Forecasting Movements in Oil Spot Prices using Data Mining Methods

M.E. Malliaris [1], **A.G. Malliaris**[2]
[1]Information Systems & Supply Chain Management Dept., Loyola University Chicago, Chicago, IL, USA
[2]Economics and Finance Departments, Loyola University Chicago, Chicago, IL, USA

Abstract *-This paper uses information about several variables related to oil fundamentals to predict the direction that the spot oil price will move in the next week. The data, was downloaded from the Energy Information Administration and spans the time from 2001 through early 2016. It is divided into four periods. We look at both the variable sensitivity and the model ability to forecast. We find that the variables' relationships alter over the periods. By using two artificial intelligence methodologies, decision trees and support vector machines, and only forecasting when they agree, we have a much better chance of being correct in identifying next week's directional move of oil price.*

Keywords: *Decision trees, Support vector machines, oil direction, forecasting*

1 Introduction

The price of oil plays an important role in the U.S. economy for many reasons. First, the price of oil and its volatility influence both the producer and the consumer price indexes and other measures of inflation. Current inflation also influences expected future inflation and interest rates. Second, the oil industry is a dynamic sector of the U.S. economy for the employment it generates, the technology it develops and its impact on other sectors such as transportation, industrial products and research and development. Finally, oil related products generate substantial tax revenues, part of which finance the U.S. highways system. Thus, understanding what moves oil prices and being able to forecast future trends has received great attention. Representative papers that document oil forecasting research can be seen in [1, 2, 3, 4, 5, 6, 7]. In this paper, we are similarly interested in forecasting the price of oil, but our interest is not in terms of time series methodology but rather the use and evaluation of data mining techniques. We also use structural breaks, or periods, to retrain the models.

2 Data and Periods

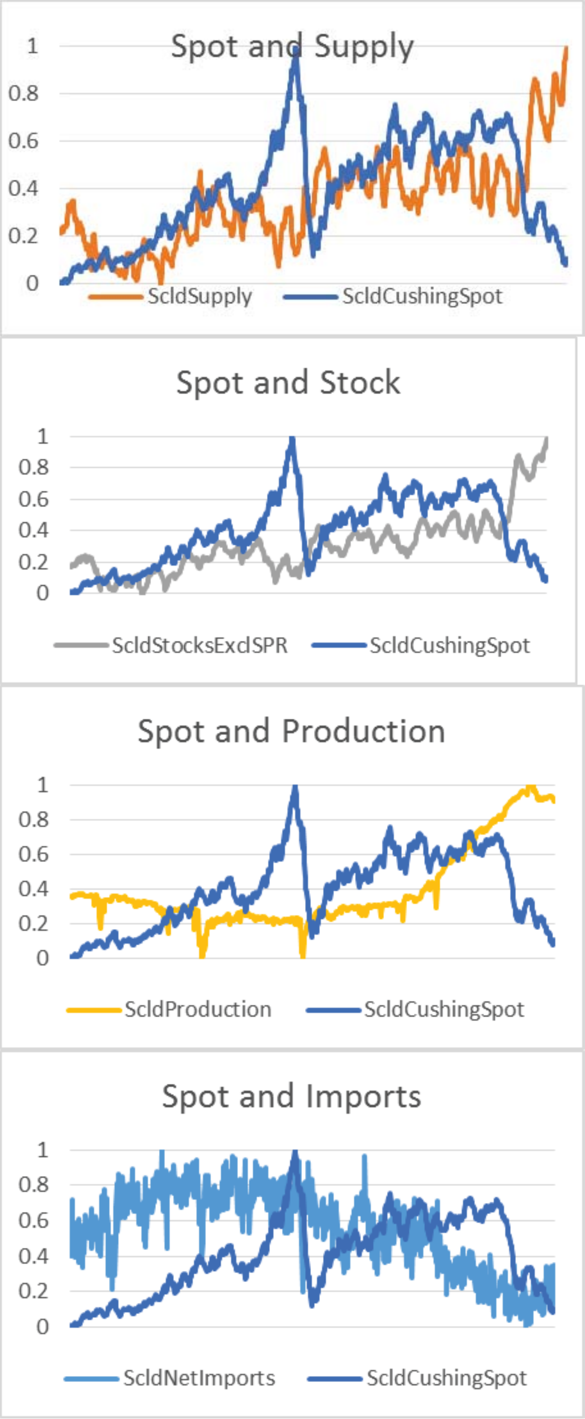
The entire data set spans the period from mid-November 2001 through the end of February 2016 and all values are weekly. Values for the five variables were downloaded from the Energy Information Administration (EIA) at http://www.eia.gov/dnav/pet/pet_sum_sndw_dcus_nus_w.htm. These variables are all based on oil and include a spot price for oil and proxies for the supply and demand for U.S. oil. The spot price is for West Texas Intermediate crude oil from Cushing Oklahoma. Demand is rather stable and changes slowly over time and we use the stock of oil as a proxy. When demand changes for seasonal reasons (summer traveling) stocks are drawn down to meet the increased demand. The stock of oil, excluding the strategic petroleum reserves, includes the inventories stored for future use and is reported in thousands of barrels on the last day of the week. Net imports are in thousands of barrels per day and include oil from the 50 states, the District of Columbia and U.S. possessions and territories. Crude oil supply is in thousands of barrels. Its components include field production, refinery production, imports, and net receipts calculated on a PAD district basis. Production is in thousands of barrels per day. The quantities are estimated by state and summed to the PADD and then the U.S. level. The paths of these variables are shown in Figure 1.

