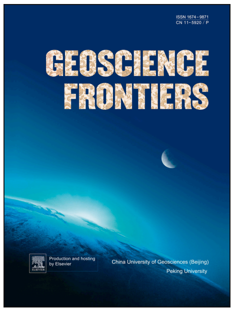
Accepted Manuscript



A new approach for crude oil price prediction based on stream learning

Shuang Gao, Yalin Lei

PII: S1674-9871(16)30086-X

DOI: 10.1016/j.gsf.2016.08.002

Reference: GSF 483

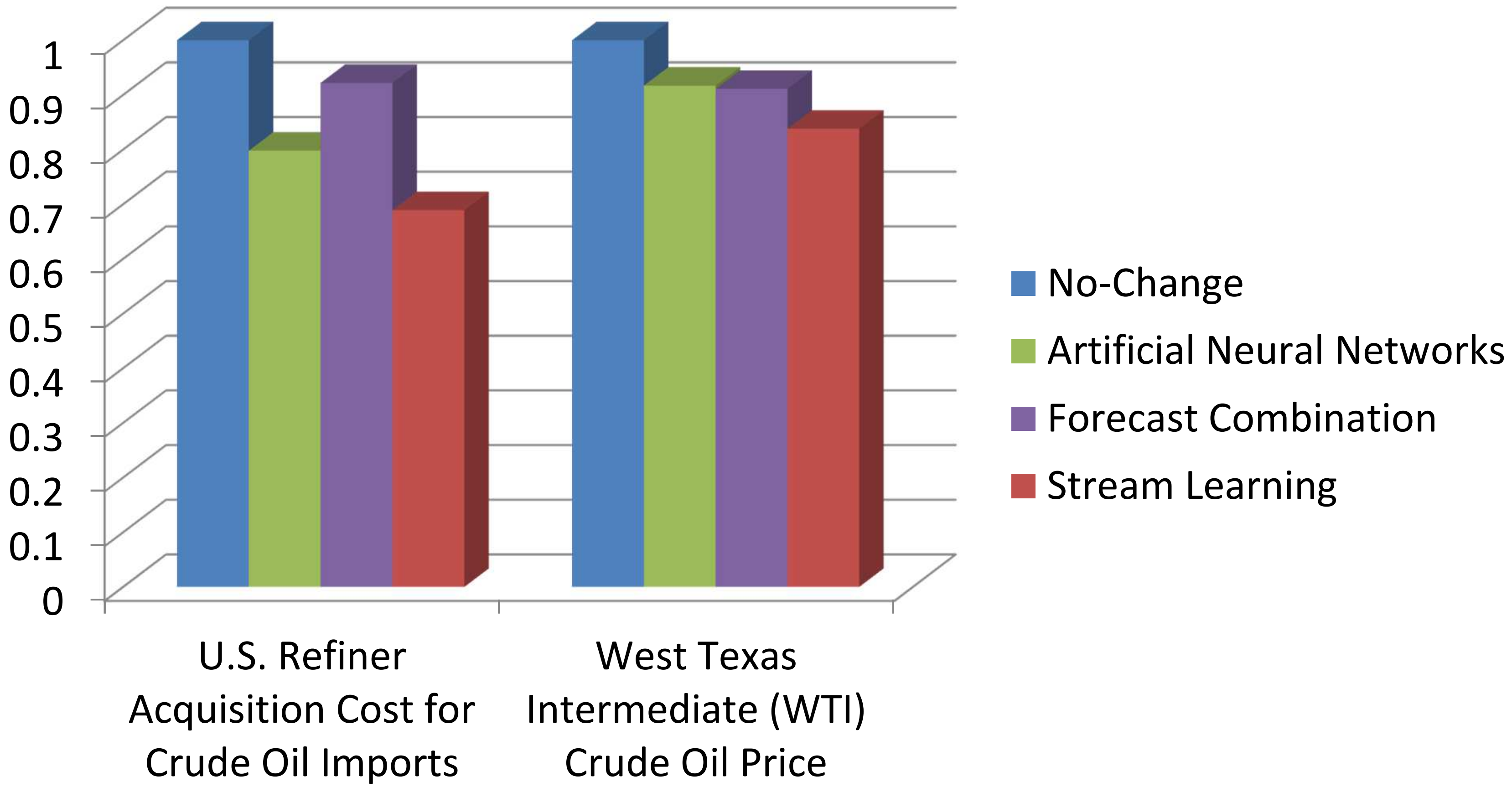
To appear in: *Geoscience Frontiers*

Received Date: 27 May 2016

Revised Date: 30 July 2016

Accepted Date: 4 August 2016

Please cite this article as: Gao, S., Lei, Y., A new approach for crude oil price prediction based on stream learning, *Geoscience Frontiers* (2016), doi: 10.1016/j.gsf.2016.08.002.


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
U.S. Refiner Acquisition Cost for Crude Oil Imports
West Texas Intermediate (WTI) Crude Oil Price
No-Change
Artificial Neural Networks
Forecast Combination
Stream Learning

A new approach for crude oil price prediction based on stream learning

Shuang Gao , Yalin Lei*

School of Humanities and Economic Management, China University of Geosciences, Beijing 100083, China

*

* Corresponding Author (Email: leiyalin@cugb.edu.cn)

Abstract

Crude oil is the world's leading fuel, and its prices have a big impact on the global environment, economy as well as oil exploration and exploitation activities. Oil price forecasts are very useful to industries, governments and individuals. Although many methods have been developed for predicting oil prices, it remains one of the most challenging forecasting problems due to the high volatility of oil prices. In this paper, we propose a novel approach for crude oil price prediction based on a new machine learning paradigm called stream learning. The main advantage of our stream learning approach is that the prediction model can capture the changing pattern of oil prices since the model is continuously updated whenever new oil price data are available, with very small constant overhead. To evaluate the forecasting ability of our streaming learning model, we compare it with three other popular oil price prediction models. The experiment results show that our stream learning model achieves the highest accuracy in terms of both mean squared prediction error and directional accuracy ratio over a variety of forecast time horizons.

