718-733.

Chao J, Corradi V, Swanson N.R. Out-of-sample tests for Granger causality. Macroeconomic Dynamics 2001; 5; 598–620.

Checkland P. Systems thinking, systems practice. Wiley: New York; 1981.

Chen Y, Kilgourb M, Hipel K.W. Multiple criteria classification with an application in water resources planning. Computers and Operations Research 2006; 33; 3301–3323.

Christensen E.H., Coombes-Betz K.M, Stein M.S. The Certified Quality Process Analyst Handbook. ASQ Quality Press: Milwaukee; 2007.

Collopy F. Armstrong J.S. Expert opinions about extrapolation and the mystery of the overlooked discontinuities. International Journal of Forecasting 1992; 8; 575-582.

Condorcet C, marquis de M.J.A.N. Essai sur l'application de l'analyse d la probabilite' des dtcisions rendues ci la pluralite' des voix. Imprimerie Royale: Paris. 1785.

Coppola A. Forecasting oil price movements: exploiting the information in the futures market. Journal of Futures Market 2008; 28; 34–56.

Crowder W. J., Hamid A. A cointegration test for oil futures market efficiency. Journal of Futures Markets 1993; 13; 933–941.

Csörgö S, Faraway J.J. On the Estimation of a Normal Variance. Statistics and Decisions 1996; 14; 23–34.

Cuñado J, Perez de Gracia F. Do Oil Price Shocks Matter? Evidence from Some European Countries. Energy Economics 2003; 25; 137-154

D'agostino R.B., Stephens M.A. Goodness-of-Fit Techniques. Marcel Dekker Inc: New York; 1986.

Dalkir K. Knowledge Management in Theory and Practice. Elsevier: Burlington; 2005.

Dallal G.E., Wilkinson L. An Analytic Approximation to the Distribution of Lilliefors Test Statistic for Normality. American Statistician 1986; 40; 294-296.

Darby M.R. The price of oil and world inflation and recession. The American Economic Review 1982; 72; 738–751.

Dalrymple D.J. Sales forecasting methods and accuracy. Business Horizons 1975; 18; 69– 73.

Dalrymple D.J. Sales forecasting practices. International Journal of Forecasting 1987; 3; 379-391.

Davis C.S., Stephens M.A. Algorithm AS 248: Empirical Distribution Function Goodness-of-Fit Tests. Applied Statistics 1989; 38; 535-543.

De Borda, J.Ch. Memoire sur les elections au scrutin. Histoire de l'Academie Royale des Sciences, annee MDCCLXXXI: Paris; 1784; p. 657-665.

De Keyser W, Peeters P. A note on the use of PROMETHEE multicriteria methods. European Journal of Operational Research 1996; 89; 457–461.

Debord B. An axiomatic characterization of Borda's k-choice function. Social Choice and Welfare, 1992; 9; 337-343.

Dent W. Analytic Approximation to Distribution of Durbin-Watson Statistic in Certain Alternative Cases. The Indian Journal of Statistics Series B 1974; 36; 163-174.

Dhillon B.S. Engineering and Technology Management Tools and Applications. Artech House: Boston; 2002.

Dias L.C., Mousseau V. Inferring ELECTRE's veto-related parameters from outranking example. European Journal of Operational Research 2006; 17; 172–191.

Doumpos M, Zopounidis C. Multicriteria Decision Aid Classification Methods. Kluwer Academic Publishers: Dordrecht; 2002.

Doyle J.R., Green R.H., Bottomley P.A. Judging relative importance: direct rating and point allocation are not equivalent. Organizational Behavior and Human Decision Processes 1997; 70; 65–72.

Dummett M. The Borda count and agenda manipulation. Social Choice and Welfare 1998; 15; 287-296.

Durbin J, Knott M. Components of Cramer-Von Mises Statistics I. Journal of the Royal Statistical Society Series B 1972; 34; 290-307.

Durbin J, Knott M, Taylor C.C. Components of Cramer Vonmises Mises Statistics II. Journal of the Royal Statistical Society Series B 1975; 37; 216-237.

Eden C, Ackermann F. Cognitive Mapping Expert Views for Policy Analysis in the Public Sector. European Journal of Operational Research 2004; 152; 615-630.

Eden C, Jones S, Sims D. Messing about in problems. Pergamon Press: Oxford; 1983.

Eden C. Cognitive Mapping. European Journal of Operational Research 1988; 36; 1-13.

Eden C. Cognitive Mapping and Problem Structuring for System Dynamics Model-Building. System Dynamics Review 1994; 10; 257-276.

Edwards W. Social utilities. Engineering Economist 1971; 6; 119-129.

Edwards W. How to use multiattribute utility measurement for social decision making. IEEE Transactions on Systems, Man and Cybernetics 1977; 7; 326-340.

Edwards W, Barron F. SMART and SMARTER: Improved simple methods for multiattribute utility measurement. Organizational Behavior and Human Decision Processes 1994; 60, 306-325.

El-Sharif I, Brown D, Burton B, Nixon B, Russell A. Evidence on the nature and extent of the relationship between oil prices and equity values in the UK. Energy Economics 2005; 27; 819–830.

Evans J.R., Lindsay W.M. The Management and Control of Quality. South-Western Thomson Learning: Ohio; 1999.