Table 1. Input and Target Variables

Derived Variables	**Role**	**Example Value**
ScldNetImports	Input	0.595
ScldSupply	Input	0.386
ScldStocksExclSPR	Input	0.324
ScldProduction	Input	0.234
ScldCushingSpot	Input	0.411
PerChgNetImp	Input	-0.015
PerChgSupply	Input	-0.013
PerChgStocksExclSPR	Input	-0.008
PerChgProduction	Input	0.002
PerChgCushingSpot	Input	0.011
DirNetImports	Input	Down
DirSupply	Input	Down
DirStocksExclSPR	Input	Down
DirProduction	Input	Up
DirCushingSpot	Input	Up
CushDirTp1	Target	Down

Using these five base variables, derived variables and a target were constructed. The derived variables included, for each of the base variables, a value scaled to be between zero and one, the percent change of the variable from week to week, and the direction the variable moved from week to week. The target variable for both model types is the direction that oil will move next week. All variables used, and their names, are shown in Table 1.

Figure 1. Cushing Oil Spot Price vs each other Variable, Scaled



This data set was divided into four periods, based on places where structural breaks have occurred, and models were built for each. Structural breaks in forecasting oil have been found important for accuracy [8, 9]. The periods are named Before, Crisis, After, and Recent. The dates used for each of the four periods are shown in Table 2. In the initial period, oil shows an overall steady positive increase in price. The Crisis period saw a steep increase in the spot price followed by an even steeper decrease. After the crisis, oil prices showed volatility with an overall path of a slight increase. In the Recent period, prices show a decreasing trend.

Table 2. Data Set time periods

Name	**Beginning**	**Ending**	**#Weeks**
Before	11/16/2001	8/31/2007	303
Crisis	9/7/2007	6/26/2009	95
After	7/3/2009	12/26/2014	287
Recent	1/2/2015	2/26/2016	61