Keywords

crude oil; economic geology; prediction model; machine learning; stream learning

1. Introduction

Crude oil is a natural liquid fossil fuel found in geological formations beneath the earth's surface. It has mostly been extracted by oil drilling, which comes after the studies of structural geology, sedimentary basin analysis, and reservoir characterization (Guerriero et al., 2012). Crude oil is one of the most important energy resources on earth. So far, it remains the world's leading fuel, with nearly one-third of global energy consumption.

Crude oil prices are determined by many factors and have a big impact on the global environment and economy. Although crude oil prices were firm in early 2014, they fell sharply from mid 2014. In January 2016, the U.S. refiner acquisition cost for crude oil imports, as a proxy for world oil price, is only $28.81 per barrel on average, and the West Texas Intermediate (WTI) crude oil spot price, as the benchmark oil price in North America, is only $31.68 per barrel on average (EIA, 2016). The prices have dropped by more than seventy percent since June 2014.

The world's environment is affected by the oil price falling. With the drop of oil prices, the fuel bills are lowered. As a result, consumers are very likely to use more oil and thus increase the carbon emission. In addition, there is less incentive to develop renewable and clean energy resources. On the other hand, sustained low oil prices could lead to a drop in global oil and gas exploration and exploitation activities.

Fluctuating oil prices also play an important role in the global economy (Husain et al., 2015). The fall in oil prices would result in a modest boost to global economic activity, although the owners of oil sectors suffer income losses. Recent research from the World Bank shows that for every 30% decline of oil prices, the global GDP (Gross Domestic Product) would be increased by 0.5%. At the same time, the drop of oil prices would reduce the cost of living, and hence the inflation rate would fall.