Faff R, Brailsford T.J. Oil price risk and the Australian stock market. Journal of Energy Finance and Development 1999; 4; 69–87.

Fan Y, Liang Q, Wei Y.M. A generalized pattern matching approach for multi-step prediction of crude oil price. Energy Economics 2006; 30; 889–904.

Fernandez V. Forecasting crude oil and natural gas spot prices by classification methods. Documentos de Trabajo from Centro de Economía Aplicada, Universidad de Chile. No 229.

Figueira J, Roy B. Determining the weights of criteria in the ELECTRE type methods with a revised Simos' procedure. European Journal of Operational Research 2002; 139; 317–326.

Figueira J, Mousseau V, Roy B. ELECTRE methods. In: Figueira J, Greco S, Ehrgott M (Eds), Multiple criteria decision analysis - State of the art surveys. Springer-Verlag: Berlin. 2005. p. 133–162.

Fischer G.W. Range sensitivity of attribute weights in multiattribute value models. Organizational Behavior and Human Decision Processes 1995; 62; 252-266.

Fishburn P.C. Utility theory for decision making. Wiley: New York; 1970.

Fishburn P.C. The theory of social choice. Princeton University Press: Princeton; 1973.

Fishburn P.C. Lexicographic orders, utilities and decision rules: A survey. Management Science 1974; 20; 1442-1471.

Fishburn P.C. Noncompensatory preferences. Synthese 1976; 33; 393-403.

Fishburn P.C. Condorcet social choice functions. SIAM Journal on Applied Mathematics 1977; 33; 469-489.

Fishburn P.C. Lexicographic additive differences. Journal of Mathematical Psychology 1980; 21; 191-218.

Fishburn P.C. Interval Orders and Interval Graphs. John Wiley & Sons: New York; 1985.

Fishburn P.C. Additive non-transitive preferences. Economic Letters 1990a; 34; 317-321.

Fishburn P.C. Continuous nontransitive additive conjoint measurement. Mathematical Social Sciences 1990b; 20; 165-193.

Fishburn P.C. Nontransitive additive conjoint measurement. Journal of Mathematical Psychology 1991a; 35; l-40.

Fishburn P.C. Nontransitive preferences in decision theory. Journal of Risk and Uncertainty 1991b; 4; 113-134.

Fishburn P.C. Additive differences and simple preference comparisons. Journal of Mathematical Psychology 1992; 36; 21-31.

Fishburn P.C. Generalisations of semiorders: A review note. Journal of Mathematical Psychology 1997; 41; 357–366.

Fishburn, P.C. Preference structures and their numerical presentations. Theoretical Computer Science 1999; 217; 359–389.

Fodor J, Marichal J.L., Roubens M. Characterization of the ordered weighted averaging operators. IEEE Transactions on Fuzzy Systems 1995; 3; 236-240.

Fodor J, Roubens M. On meaningfulness of means. Journal of Computational and Applied Mathematics 1995; 64; 103-115.

Foster S.T. Managing Quality: An Integrative Approach. Prentice-Hall: New Jersey; 2001.

Fortemps Ph, Pirlot M. Conjoint axiomatization of min, discrimin and leximin. Fuzzy Sets and Systems 2004; 148; 211-229.

Friend J.K., Hickling A. Planning under pressure: The strategic choice approach. Pergamon Press: New York; 1987.

Friend J.K., Jessop W.N. Local government and strategic choice. Tavistock: London; 1969.

Galeotti M, Lanza A, Manera M. Rockets and feathers revisited: an international comparison on European gasoline markets. Energy Economics 2003; 25; 175–190.

Garcia-Lapresta J.L., Llamazares B. Aggregation of fuzzy preferences: Some rules of the mean. Social Choice and Welfare 2000; 17; 673-690

Gitlow H, Oppenheim A, Oppenheim R. Quality Management: Tools and Methods for Improvement. Burr Ridge: Irwin; 1995.

Gisser M, Goodwin T.H. Crude oil and the macroeconomy: Tests of some popular notions. Journal of Money, Credit and Banking 1986; 18; 95–103.

Ghaffari A, Zare S. A novel algorithm for prediction of crude oil price variation based on soft computing. Energy Economics 2009; 31; 531-536

Godby R, Lintner A.M., Stengos T, Wandschneider B. Testing for asymmetric pricing in the Canadian retail gasoline market, Energy Economics; 2000; 22; 349–368.

Green S.L., Mork K.A. Toward efficiency in the crude oil market. Journal of Applied Econometrics 1991; 6; 45–66.

Gryna F.M., Juran J.M. Quality Planning and Analysis – From Product Development through Use. McGraw-Hill: Boston; 2001.

Guitouni A, Martel J.M. Tentative guidelines to help choosing an appropriate MCDA Method. European Journal of Operational Research 1998; 109, 501-521.

Guitouni A, Martel J.M., Vincke P. A Framework to Choose a Discrete Multicriterion Aggregation Procedure, Technical Report TR/SMG/2000-003, SMG, Universite Libre de Bruxelles, 2000.

Gulen S.G. Regionalization in world crude oil markets: Further evidence. The Energy Journal 1999; 20; 125–139.