Table 3. Correlations between variables per period, negative cells shaded

Before	***ScldNtImp***	***ScldSup***	***ScldStks***	***ScldPrd***
ScldNetImports	1.000			
ScldSupply	0.075	1.000		
ScldStocksExclSPR	0.309	0.887	1.000	
ScldProduction	-0.371	-0.449	-0.413	1.000
ScldCushingSpot	0.528	0.557	0.671	-0.787
Crisis				
ScldNetImports	1.000			
ScldSupply	-0.323	1.000		
ScldStocksExclSPR	-0.258	0.925	1.000	
ScldProduction	0.136	0.318	0.569	1.000
ScldCushingSpot	0.207	-0.698	-0.687	-0.261
After				
ScldNetImports	1.000			
ScldSupply	-0.089	1.000		
ScldStocksExclSPR	-0.419	0.600	1.000	
ScldProduction	-0.811	0.017	0.633	1.000
ScldCushingSpot	-0.309	0.108	0.270	0.346
Recent				
ScldNetImports	1.000			
ScldSupply	0.218	1.000		
ScldStocksExclSPR	0.302	0.894	1.000	
ScldProduction	-0.299	-0.137	-0.018	1.000
ScldCushingSpot	-0.616	-0.327	-0.349	0.655

Each of the periods contains a shift in the relationships among the variables. Table 3 indicates the positive and negative correlations in each period, with the negative correlations shaded. Notice, for example, that the correlation between the Cushing Spot price and Supply is positive in the Before period, negative in the Crisis period, positive in the After period, and negative in the Recent period. In order for a single model to remain stable and robust over time, the relationships among the variables also need to be stable. Because each of these variables has an impact on the price of oil, but this impact shifts throughout the periods, we build separate models on each period. After building these models, we look at the impact of the most important variables in each of the periods, and compare the forecasting results for the oil spot price movement.

3 Methods

Two data mining methodologies are used for this paper, decision trees and support vector machines. All models were built using IBM's SPSS Modeler 17.0 data mining software. A decision tree works in a stepwise fashion to successively divide the data into parts that are more single-valued on the target variable. In the beginning, all data is held in the root node. All models were built using IBM's SPSS Modeler 17.0 data mining software.

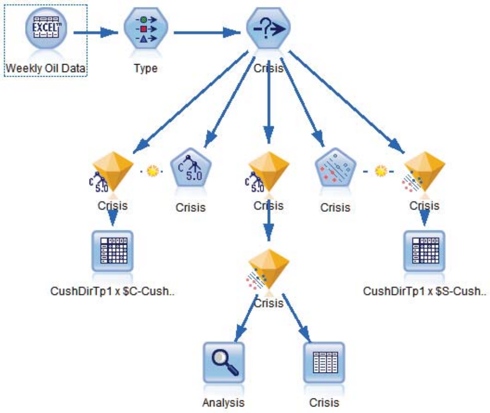
For the decision tree, the C5.0 algorithm was applied. This algorithm first splits the data set on the input field that gives the greatest information gain. That is, that provides the split with resulting groups that have higher percentages of a single value of the target than the original node did. Each of the nodes resulting from the first split are then split again, typically using a different field. This process continues until the nodes cannot have any possible split that could improve the model. Each terminal node forms a path from the root that describes a subset of the training data and generates a single value of the target for any new data that conforms to the path.

The second model used is the support vector machine methodology. This methodology applies a transformation to the input data that enables the two values of the target variable (in this case, Up and Down) to be separated by a plane so that each side of the plane has a single target value. Each row in the data set is graphed in n-dimensional space. Initially, the target variable values are not divisible. By applying a transformation that moves the data to an n+1 dimensional space, the target values form an arrangement that is separable. Modeler offers a choice of four transformations to the input data: radial basis function, polynomial, sigmoid, and linear. In this case, the polynomial transformation yielded the best results.

After each model has finished training, a sensitivity analysis is performed on all the variables. This generates a relative set of values for predictor importance. Various values of one variable are fed through the trained model while all other variable values are held fixed. The impact on the target variable is noted. This is repeated for each variable. Then, all variables are ranked according to the relative change to the target using the trained model. These predictor importance values range all sum to 1. Comparing these across models allows us to see the difference in how each of the models values the input of each of the variables, and how the variables impact the final target value.