There is no doubt that crude oil price forecasts are very useful to industries, governments as well as individuals. Thus, forecasting crude oil prices has been the subject of research by both academia and industry. Many methods and approaches have been developed for predicting oil prices. However, due to the high volatility of oil prices (Regnier, 2007), it remains one of the most challenging forecasting problems.

In recent years, machine learning techniques have been used in many applications in geosciences (Alavi et al., 2016). Machine learning provides powerful computational tools and algorithms that can learn from and make predictions on data. In this paper, we propose a novel approach for crude oil price prediction based on a new machine learning paradigm called stream learning (Gama et al., 2009; Bifet et al., 2010). The main advantage of our stream learning approach is that the prediction model can capture the changing pattern of oil prices since the model is continuously updated whenever new oil price data are available, with very small constant overhead. We compare our streaming learning model with three other popular oil price prediction models for predicting two types of oil prices (the U.S. refiner acquisition cost for crude oil imports and the WTI crude oil spot price). The experiment results show that our stream learning model achieves the highest accuracy in terms of both mean squared prediction error and directional accuracy ratio over a variety of forecast time horizons.

2. Methods

2.1. Literature review

We divide crude oil price forecasting approaches into three categories: 1) heuristic approaches; 2) econometric models; and 3) machine learning techniques.

Heuristic approaches for oil price prediction include professional and survey forecasts, which are mainly based on professional knowledge, judgments, opinion and intuition. Another heuristic approach, the so-called no-change forecast, uses the current price of oil as the best prediction of future oil prices. Despite its simplicity, the no-change forecast appeared to be a good baseline approach for oil price prediction and was better than other heuristic judgmental approaches (Alquist et al., 2013).

Econometric models are the most widely used approaches for oil price prediction, which include autoregressive moving average (ARMA) models and vector autoregressive (VAR) models, with possibly different input variables (Pindyck, 1999; Frey et al., 2009). These econometric models provide more accurate prediction than the no-change model at least at some horizons (Alguist et al., 2013; Baumeister and Kilian, 2015). Recently, a forecast combination approach was proposed by Baumeister and Kilian (2015), which combines 6 different oil price prediction models including both econometric models (such as the VAR model) and the no-change model. It should be noted that most of the econometric models are linear models and are not be able to capture the nonlinearity of oil prices.

Several machine learning techniques were proposed for oil price prediction, such as artificial neural networks (ANN) (Yu et al., 2008; Kulkarni and Haidar, 2009), and support vector machine (SVM) (Xie et al., 2006). These are nonlinear models which may produce more accurate predictions if the oil price data are strongly nonlinear (Behmiri and Pires Manso, 2013). However, these machine learning techniques, like other traditional machine learning techniques, rely on a fixed set of training data to train a machine learning model and then apply the model to a test set. Such an approach works well if the training data and the test data are generated from a stationary process, but may not be effective for non-stationary time series data such as oil price data.

2.2. A new stream learning approach

In this paper, we proposed a new stream learning approach for oil price prediction. Unlike traditional machine learning algorithms that use "one-shot" data analysis and focus on homogeneous and stationary data, stream learning algorithms have been developed to handle applications where continuous data streams are generated from non-stationary processes (Gama et al., 2009; Bifet et al., 2010).

Under the traditional supervised machine learning framework, one typically splits labeled data examples into a training set and a test set. The training set is used to train a machine learning model using a machine learning technique such as ANN or SVM. The test set then is used to test the performance of the machine learning model (note that sometimes there is also a development set which is used to tune the parameters of the machine learning model). For such an approach to be useful, there is an underlying assumption that the data examples in the training set and in the test set are homogeneous (e.g., generated from a stationary process), so that the trained machine learning model can capture the pattern of the data examples in the test set and produce accurate predictions for them. However, oil price time series data are not stationary, and thus traditional machine learning approaches may not produce accurate predictions.