Hamilton J.D. Oil and the macroeconomy since World War II. Journal of Political Economy 1983; 91; 228–248.

Hamilton J.D. Historical causes of postwar oil shocks and recessions. Energy Journal 1985; 6; 97–116.

Hamilton J.D. This is what happened to the oil price–macroeconomy relationship. Journal of Monetary Economics 1996; 38; 215–220.

Hanke J.E., Reitsch A.G. Business forecasting. Prentice-Hall: New Jersey; 1995.

Hansen H, Johansen S. Some tests for parameter constancy in cointegrated VAR-models. Econometrics Journal 1999; 2; 306–333.

Harmon P. Business Process Change: A Guide for Business Managers and BPM and Six Sigma Professionals. Elsevier/Morgan Kaufmann Publishers: Boston; 2007.

Harrison F.L., Lock D. Advanced Project Management: A Structured Approach. Gower: Burlington; 2004.

Hillson D, Simon P. Practical Project Risk Management: The Atom Methodology. Management Concepts: Virginia; 2007.

Hokkanen J, Salminen S. Choice of a Solid Waste Management System by Using the ELECTRE III Method, Applying MCDA for Decision to Environmental Management. Kluwer Academic Publishers: Holland; 1994.

Huang R.D., Masulis R.W., Stoll H.R. Energy shocks and financial markets. Journal of Futures Markets 1996; 16; 1−27.

Hugonnard J, Roy B. Le plan d'extension du métro en banlieue parisienne, uncas type d'application de l'analyse multicritère. Les Cahiers Scientifiques de la Revue Transports 1982; 6; 77–108.

Ishikawa K. What Is Total Quality Control? The Japanese Way. Prentice-Hall: New Jersey; 1985.

Jacquet-Lagrèze E. An application of the UTA discriminant model for the evaluation of R&D projects. In: Pardalos PM, Siskos Y, Zopounidis C (Eds), Advances in Multicriteria Analysis. Kluwer Academic Publishers: Dordrecht; 1995. p. 203–211.

Jacquet-Lagrkze F, Siskos Y. Assessing a set of additive utility functions for multicriteria decision making: The UTA method. European Journal of Operational Research 1982; 10; 151-164.

Jacquet-Lagreze E, Siskos Y. Preference disaggregation: 20 years of MCDA experience. European Journal of Operational Research 2001; 130; 233-245.

Kaboudan M.A. Compumetric forecasting of crude oil prices. Proceedings of the 2001 Congress on Evolutionary Computation 2001; 1, 283-287.

Kaufmann R.K., Laskowski C. Causes for an asymmetric relation between the price of crude oil and refined petroleum products. Energy Policy 2005; 33; 1587–1596.

Keeney R.L. Value-focused thinking. A path to creative decision making. Harvard University Press: Cambridge; 1992.

Keeney R.L., Raiffa H. Decisions with Multiple Objectives: Preferences and Value Trade-offs. Wiley: New York; 1976.

Keeney R.L., Raiffa H. Decisions with Multiple Objectives – Preferences and Value Tradeoffs. Cambridge University Press: Cambridge; 1993.

Kiker G.A., Todd S, Varghese B.A., Seager T.P., Linkov I. Application of Multicriteria Decision Analysis in Environmental Decision Making. Integrated Environmental Assessment and Management 2005; 1; 95–108.

Kilian L. Exogenous oil supply shocks: how big are they and how much do they matter for the U.S. economy? Review of Economics and Statistics 2008a; 90; 216–240.

Kilian L. A comparison of the effects of exogenous oil supply shocks on output and inflation in the G7 countries. Journal of the European Economic Association 2008b; 6; 78–121.

Kilian L. Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market. American Economic Review 2009; 99; 1053-1069.

Kilian L, Park C. The Impact of Oil Price Shocks on the U.S. Stock Market. Forthcoming: International Economic Review.

Kilian L, Rebucci A, Spataford N. Oil Shocks and External Balances. Journal of International Economics 2009; 77; 181-194.

Kiss L, Martel J.M., Nadeau R. ELECCALC – an interactive software for modelling the decision maker's preferences. Decision Support Systems 1994; 12; 757–777.

Kitchin R, Freundschuh S. Cognitive Mapping: Past, Present and Future. Routledge: London; 2000.

Knetsch T.A. Forecasting the price of crude oil via convenience yield predictions. Journal of Forecasting 2007; 6; 527-549.

Knoke D, Yang S. Social Network Analysis. Sage Publications: California; 2008.

Kolmogoroff A. Sur la notion de moyenne. Atti delle Reale Accademia Nazionale dei Lincei 1930; 12; 388-391.

Krantz D.H. Conjoint measurement: The Luce-Tukey axiomatization and some extensions. Journal of Mathematical Psychology 1964; 1; 248-277.

Krantz D.H., Luce R.D., Suppes P., Tversky A. Foundations of measurement: Additive and polynomial representations. Academic Press: New York; 1971.

Lalonde R, Zhu Z, Demers F. Forecasting and Analyzing World Commodity Prices. Bank of Canada 2003; No. 24.

Lanza A, Manera M, Grasso M, Giovannini M. Long-run Models of Oil Stock Prices. Environmental Modelling and Software 2005; 20; 1423-1430.