Figure 2 illustrates the IBM Modeler data mining stream for the Crisis data period. The data set is read in with an Excel node and flows to a Type node where each field is assigned its role of input or target. The data for a particular period is then selected in the hexagon-shaped Select node, and sent to the pentagon-shaped modeling nodes, C5.0 and SVM, for training. Within the model training nodes, settings the affect the training run can be specified. This training process generates two diamond-shaped trained models which are shown as nuggets connected to their original training model. The results from the trained models are then sent to matrix nodes for evaluation of the output. In addition, copies of the trained models can be connected in succession, as shown here in the center of the figure. The data is sent through each of the trained models to an analysis node where the comparative forecasts can be studied. The data is also sent to a table node so that the forecasts can be exported to Excel for further analysis if desired.

Figure 2. Structure of the Modeler stream for one data set.



3.1 Decision tree results

The decision trees were developed separately for each of the four data sets. Figure 3 shows the trained decision tree maps for each of these data set periods. The Before and After maps are wider and deeper, but these data sets are also much larger than the Crisis and Recent sets are.

More important than just the complexity of the tree is how well they did in forecasting and explaining the paths to the forecasts. The results for forecasting accuracy are detailed in Table 4.

Figure 3. Decision Tree maps for each of the data sets

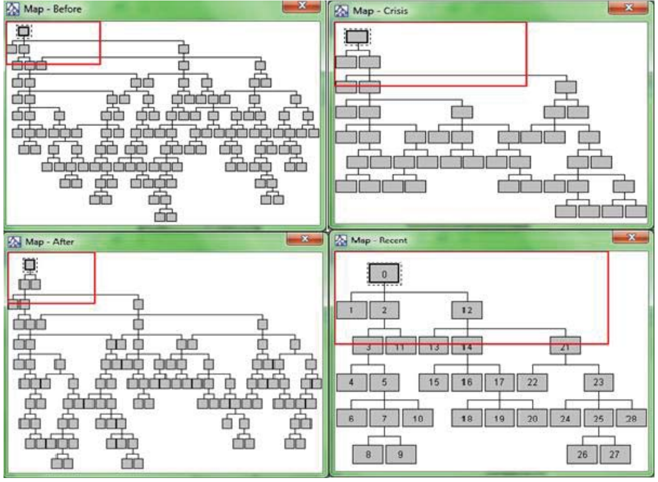


Table 4. Decision tree forecasting accuracy, correct percentages in bold.

		Actual Dir		Actual Dir		
Period		Down	% Cor	Up	% Cor	Total
Before	Dn	124	**89.2**	15		139
	Up	6		158	**96.3**	164
Crisis	Dn	40	**95.2**	2		42
	Up	6		47	**88.7**	53
After	Dn	135	**93.8**	9		144
	Up	8		135	**94.4**	143
Recent	Dn	33	**89.2**	4		37
	Up	2		22	**91.7**	24

In this table, we see that the forecasting accuracy for predicting that oil would move Down the next week varied from a low of 89.2% to a high of 95.2%. For the times that Up was predicted the worst period was correct 88.7% of the time, and the best was correct 96.3% of the time. The models are doing well in all periods and well in each direction of the forecast. We emphasize that these are "in sample" performances that do not guarantee similar results for "out of sample" performances. However, this forecasting accuracy demonstrates that the two techniques used were successful in identifying across 4 regimes the appropriate variables determining the in sample performance.

Table 5 shows the relative importance of each of the variables to the model trained in that period. These relative importance values sum to 1 within each model and are not related to model accuracy, but just to the way the trained model used the variables to forecast the target variable, the direction that oil would move in the following week. Blank cells indicate that the variable was not used. The top five variables for each model are shown in bold, and the most important variable is shaded. Since we saw in Table 3 that the correlations between the variables vary over time, we would expect variables to go in and out of importance in Table 5.

Table 5. Relative variable importance.