In recent years, a new machine learning paradigm called stream learning has emerged to handle real-world applications where 1) there is a continuous flow of data as opposed to a fixed sample of independent and identically distributed (i.i.d.) examples, and 2) the data are generated by a non-stationary process instead of a stationary process (potentially at very high speed). Examples of stream learning applications include social networks, web mining, scientific data, financial data, etc.

Suppose we start with a set of initial training data which include a sequence of historical oil prices and the associated input data vectors, denoted by $\boldsymbol{D} = \{(\boldsymbol{x}_{-m+1}, y_{-m+1}), (\boldsymbol{x}_{-m+2}, y_{-m+2}), ..., (\boldsymbol{x}_{-1}, y_{-1}), (\boldsymbol{x}_0, y_0)\}$, where y_t is the oil price at time slot t, and $\boldsymbol{x}_t$ is a vector of input data variables for predicting y_t. In this paper, we use the oil prices of the previous o time slots to predict the oil price of time slot t, i.e., $\boldsymbol{x}_t = (y_{t-o}, ..., y_{t-2}, y_{t-1})$, where o ranges from 1 to 10. Our framework is applicable to general time slot unit (also called forecast time horizon), which could be daily, weekly, monthly, quarterly, etc.

Suppose we want to predict the oil prices for the next n time slots, denoted by y_1, y_2, ..., y_n and again let $\boldsymbol{x}_t$ be input data vector for predicting y_t, t=1,..., n. We propose the following stream learning procedure for predicting oil prices.

Streaming learning procedure

1. Use the data in the initial training set $\boldsymbol{D}$ to train a machine learning model, denoted by $\boldsymbol{M}_1$, and use $\boldsymbol{M}_1$ to predict the oil price for time slot 1.
2. For time slot t, t = 2, ..., n:

   Add ($\boldsymbol{x}_{t\text{-}1}$, $y_{t\text{-}1}$) to the training set $\boldsymbol{D}$ and update the machine learning model, denoted by $\boldsymbol{M}_t$, and use $\boldsymbol{M}_t$ to predict the oil price for time slot t.

Our stream learning approach is a supervised machine learning method which uses a set of labeled training data to train an initial model. The main features and advantages of using stream learning for oil price prediction include:

1) The machine learning model will be updated whenever new oil price data are available, so the model continuously evolves over time, and can capture the changing pattern of oil prices.

2) For non-stationary time series data such as oil prices, a forgetting mechanism (e.g., sliding windows, fading factors) will be deployed when updating the machine learning model.

3) Updating the model requires only a small constant time per new data example, as opposed to re-training the model using the entire training data set.

MOA (Massive Online Analysis) is an open-source framework software that allows to build and run experiments of machine learning or data mining on evolving data streams (Bifet et al., 2010; Bifet et al., 2012). We use MOA to develop stream learning models for oil price prediction. We have tried several machine learning models, including linear regression, perceptron and regression trees. To tune the parameters of a specific model in searching for the optimal solution, we split the initial training data into 90% as the training set and 10% as the development set. The development set is used to tune the model parameters, which is a standard approach in machine learning. We find that peceptron achieves the best accuracy for predicting oil prices and hence we use it as the core machine learning method within MOA.

2.3. Data transformation

A stationary process is a stochastic process whose joint probability distribution does not change when shifted in time (Priestley, 1988). Therefore, parameters of a stationary process such as the mean and variance do not change over time and do not follow any trends. A weaker form of stationarity is known as weak-sense stationarity, or wide-sense stationarity, which only requires that the mean and autocovariance do not vary with respect to time.

Under the traditional machine learning framework, to build a statistically sound machine learning model, the data time series need to be stationary or at least weakly stationary over the whole training and evaluation period. However, the oil price time series are not stationary. Under our stream learning approach, we no longer require that the oil price time series be stationary over the whole training and evaluation period, because the machine learning model will be continuously updated over time to capture the changing pattern of oil price time series. Nevertheless, when we update the machine learning model and apply the model to predict the oil price for the next time slot, there is still an underlying assumption that oil prices are stationary over the relatively shorter time period covering the current training data time slots and the next time slot.