Landry M, Malouin J.L., Oral M. Model validation in Operations Research. European Journal of Operational Research 1983; 14; 207-220.

Lardic S, Mignon V. Oil prices and economic activity: An asymmetric cointegration approach. Energy Economics 2008; 30; 847-855.

Lee K, Ni S, Ratti R. Oil shocks and the macroeconomy: The role of price variability. The Energy Journal 1995; 16; 39–56.

Lewis J.P. Fundamentals of Project Management. American Management Association: New York; 2007.

Lewis P.A. Distribution of Anderson-Darling Statistic. Annals of Mathematical Statistics 1961; 32; 1118-1124.

Lilliefors H. On Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Journal of the American Statistical Association 1967; 62; 399-402.

Liu J.L., Bai Y, Li B. A New Approach to Forecast Crude Oil Price Based on Fuzzy Neural Network. Fourth International Conference on Fuzzy Systems and Knowledge Discovery 2007; 3; 273-277.

Longo C, Manera M, Markandya A, Scarpa E. Evaluating the empirical performance of alternative econometric models for oil price forecasting. FEEM Working Paper 2007; No 4.

Loungani P. Oil Price Shocks and the Dispersion Hypothesis. Review of Economics and Statistics 1986; 68; 536–539.

Luce R.D., Raiffa H. Games and decisions: Introduction and critical survey. Wiley: New York; 1957.

Mahmoud E. Accuracy in forecasting: A survey. Journal of Forecasting 1984; 3; 139–159.

Mahmoud E, Rice G. Malhotra N. Emerging issues in sales forecasting and decision support systems. Journal of Academy of Marketing Science 1986; 16; 47-61.

Marchant Th. Valued relations aggregation with the Borda method. Journal of Multi-Criteria Decision Analysis 1996; 5; 127-132.

Marchant Th. Cardinality and the Borda score. European Journal of Operational Research 1998; 108; 464-472.

Marchant Th. Does the Borda rule provide more than a ranking? Social Choice and Welfare 2000; 17; 381-391.

Marchant Th. The probability of ties with scoring methods: Some results. Social Choice and Welfare 2001; 18; 709-735.

Marchant Th. Towards a theory of MCDM: Stepping away from social choice theory. Mathematical Social Sciences 2003; 45; 343-363.

Marchant Th. The measurement of membership by comparisons. Fuzzy Sets and Systems 2004a; 148; 157-177.

Marchant Th. The measurement of membership by subjective ratio estimation. Fuzzy Sets and Systems 2004b; 148; 79-199.

Marquering W, Verbeek M. A multivariate nonparametric test for return and volatility timing. Finance Research Letters 2004; 1; 250–260.

Massey F.J. The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association 1951; 46; 68-78.

May K.O. A set of independent necessary and sufficient conditions for simple majority decisions. Econometrica 1952; 20; 680-684.

Maystre L, Pictet J, Simos J. Methodes Multicriteres ELECTRE, Description, conseils pratiques et cas d'application a la gestion environnementale. Presses Polytechniques et Universitaires Romandes: Lausanne; 1994.

McCarthy T.M., Davis D.F., Golicic S.L., Mentzer J.T. The Evolution of sales forecasting management: A 20-year longitudinal study of forecasting practices. Journal of Forecasting 2006; 25; 303-324.

McKellar B.H.J. Light-Scattering Determination of the Size Distribution of Cylinders - an Analytic Approximation. Journal of the Optical Society of America 1982; 72; 671-672.

McLean I, Urken A.B. Classics of social choice. University of Michigan Press: Ann Arbor; 1995.

Mentzer J.T., Cox J. A model of the determinants of achieved forecast accuracy. Journal of Business Logistics. Journal of Business Logistics 1984a; 5; 143–155.

Mentzer J.T., Cox J. Familiarity, application, and performance of sales forecasting techniques. Journal of Forecasting 1984b; 3; 27-36.

Mentzer J.T. Kahn K.B. Forecasting technique familiarity, satisfaction, usage, and application. Journal of Forecasting 1995; 14; 465–476.

Mirmirani S, Li H.C. A comparison of VAR and neural networks with genetic algorithm in forecasting price of oil. Advances in Econometrics 2004; 19; 203–223.

Mitra A. Fundamentals of Quality Control and Improvement. Prentice-Hall: New Jersey; 1998.

Mizuno S. Management for Quality Improvement: The Seven New QC Tools. Productivity Press: Cambridge; 1988.

Mork K.A. Oil and the macroeconomy when prices go up and down: An extension of Hamilton's results. Journal of Political Economy 1989; 91; 740–744.

Morgan W.A. A test for significance of the difference between two variances in a sample from a normal bivariate population. Biometrika 1939; 31; 13–19.

Moshiri S, Foroutan F. Forecasting nonlinear crude oil futures prices. The Energy Journal 2006; 27; 81–95.

Mousseau V. Eliciting Information Concerning the Relative Importance of Criteria. In: Pardalos PM, Siskos Y, Zopounidis C (Eds), Advances in Multicriteria Analysis: Nonconvex Optimization and its Applications, vol 5. Kluwer Academic Publishers: Dordrecht; 1995. p.17–43.