Variable	DT Before	DT Crisis	DT After	DT Recent
DirCushingSpot	**0.1242**	**0.1062**		0.0196
DirNetImports	**0.1126**	0.0541	**0.1247**	**0.3594**
DirProduction	0.0341	**0.2725**	**0.1234**	**0.1276**
DirStocksExclSPR	**0.0898**	0.0479	0.0580	0.0222
DirSupply	0.0881	0.0585	0.0214	0.0637
PerChgCushingSpot			0.0037	**0.1444**
PerChgNetImp	0.0635	**0.0775**	0.0000	**0.1149**
PerChgProduction		**0.1066**	**0.1429**	0.0099
PerChgStocksExclSPR	**0.0937**			
PerChgSupply	0.0484	**0.1905**	0.1117	
ScldCushingSpot	0.0498		**0.1441**	
ScldNetImports	0.0504	0.0657	0.0425	**0.1290**
ScldProduction	0.0396		**0.1356**	0.0093
ScldStocksExclSPR	0.0875	0.0205	0.0836	
ScldSupply	**0.1183**		0.0084	

The most important variable for each period is different, oil's direction is top in the Before period, the change in supply matters most during the Crisis period, After the crisis the change in production tops the importance list, and in the Recent period, the direction of net imports has over a third of the relative importance. We see similarities in the Before and After sets in that they use almost all of the variables. In the Crisis and Recent sets, a third of the variables are not used at all by the model.

Another way to think about the impact of an input is by summing the individual importance values of each variable related to the same base variable. For example, the three variables derived from the price of Cushing oil are oil's direction, oil's percent change in price, and its value as scaled from 0 to 1. These are represented by DirCushingSpot, PerChgCushingSpot, and ScldCushingSpot. Combining the variables in this way reduces the number of variables to four and allows us to see their overall impact regardless of the form in which they were presented to the models. Table 6 gives these summed values for the inputs related to each base variable.

When looked at as summed values we see that stocks of oil (excluding the strategic petroleum reserves) had the greatest overall influence in the Before period. Oil production was most important in both the Crisis and After periods, and net imports climbed to the top spot with an overall relative importance of .6 in the Recent period.

Table 6. Sum of Dir, PerChg and Scld for each of the base variables, per period.

Variable	DT Before	DT Crisis	DT After	DT Recent
CushingSpot	0.1740	0.1062	0.1478	0.1640
NetImports	0.2265	0.1973	0.1672	**0.6033**
Production	0.0737	**0.3791**	**0.4019**	0.1468
StocksExclSPR	**0.2710**	0.0684	0.1416	0.0222
Supply	0.2548	0.2490	0.1415	0.0637

3.2 Support vector machine results

The support vector machines used a polynomial kernel, with the regularization parameter C set to 10 and epsilon at 0.1. The model accuracy on each period is shown in Table 6. Accuracy on weeks where the direction of oil was Down varied from a low of 90.8% to a high of 100%. For weeks where oil moved Up, accuracy varied from 92.6% to 100%. Lower accuracies occurred in the larger data sets.

Table 6. Support vector machine forecasting accuracy, correct percentages in bold.

		Actual Dir		Actual Dir		
		Down	%Cor	Up	% Cor	Total
Before	Dn	118	**90.8**	12		130
	Up	12		161	**93.1**	173
Crisis	Dn	45	**100**	0		45
	Up	1		49	**98**	50
After	Dn	132	**95**	7		139
	Up	11		137	**92.6**	148
Recent	Dn	35	**100**	0		35
	Up	0		26	**100**	26

The accuracies are higher using the support vector machine than the decision trees and even hit 100% in three of the eight summary cells. Next, we look at the use of the variables by the models. Table 7 lists the relative importance of the variables in each model. Again, the top five variables are in bold and the highest variable is shaded. The variable with the most impact in determining the direction that oil will move next week in the Before set is the percent change in the value of net imports; for the Crisis data, it moves to the direction of the Cushing spot price; in the After period it switches to the percent change in the Cushing spot price, and in the Recent period, production has over twenty-five percent of the importance. None of the most important single variables are the same in the support vector machines as they were in the decision trees, an indication that the way the models are making their decisions is very different. If we look at these variables in a condensed format, obtained by summing the variables built on the same base variables, we get the results shown in Table 8. Here we have the most important variable in bold and shaded. Net imports came in first in the Before period, Cushing spot price in both the Crisis and After periods, and Production in the Recent period.