In statistics, data transformation techniques are usually applied so that the data are more likely to meet the assumptions of a statistical inference procedure. For time series data, it is common to difference the data to improve stationarity (Priestley, 1988; Witten, 2011). Therefore, we apply the following data transformation technique for the oil price time series:

$$y'_t = y_t - y_{t-1}$$

and we find that it improves the accuracy of oil price prediction models.

3. Results and discussion

In this section, we evaluate the performance of our streaming learning model and compare it with three popular oil price prediction models.

3.1. Data and other prediction models

We use two types of oil price data to evaluate the accuracy of different oil price prediction models. The first one is the U.S. refiner acquisition cost for crude oil imports, which is the weighted average cost of all oil imported into the U.S. It can be viewed as a proxy for world oil price. The second one is the WTI crude oil spot price, which is used as a benchmark in oil pricing in North America, and is commonly cited in the press. Both data can be obtained from the U.S. Energy Information Administration (EIA) website (EIA, 2016).

We compared our streaming learning model with three other oil price prediction models. The first one is the no-change model, which is a commonly used heuristic model for oil price prediction. The second one is a model trained with artificial neural networks (ANN), which is a classical machine learning model for oil price prediction (Yu et al., 2008; Kulkarni and Haidar, 2009). The third one is the forecast combination model proposed by Baumeister and Kilian (2015), which is a state-of-the-art econometric oil price prediction model.

To have a robust evaluation, we applied the same long evaluation period from January 1992 to September 2012 as in Baumeister and Kilian (2015). For a fair comparison, we used the same sets of training data and test data for evaluating our stream learning method and the other two state-of-the-art supervised learning methods (the ANN model and the forecast combination model). Specifically, for predicting WTI crude oil price, we have used oil prices from January 1982 to December 1991 as training data, and oil prices from January 1992 to September 2012 as test data. For predicting U.S. refiner acquisition cost for crude oil imports, we have used oil prices from January 1974 to December 1991 as training data, and oil prices from January 1992 to September 2012 as test data.

The forecast include 1-month, 3-month, 6-month, 9-month and 12-month time horizons. Note that monthly oil prices are calculated by EIA based on daily data by taking an unweighted average of the daily closing spot prices for a given product over the specified time period. In Fig. 1 we plot monthly WTI crude oil spot price (both original oil price time series and time series after difference transformation) from January 1986 to January 2016.

3.2. Performance metrics

We used two standard performance metrics in the oil price prediction literature for comparing different oil price prediction models. The first metric is Mean Squared Prediction Error (MSPE). MSPE of a prediction model measures the average of the squares of the prediction errors. The prediction error is the difference between the true value and the predicted value. Let $y_1, y_2, ..., y_n$ be the true oil prices and $\hat{y}_1, \hat{y}_2, ..., \hat{y}_n$ be the predicted oil prices under an oil price prediction model, then the MSPE of that model is:

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

For comparison purposes, we use the no-change model as the baseline model and express the MSPE of another model as a ratio relative to the MSPE of the no-change model. If the (MSPE) ratio of a model is less than 1, then the model is more accurate than the no-change model in terms of MSPE.

The second metric is Directional Accuracy Ratio (DAR), which measures the accuracy of predicting the direction of oil price change (i.e., whether oil price increases or decreases in the next time slot). It can be computed as follows:

$$DAR = \frac{1}{n}\sum_{i=1}^{n} d_t$$

where $d_t = 1$ if $(\hat{y}_t - y_{t-1})(y_t - y_{t-1}) > 0$ and $d_t = 0$ otherwise. Note that if we do a random guess of the oil price direction by tossing a fair coin, the DAR would be 0.5. Thus, if the DAR of a model is greater than 0.5, then the model is better than a random guess.