Mousseau V, Figueira J, Naux J-Ph. Using assignment examples to infer weights for ELECTRE TRI method: Some experimental results. European Journal of Operational Research 2001; 130; 263-275.

Mousseau V, Slowinski R. Inferring an ELECTRE TRI model from assignment examples. Journal of Global Optimization 1998; 12; 157–174.

Mousseau V, Slowinski R, Zielniewicz P. A user-oriented implementation of the ELECTRE TRI method integrating preference elicitation support. Computers and Operations Research 2000; 27; 757–777.

Mukherjee P.N. Total Quality Management. Prentice Hall: India; 2006.

Mukhopadhyaya A.K. Value Engineering: Concepts, Techniques and Applications. Response: New Delhi; 2003.

Nadler G. The Planning and Design Approach. Wiley: Chichester; 1981.

Nadler G. Hibino S. Breakthrough Thinking: Seven Principles of Creative Problem Solving. Prima Publishing: California; 1998.

Ngo The A, Mousseau V. Using assignment examples to infer category limits for the ELECTRE TRI method. Journal of Multi-Criteria Decision Analysis 2002; 11; 29–43.

Nitzan S, Rubinstein A. A further characterization of Borda ranking method. Public choice 1981; 36; 153-158.

Olson D.L., Dorai V.K. Implementation of the centroid method of Solymosi and Dombi. European Journal of Operational Research 1992; 60; 117-129.

Oral M, Kettani O. The facets of the modeling and validation process in operations research. European Journal of Operational Research 1993; 66; 216-234.

Ormerod R. Mixing methods in practice. In: Rosenhead J, Mingers J (Eds), Rational Analysis for a Problematic World Revisited: Problem Structuring Methods for Complexity, Uncertainty and Conflict. Wiley: Chichester. 2001. p. 311–336.

Ott E. R., Schilling E.G., Neubauer D.V. Process Quality Control: Troubleshooting and Interpretation of Data. McGraw Hill: New York; 2000.

Ozernoy V.M. Choosing the ‘Best’ Multiple Criteria Decision Making Method. INFOR 1992; 30; 159-171.

Oztürk M, Tsoukiàs A, Vincke Ph. Preference Modelling. In: Figueira J, Greco S, Ehrgott M (Eds), Multiple Criteria Decision Analysis: State of the Art Surveys. Springer-Verlag: Berlin; 2005. p. 27-72.

Panagiotidis T, Rutledge E. Oil and Gas Markets in the UK: A cointegrating approach. Energy Economics 2007; 29; 329-347.

Papapetrou E. Oil price shocks, stock markets, economic activity and employment in Greece. Energy Economics 2001; 23; 511−532.

Perny P, Roy B. The use of fuzzy outranking relations in preference modelling. Fuzzy Sets and Systems 1992; 49; 33-53.

Pictet J, Bollinger D. Extended use of the cards procedure as a simple elicitation technique for MAVT. Application to public procurement in Switzerland. European Journal of Operational Research 2008; 185; 1300-1307.

Pindyck R. S. The long-run evolution of energy prices. The Energy Journal 1999; 20; 1-27.

Pirlot M. A characterisation of 'min' as a procedure for exploiting valued preference relations and related results. Journal of Multi-Criteria Decision Analysis 1995; 4; 37-56.

Pirlot M. A common framework for describing some outranking procedures. Journal of Multi-Criteria Decision Analysis 1997; 6; 86-93.

Radchenko S. Lags in the response of gasoline prices to changes in crude oil prices: the role of short-term and long-term shocks. Energy Economics 2005; 27; 573–502.

Raiffa H. Preference for multi-attributed alternatives. RAND Memorandum, RM-5868-DOT/RC, Santa Monica. 1969.

Raiffa H. Decision analysis: Introductory lectures on choices under uncertainty. Addison Wesley: Reading; 1970.

Ramakrishna H.V., Brightman H.J. The Fact-Net Model: A Problem Diagnosis Procedure. Interfaces 1986; 16; 86-94.

Rautava J. The role of oil prices and the real exchange rate in Russia's economy-a cointegration approach. Journal of Comparative Economics 2004; 32; 315-327.

Regenwetter M, Grofman B. Approval voting, Borda winners and Condorcet winners: Evidence from seven elections. Management Science 1998; 44; 520-533.

Roberts K.W.S. Interpersonal comparability and social choice theory. Review of Economic Studies 1980; 47; 421-439

Rocha C. J. Essentials of Social Work Policy Practice. John Wiley & Sons: New Jersey; 2007.

Rogers M, Bruen M. Choosing Realistic Values of Indifference, Preference and Veto Thresholds for use with Environmental Criteria within ELECTRE. European Journal of Operational Research 1998a; 107; 542-551.

Rogers M, Bruen M. A new system for weighting environmental criteria for use within ELECTRE III. European Journal of Operational Research 1998b; 107; 552-563.

Rogers M, Bruen M, Maystre L. ELECTRE and Decision Support. Kluwer Academic Press: Boston. 2000.

Roy B. Classement et choix en presence de points de vue multiples: La methode ELECTRE. R.I.R.O 1968; 8; 57–75.

Roy B. Algèbre Moderne et Théorie des Graphes Orientées vers les Sciences Economiques et Sociales. Dunod: Paris; 1969 (1); 1970 (2).