Table 7. Relative variable importance.

Variable	SVM Before	SVM Crisis	SVM After	SVM Recent
DirCushingSpot		**0.1323**	0.0149	0.0552
DirNetImports		**0.1158**	0.0369	**0.1235**
DirProduction		0.0905	0.0506	**0.2584**
DirStocksExclSPR	0.0454	**0.0912**	0.0065	0.0493
DirSupply	0.0240	**0.0957**	0.0069	0.0257
PerChgCushingSpot	**0.1055**	0.0737	**0.1737**	0.0614
PerChgNetImp	**0.1875**	0.0692	0.0299	**0.1174**
PerChgProduction	0.0447	**0.0944**	**0.0729**	
PerChgStocksExclSPR		0.0487	**0.1541**	0.0575
PerChgSupply	**0.1263**	0.0707	**0.1000**	**0.0659**
ScldCushingSpot	0.0506	0.0444	**0.1456**	**0.0829**
ScldNetImports	**0.1691**	0.0169	0.0675	0.0352
ScldProduction	**0.1152**	0.0089	0.0438	0.0197
ScldStocksExclSPR	0.0406	0.0158	0.0469	
ScldSupply	0.0911	0.0318	0.0498	0.0479

None of these match the most important base variables in the decision tree models. Like the decision tree models, in both the Crisis and After periods, the most important base variable was the same.

Table 8. Table 6. Sum of Dir, PerChg and Scld for each of the base variables, per period.

Variable	SVM Before	SVM Crisis	SVM After	SVM Recent
CushingSpot	0.1561	**0.2504**	**0.3342**	0.1995
NetImports	**0.3566**	0.2019	0.1343	0.2761
Production	0.1599	0.1938	0.1673	**0.2781**
StocksExclSPR	0.0860	0.1557	0.2075	0.1068
Supply	0.2414	0.1982	0.1567	0.1395

3.3 Both methods

We have seen in the previous sections that the decision tree models and the support vector machine models both do well, but they value and use the variables in different ways. Neither model is always right across all periods. It would be of interest to see whether they are wrong at the same times. Figure 4 shows a graph for each period of when each of the models is wrong. The horizontal axis in each case spans from the beginning to the ending date of the period. We see that some of the time, both models are wrong, however much of the time, these incorrect forecasts occur in different weeks.

Figure 4. Dates when each of the models predicted the wrong direction

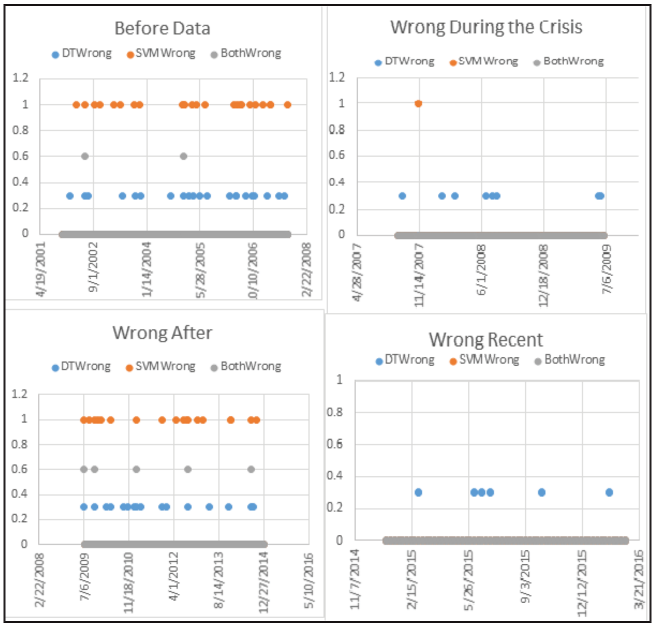


Table 8 summarizes the percent of time that the individual models are correct and are wrong, and, when they agree, the percent of time that they are both correct. We can see that, in the Before period, agree occurs 262 out of 303 weeks. In these 262 weeks, the models are correct 260 out of the 262 time, that is, 99.24% of the time.