3.3. Evaluation

Tables 1 and 2 summarize the results for predicting the U.S. refiner acquisition cost for crude oil imports. Our stream learning model achieves the lowest MSPE and the highest DAR among all prediction models for the 1-month, 6-month, 9-month and 12-month forecast time horizons. For example, for the 1-month time horizon, the MSPE of our stream learning model is 11.87, with an error reduction of 31.1% compared with the no-change model (17.23), an error reduction of 13.7% compared with the ANN model (13.75), and an error reduction of 25.3% compared with the forecast combination model in Baumeister and Kilian (2015) (15.89). The DAR of our streaming learning model is 0.594, which is higher than both ANN model (0.582) and forecast combination model (0.570). For the 3-month horizon, although our model has a lower accuracy than the forecast combination model, it is more accurate than the no-change model, and more accurate than the ANN model in terms of MSPE.

Tables 3 and 4 summarize the results for predicting the WTI crude oil spot price. Again, our stream learning model achieves the highest accuracy among all prediction models for all forecast time horizons except for the 3-month time horizon. For example, for the 1-month time horizon, the MSPE of our streaming learning model is 17.16, with an error reduction of 16.2% compared with the no-change model (20.48), an error reduction of 8.7% compared with the ANN model (18.79), and an error reduction of 8.0% compared with the forecast combination model (18.66). The DAR of our streaming learning model is 0.570, which is also more accurate than both the ANN model (0.542) and the forecast combination model (0.521).

4. Conclusion

Forecasting crude oil prices is a very challenging problem due to the high volatility of oil prices. In this paper, we developed a new oil price prediction approach using ideas and tools from stream learning, a machine learning paradigm for analysis and inference of continuous flow of non-stationary data. Our stream learning model will be updated whenever new oil price data are available, so the model continuously evolves over time, and can capture the changing pattern of oil prices. In addition, updating the model requires only a small constant time per new data example, as opposed to re-training the model using the entire training data set. The experiment results show that our streaming learning model outperformed three other popular oil price prediction models over a variety of forecast time horizons.

Acknowledgements

This research was supported by grants from the National Natural Science Foundation of China (Grant Number: 71173200), and the Strategic Research Center for Oil and Gas Resources, Ministry of Land and Resources of the People's Republic of China (Grant Number: 2014BJYQ03).

References

Alavi, A.H., Gandomi, A.H., Lary, D.J., 2016. Progress of machine learning in geosciences: Preface. Geoscience Frontiers 7 (1), 1–2.

Alquist, R., Kilian, L., Vigfusson, R.J., 2013. Forecasting the price of oil. In: Elliott G, Timmermann A, editors. Handbook of economic forecasting. Amsterdam, North Holland, 427–507.

Baumeister, C., Kilian, L., 2015. Forecasting the real price of oil in a changing world - a forecast combination approach. Journal of Business and Economic Statistics 33 (3), 338–351.

Behmiri, N.B., Pires Manso, J.R., 2013. Crude oil price forecasting techniques: a comprehensive review of literature. Alternative Investment Analyst Review 3, 18–29.

Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B., 2010. MOA: massive online analysis. The Journal of Machine Learning Research 11, 1601–1604.

Bifet, A., Kirkby, R., Kranen, P., Reutemann, P., 2012. Massive online analysis manual. Centre for Open Software Innovation. Available: http://ufpr.dl.sourceforge.net/project/moa-datastream/documentation/Manual.pdf.

EIA, 2016. U.S. Energy Information Administration [Internet]. [cited 2016 Mar 12]. Available: http://www.eia.gov/petroleum/data.cfm#prices.

Frey, G., Manera, M., Markandya, A., Scarpa, E., 2009. Econometric models for oil price forecasting: a critical survey. CESifo Forum, 1: 29–44.