Roy B. Partial preference analysis and decision aid: The fuzzy outranking relation concept. In: Bell DE, Keeney RL, Raiffa H (Eds), Conflicting Objectives in Decisions. John Wiley & Sons: New York; 1977. p. 40–75.

Roy B. ELECTRE III: Un algorithme de classements fondé sur une représentation flouedes préférences en présence de critères multiples. Cahiers du CERO 1978; 20; 3–24.

Roy B. Méthodologie Multicritère d'aide à la Décision. Economica: Paris; 1985.

Roy B. The outranking approach and the foundations of ELECTRE methods. In: Bana e Costa CA (Eds), Reading in multiple criteria decision aid. Springer-Verlag: Berlin; 1990. p. 155–183.

Roy B. The outranking approach and the foundations of ELECTRE methods. Theory and Decision 1991; 31; 49–73.

Roy B. Multicriteria Methodology for Decision Aiding. Nonconvex Optimization and its Applications. Kluwer Academic Publishers: Dordrecht; 1996.

Roy B, Bertier P. La methode ELECTRE II: Une methode au media-planning. In: Ross M (Eds), Operational research 1972. North-Holland: Amsterdam; 1973. p. 291–302.

Roy B, Bouyssou D. Aide Multicrit`ere `a la D´ecision: M´ethodes et Cas. Economica: Paris; 1993.

Roy B, Mousseau V. A theoretical framework for analysing the notion of relative importance of criteria. Journal of Multi-Criteria Decision Analysis 1996; 5; 145–159.

Roy B, Present M, Silhol D. A programming method for determining which Paris metro stations should be renovated. European Journal of Operational Research 1986; 24; 318-334.

Roy B, Skalka J. ELECTRE IS: Aspects méthodologiques et guide d'utilisation. Document du LAMSADE 30, Université Paris Dauphine, 1984.

Roy B, Vincke Ph. Multicriteria analysis: survey and new directions. European Journal of Operational Research 1981; 8; 207-218.

Roy B, Vincke Ph. Relational systems of preference with one or more pseudo criteria: Some new concepts and results. Management Science 1984; 30; 1323–1335.

Roy B, Vincke Ph. Pseudo-orders: Definition, properties and numerical representation. Mathematical Social Sciences 1987; 14; 263–274.

Pöyhönen M, Hämäläinen R.P. Notes on the weighting biases in value trees. Journal of Behavioral Decision Making 1998; 11; 139–150.

Pöyhönen M, Hämäläinen R.P. On the convergence of multiattribute weighting methods. European Journal of Operational Research 2001; 129; 569-585.

Proctor T. Creative problem solving for managers: Developing Skills for Decision-Making and Innovation. Routledge: London; 2005.

Sadorsky P. Oil price shocks and stock market activity. Energy Economics 1999; 21; 449−469.

Sadorsky P. Risk factors in stock returns of Canadian oil and gas companies. Energy Economics 2001; 23; 17–28.

Sanders N.R., Manrodt K.S. Forecasting practices in U.S. corporations. Interfaces 1994; 24; 92-100.

Schoemaker P.J., Waid C.C. An experimental comparison of different approaches to determining weights in additive value models. Management Science 1982; 28; 182-196.

Schwartz T.V., Szakmary A.C. Price discovery in petroleum markets: Arbitrage, cointegration and the time interval of analysis. The Journal of Futures Markets 1994; 14; 147–167.

Sequeira J.M., McAleer M. A market-augmented model for SIMEX Brent crude oil futures contracts. Applied Financial Economics 2000; 10; 543–552.

Serletis A. A cointegration analysis of petroleum futures prices. Energy Economics 1994; 16; 93–97.

Shapiro S.S., Wilk M.B. An Analysis of Variance Test for Normality (Complete Samples). Biometrika1965; 52; 591-611.

Shambora W.E., Rossitera R. Are there exploitable inefficiencies in the futures market for oil? Energy Economics 2007; 29; 18-27.

Siegel S, Tukey J.W. A Nonparametric Sum of Ranks Procedure for Relative Spread in Unpaired Samples. Journal of the American Statistical Association 1960; 55; 429-445.

Siliverstovs B. Neuman A. L'Hégaret G, von Hirschhausen C. International Market Integration for Natural Gas? A Cointegration Analysis of Prices in Europe, North America and Japan. Energy Economics 2005; 27; 603–615.

Simon P, Hillson D, Newland K. Project Risk Analysis and Management Guide. APM Publishing: Norwich; 1997.

Simos J. L'e'valuation environnementale: un processus cognitif nigocii. These de doctorat, DGF-EPFL, Lausanne.1990.

Siskos J, Spyridakos A. Intelligent multicriteria decision support: Overview and perspectives. European Journal of Operational Research 1999; 113; 236-246.

Solymosi T, Dombi J. A method for determining the weights of criteria: The centralized weights. European Journal of Operational Research 1986; 26; 35–41.

Stamatis D.H. TQM Engineering Handbook. Marcel Dekker Inc: New York; 1997.

Stephens M.A. Tests Based on Regression and Correlation. In: D'agostino RB, Stephen MA (Eds), Goodness-of-Fit Techniques. Marcel Dekker Inc: New York; 1986.