During the Crisis period, there is agreement 86 out of 95 weeks. When the models agree, they are correct always. In the After period, the models agree 262 out of 287 weeks and, when such agreement occurs, are correct 98.09% of the time. In the last period, Recent, the models agree 55 out of 61 weeks, and are correct 90.16% of the agreement time.

Thus, in all periods but the last one, we improve our forecasting accuracy by looking only at the weeks when both of the models agree. This gives us an indication that we might be more successful in trading during weeks when the model signals are in agreement.

Table 8. Model agreement.

	DT alone	SVM alone	When Agreed	#Weeks Agreed
Before: 303 weeks				
Correct	93.07%	92.08%	99.24%	260
Wrong	6.93%	7.92%	0.76%	2
Crisis: 95 weeks				
Correct	91.58%	98.95%	100%	86
Wrong	8.42%	1.05%	0%	0
After: 287 weeks				
Correct	94.08%	93.73%	98.09%	257
Wrong	5.92%	6.27%	1.01%	5
Recent: 61 weeks				
Correct	90.16%	100.0%	90.16	55
Wrong	9.84%	0.00%	9.84	0

4 Conclusions

The purpose of this study is to explore two questions: what determines the forecastability of oil prices, and whether two data mining models that "think" very differently can be used to generate a better forecast that either model alone. In contrast to market efficiency that proposes that financial markets rationally collect all relevant fundamental data to accurately determine spot prices without any ability to forecast future prices, this study explores two strategies for understanding the formation of the direction of future prices a week ahead. This study used in-sample data in the analysis. The models help to identify the relative importance of the inputs.

The study considers four regimes and finds that the economic fundamentals play a different role in each regime. In addition, it makes use of two different methodologies to assess the changing role of fundamentals. The high accuracy of the in-sample forecastability indicates that the existing structure among fundamentals was successfully detected by the two methods used and actually by jointly employing these two methods, the overall results improve further.

5 References

[1] Baumeister, C., & Kilian, L. (2014). What central bankers need to know about forecasting oil prices, International Economic Review, 55(3), 869-889.

[2] Baumeister, C., Kilian, L. (2015) Forecasting the Real Price of Oil in a Changing World: A Forecast Combination Approach, Journal of Business & Economic Statistics, 33(3), pp 338-351.

[3] Chen, S. (2014) Forecasting Crude Oil Price Movements with Oil-Sensitive Stocks, Economic Inquiry, 52(2), pp 830-844.

[4] Jammazi, R., Aloui, C. (2012) Crude oil price forecasting: Experimental evidence from wavelet decomposition and neural network modeling, Energy Economics, 34(3), pp 828-841.

[5] Shin, H., Hou, T., Park, K., Park, C., and Choi, S. (2013) Prediction of movement direction in crude oil prices based on semi-supervised learning, Decision Support Systems, 55(1), pp 348-358.

[6] Tsai, C. (2015) How do U.S. Stock Returns Respond Differently to Oil Price Shocks Pre-Crisis, Within the Financial Crisis, and Post-Crisis?, Energy Economics, 50, pp 47-62.

[7] Xiong, T., Bao, Y., Hu, Z (2013) Beyond one-step-ahead forecasting: Evaluation of alternative multi-step-ahead forecasting models for crude oil prices, Energy Economics, 40, pp 405-415.

[8] Noguera, J. (2013) Oil Prices: Breaks and Trends, Energy Economics, 37, pp. 60-67.

[9] Salisu, A., & Fasanya, I (2013) Modelling Oil Price Volatility With Structural Breaks, 52, pp. 554-562.