Gama, J., Sebastiao, R., Rodrigues, P.P., 2009. Issues in evaluation of stream learning algorithms. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Jun 28-Jul 1, Paris, France, 329–338.

Guerriero, V., et al., 2012. A permeability model for naturally fractured carbonate reservoirs. Marine and Petroleum Geology 40 (2), 115–134.

Husain, A.M., et al., 2015. Global implications of lower oil prices. IMF Survey Magazine, 1–41.

Kulkarni, S., Haidar, I., 2009. Forecasting model for crude oil price using artificial neural networks and commodity future prices. International Journal of Computer Science and Information Security 2 (1).

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

Priestley, M.B., 1988. Non-linear and non-stationary time series analysis. Academic Press.

Regnier, E., 2007. Oil and energy price volatility. Energy Economics 29 (3), 405–427.

Witten, I.H., Frank, E., Hall, M.A., 2011. Data mining: practical machine learning tools and techniques, 3rd ed. Morgan Kaufmann.

Xie, W., Yu, L., Xu, S., Wang, S., 2006. A new method for crude oil price forecasting based on support vector machines. Proceedings of the 6th International Conference on Computational Science (ICCS), May 28-31, Reading, UK, 444–451.

Yu, L., Wang, S.Y., Lai, K.K., 2008. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Economics 30 (5), 2623–2635.

Figure captions

Fig. 1. Monthly WTI crude oil spot price in U.S. dollar per barrel (original oil price time series and time series after difference transformation) from January 1986 to January 2016.

Table captions

Table 1 MSPE of different oil price prediction models: U.S. Refiner Acquisition Cost for Crude Oil Imports

Table 2 DAR of different oil price prediction models: U.S. Refiner Acquisition Cost for Crude Oil Imports

Table 3 MSPE of different oil price prediction models: West Texas Intermediate (WTI) Crude Oil Price

Table 4 DAR of different oil price prediction models: West Texas Intermediate (WTI) Crude Oil Price

Table 1. MSPE of different oil price prediction models: U.S. Refiner Acquisition Cost for Crude Oil Imports

Time Horizon		No-Change	ANN	Forecast Combination	Stream Learning
1-month	MSPE	17.23	13.75	15.89	**11.87**
	ratio	1	0.798	0.922	**0.689**
3-month	MSPE	97.13	96.92	**88.00**	95.37
	ratio	1	0.998	**0.906**	0.981
6-month	MSPE	221.75	205.21	212.21	**156.07**
	ratio	1	0.925	0.957	**0.704**
9-month	MSPE	277.09	276.06	262.68	**183.79**
	ratio	1	0.996	0.948	**0.663**
12-month	MSPE	307.72	303.23	280.64	**176.02**
	ratio	1	0.985	0.912	**0.572**

NOTES: MSPE stands for Mean Squared Prediction Error, lower is more accurate. Boldface indicates the best performance score among all four prediction models for a given forecast time horizon.

Table 2. DAR of different oil price prediction models: U.S. Refiner Acquisition Cost for Crude Oil Imports

Time Horizon		Random Guess	ANN	Forecast Combination	Stream Learning
1-month	DAR	0.5	0.582	0.570	**0.594**
3-month	DAR	0.5	0.586	**0.592**	0.562
6-month	DAR	0.5	0.510	0.556	**0.598**
9-month	DAR	0.5	0.594	0.575	**0.602**
12-month	DAR	0.5	0.643	0.627	**0.667**

NOTES: DAR stands for Directional Accuracy Ratio, higher is more accurate. Boldface indicates the best performance score among all four prediction models for a given forecast time horizon.