Stewart T. A critical survey on the status of multiple criteria decision making theory and practice. Omega 1992; 20; 569-586.

Stillwell W.G., Winterfeldt D, John R.S. Comparing hierarchical and non-hierarchical weighting methods for eliciting multiattribute value models. Management Science 1987; 33; 442–450.

Suriya K. Forecasting Crude Oil Price Using Neural Networks. CMU Journal 2006; 5; 377-386.

Tashman L.J. Out of sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting 2000; 16; 437-450.

Tague N.R. The Quality Toolbox. ASQ Quality Press: Milwaukee; 2005.

Thompson P.A. An MSE statistic for comparing forecast accuracy across series. International Journal of Forecasting 1990; 6; 219-227.

Torgerson W.S. Theory and methods of scaling. Wiley: New York; 1958.

Tsoukiàs A. From Decision Theory to Decision Aiding Methodology. European Journal of Operational Research 2008; 187; 138 - 161.

Tsoukiàs A, Vincke Ph. A survey on non conventional preference modelling. Ricerca Operativa 1992; 61; 5–49.

Tsoukiàs A, Vincke Ph. A new axiomatic foundation of partial comparability. Theory and Decision 1995; 39; 79–114.

Tsoukiàs A, Vincke Ph. Extended preference structures in MCDA. In: Climaco J (Eds), Multicriteria Analysis. Springer Verlag: Berlin; 1997.p. 37–50.

Tsoukiàs A, Vincke Ph. A characterization of pqi interval orders. Discrete Applied Mathematics 2003; 127; 387–397.

Van Newenhizen J. The Borda method is most likely to respect the Condorcet principle. Economic Theory 1992; 2; 69-83.

Vanderpooten D. The construction of prescriptions in outranking methods. In: Bana e Costa CA (Eds), Reading in Multiple Criteria Decision Aid, Springer-Verlag, Berlin, 1990. p. 184–215.

Vincke Ph. Analysis of multicriteria decision aid in Europe - Invited Review. European Journal of Operational Research 1986; 25; 160-168.

Vincke Ph. Multicriteria Decision-Aid. John Wiley and Sons: Chichester; 1992.

Vincke Ph. Preferences and numbers. In: Colorni A, Paruccini M, Roy B (Eds), *A-MCDA - Aide Multi Critère à la Décision - Multiple Criteria Decision Aiding.* The European Commission Joint Research Center 2001. p. 343–354.

Vind K. Independent preferences. Journal of Mathematical Economics 1991; 20; 119-135.

Virine L, Trumper M. Project Decisions: The Art and Science. Management Concepts: Virginia; 2008.

von Winterfeldt D, Edwards W. Decision analysis and behavorial research. Cambridge University Press: Cambridge; 1986.

von Winterfeldt D, Edwards W. Defining a decision analytic structure. In: Edwards W, Miles RF, von Winterfeldt D (Eds), Advances in Decision Analysis. Cambridge University Press: Cambridge; 2007. p. 81–103.

von Winterfeldt D, Fasolo B. Structuring decision problems: A case study and reflections for practitioners. European Journal of Operational Research 2009; 199; 857–866.

Wang S.Y., Yu L, Lai K.K. A novel hybrid AI system framework for crude oil price forecasting. Lecture Notes in Computer Science 2004; 3327; 233–242.

Wang S.Y., Yu L, Lai K.K. Crude oil price forecasting with Tei@I methodology, Journal of Systems Science and Complexity 2005; 18; 145–166.

Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge University Press: Cambridge; 1994.

Weber M, Borcherding K. Behavioral influences on weight judgments in multiattribute decision making. European Journal of Operational Research 1993; 67; 1-12.

Williamson D. Strategic Management and Business Analysis. Elsevier, Butterworth-Heineman: Boston; 2004.

Winklhofer H, Diamantopoulos A, Witt S.F. Forecasting practice: A review of the empirical literature and an agenda for future research. International Journal of Forecasting 1996; 12; 193– 221.

Winklhofer H, Diamantopoulos A. Managerial evaluation of sales forecasting effectiveness: A MIMIC modeling approach. International Journal of Research in Marketing 2002; 19; 151–166.

Xie W, Yu L, Xu S.Y., Wang S.Y. A new method for crude oil price forecasting based on support vector machines. Lecture Notes in Computer Science 2006; 3994; 441–451.

Ye M, Zyren J. Shore J. Forecasting crude oil spot price using OECD petroleum inventory levels. International Advances in Economic Research 2002; 8; 324–334.

Ye M, Zyren J. Shore J. A monthly crude oil spot price forecasting model using relative inventories. International Journal of Forecasting 2005; 21; 491–501.

Ye M, Zyren J. Shore J. Forecasting short-run crude oil price using high and low-inventory variables. Energy Policy 2006a; 34; 2736–2743.

Ye M, Zyren J. Shore J. Short-run crude oil price and surplus production capacity. International Advances in Economic Research 2006b; 12; 390–394.

Yokum J, Armstrong J. Beyond Accuracy: Comparison of Criteria Used to Select Forecasting Methods. International Journal of Forecasting 1995; 11; 591-97.

Younker D.L. Value Engineering: Analysis and Methodology. Marcel Dekker Inc: New York; 2003.