Table 3. MSPE of different oil price prediction models: West Texas Intermediate (WTI) Crude Oil Price

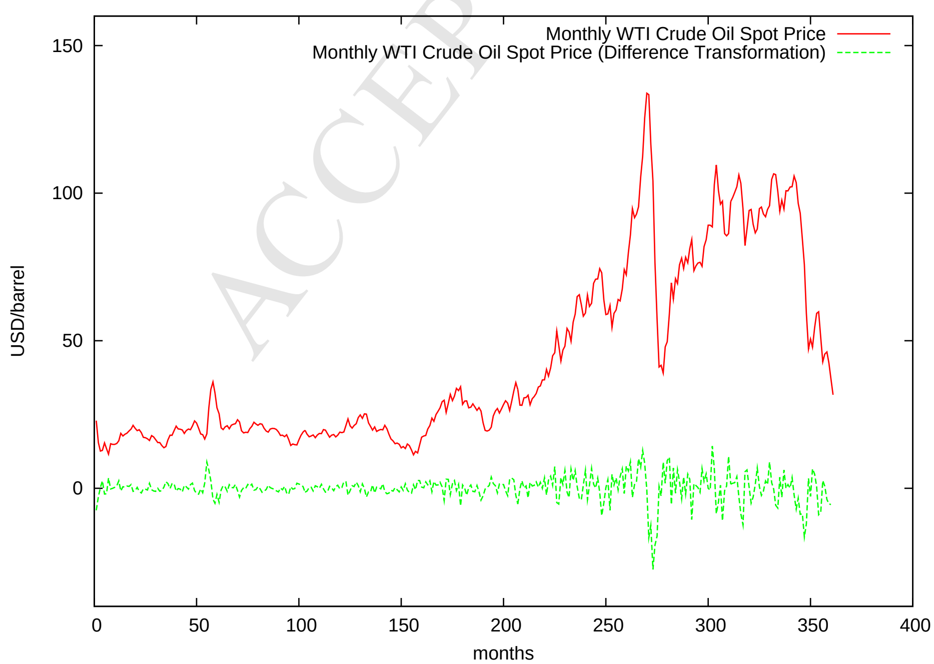
Time Horizon		No-Change	ANN	Forecast Combination	Stream Learning
1-month	MSPE	20.48	18.79	18.66	**17.16**
	ratio	1	0.917	0.911	**0.838**
3-month	MSPE	101.81	101.77	**92.24**	99.58
	ratio	1	0.999	**0.906**	0.978
6-month	MSPE	233.03	226.61	224.17	**181.47**
	ratio	1	0.972	0.962	**0.778**
9-month	MSPE	293.54	290.25	279.45	**214.34**
	ratio	1	0.989	0.952	**0.730**
12-month	MSPE	328.31	307.62	300.73	**234.98**
	ratio	1	0.937	0.916	**0.716**

NOTES: MSPE stands for Mean Squared Prediction Error, lower is more accurate. Boldface indicates the best performance score among all four prediction models for a given forecast time horizon.

Table 4. DAR of different oil price prediction models: West Texas Intermediate (WTI) Crude Oil Price

Time Horizon		Random Guess	ANN	Forecast Combination	Stream Learning
1-month	DAR	0.5	0.542	0.521	**0.570**
3-month	DAR	0.5	**0.586**	0.576	0.546
6-month	DAR	0.5	0.469	0.543	**0.602**
9-month	DAR	0.5	0.502	0.571	**0.606**
12-month	DAR	0.5	0.526	0.605	**0.610**

NOTES: DAR stands for Directional Accuracy Ratio, higher is more accurate. Boldface indicates the best performance score among all four prediction models for a given forecast time horizon.


Monthly WTI Crude Oil Spot Price
Monthly WTI Crude Oil Spot Price (Difference Transformation)
USD/barrel
months
150
100
50
0
0
50
100
150
200
250
300
350
400

Highlights

- Proposing a new approach for oil price prediction based on stream learning
- Updating the model whenever new oil price data are available to capture the changing pattern of oil prices
- Achieving the highest accuracy over a variety of forecast time horizons compared with three popular oil price prediction models