Yousefi S, Weinreich I. Reinarz D. Wavelet-based prediction of oil prices. Chaos, Solitons and Fractals 2005; 25; 265–275.

Yu L, Lai K.K., Wang S, He K. Oil price forecasting with an EMD-based multi-scale neural network learning paradigm. Lecture Notes in Computer Science 2007; 4489; 925–932.

Yu L, Wang S, Lai K.K. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Economics 2008; 30; 2623–2635.

Yu W. ELECTRE TRI: Aspects me thodologiques et manuel d'utilisation. Document du LAMSADE n074, Universite Paris-Dauphine ;1992a.

Yu W. Aide multicrit`re a la de cision dans le cadre de la proble matique du tri: concepts, me thodes et applications, PhD dissertation, Universite Paris-Dauphine.1992b.

Zeng T, Swanson N.R. Predictive evaluation of econometric forecasting models in commodity futures markets. Studies in Nonlinear Dynamics and Econometrics 1998; 2; 1037-1037.

Zopounidis C, Doumpos M. Building additive utilities for multi-group hierarchical discrimination: the M.H.DIS method. Optimization Methods and Software 2000; 14; 219–240.

Chapter 5:

Conclusion

Conclusion

A common issue faced by both academics and practitioners is related to the performance evaluation of competing forecasting models; to be more specific, although the performance evaluation exercise requires one to take account of several criteria at the same time, to the best of our knowledge, in the field of forecasting there is no published multidimensional framework designed for this purpose. Consequently, conflicting results about the performance of specific forecasting models are often reported in that some models perform better than others with respect to a specific criterion but worse with respect to other criteria; thus, leading to a situation where one cannot make an informed decision as to which model performs best overall by taking all criteria into account. Therefore, the aim of my PhD research is to contribute to the field of forecasting from a methodological perspective by proposing multidimensional frameworks to performance evaluation of competing forecasting models, and use crude oil as an application area to illustrate the use of the proposed frameworks. In sum, two main methodological contributions are proposed.

The first contribution consists of proposing a mathematical programming based approach, commonly referred to as Data Envelopment Analysis (DEA), as a multidimensional framework for relative performance evaluation of competing forecasting models or methods. In order to present and discuss how one might adapt this framework to measure and evaluate the relative performance of competing forecasting models, we first survey and classify the literature on performance criteria and their measures – including statistical tests – commonly used in evaluating and selecting forecasting models or methods to assist in selecting the appropriate metrics to measure the criteria under consideration. This classification will serve as a basis for the operationalisation of DEA. In sum, context-dependent DEA is proposed and its application is illustrated in evaluating and selecting models to forecast crude oil prices. 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, linear

regression models such as REG2, REG3 and REG5, Holt-Winter Exponential Smoothing with Multiplicative Seasonality and Random Walk adjusted for Trend tend to have ranks that are not sensitive to performance measures, which suggest that the rankings of these models are reliable. Furthermore, REG5, Holt-Winter Exponential Smoothing with Multiplicative Seasonality and Random Walk adjusted for Trend are superior to the remaining models. Finally, in practice, we recommend that the forecasts produced by models with similar performance such as these should be combined and compared to all forecasts produced by individual models before deciding on the forecasting model or method to implement.

The second contribution consists of proposing a Multi-Criteria Decision Analysis (MCDA) based approach as a multidimensional framework for relative performance evaluation of the competing forecasting models or methods. In order to present and discuss how one might adapt such framework, we first revisit MCDA methodology, propose a revised methodological framework that consists of a sequential decision making process with feedback adjustment mechanisms. Second, we provide guidelines as to how to operationalise it. Third, we survey and classify the literature on performance criteria and their measures – including statistical tests – commonly used in evaluating and selecting forecasting models or methods to assist in selecting the appropriate metrics to measure the criteria under consideration. Finally, we discuss how one might adapt such MCDA framework to address the problem of relative performance evaluation of competing forecasting models of crude oil prices. Three outranking methods have been used in our empirical experiments; namely, ELECTRE III, PROMETHEE I and PROMETHEE II, and our main conclusions 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, when relatively high importance weights are assigned to a specific criterion, some multidimensional outranking methods tend to produce rankings that are similar to the one obtained by a unidimensional ranking method. Third, in practice, one should use several outranking methods, compare their outputs, and eventually combine the forecasts of some of the models under consideration, which are incomparable, before deciding on

the forecasting model or method to implement. Finally, as far as the evaluation of the relative performance of the valid forecasting models considered in this study is concerned, models such as the linear regression model REG5 and Holt-Winter Exponential Smoothing with Multiplicative Seasonality tend to have ranks that are not sensitive to importance weights or outranking methods, which suggest that these models are superior.

As to future research, we aim first to apply these new methodological frameworks for performance evaluation with applications in the corporate finance, such as portfolio selection and management, and corporate performances evaluations. Second, we would like to contribute on the field of forecasting from the methodological perspective by proposing new models from artificial intelligence and/or hybrid approaches, with the applications on strategic commodities and financial markets. Last, but not the least, we are particularly interested in work on the area of combination forecast, such as proposing new approaches to combine different models and compare their performances to all forecasts produced by individual models within our multidimensional frameworks.