Modelling and Forecasting Monthly Crude Oil Price of Pakistan: A Comparative Study of ARIMA, GARCH and ARIMA Kalman Model

Muhammad Aamir [1, a)] and Ani Shabri [1, b)]

[1] *Department of Mathematical Sciences, Faculty of Science*
Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia.

[a)] Corresponding author: aamirkhan@awkum.edu.pk
[b)] ani@utm.my

Abstract. Crude oil is one of the most important commodity in the world and it is meaningful for every individual. This study aims at developing a more appropriate model for forecasting the monthly crude oil price of Pakistan. In this study, three-time series models are used namely Box-Jenkins ARIMA (Auto-regressive Integrated Moving Average), GARCH (Generalized Auto-regressive Conditional Hetero-scedasticity) and ARIMA Kalman for modelling and forecasting the monthly crude oil price of Pakistan. The capabilities of ARIMA, GARCH and ARIMA-Kalman in modelling and forecasting the monthly crude oil price are evaluated by MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error). It is concluded that the hybrid model of ARIMA Kalman perform well as compared to the Box-Jenkins ARIMA and GARCH models.

Keywords: ARIMA, Crude oil, Forecasting, GARCH, Kalman Filter.
PACS: 05.45.Tp

INTRODUCTION

The most important commodity which affects the daily life of every one in a number of ways is the crude oil. In today's world, crude oil is as much important as the blood for human body. Due to high demand of crude oil in every field of life needs more attention as compared to other commodities. Oil is a non-renewable commodity but the world consumes it in different ways thus it's a challenge for mathematician, statistician and econometrician to develop a better strategy for understanding the price changing aspect of crude oil. Having better strategies, agencies and suppliers consuming and supplying oil can take more accurate and up-to date decisions. Crude oil price forecasting is also very necessary for government agencies and investors to plan their activities in an effective manner. Due to the instable nature of global socio-economic and geo political events the crude oil prices are very sensitive. The fluctuations in the prices of crude oil is not only due to the demand, inventory, supply and consumption factors but also depends on so many other unpredictable and irregular factors that are random in nature. Designing appropriate models for forecasting the crude oil price is a challenging and complex task for mathematician due to stochastic and irregular form of the price of crude oil. In contrast, the compound and complex nature of the crude oil price, it is a widely open area for research and now a day's most of the researchers use a variety of different procedures for better forecasting of crude oil price. However, mostly two different techniques are used for crude oil price forecasting. In the first approach, the framework which is used for forecasting is akin to cause and effect, whereas the dependent variable is supposed to be affected by more variables generally called covariates. Sometimes this approach is also called fundamental analysis. Logically, this approach is very attractive by placing the reasons for ups and downs in price forecasting. So, many studies have found such as [1] and [2] used this technique. They expend the model of crude oil price and examine the nonlinear effect of processing plant utilization, OPEC capability utilization and future environment in markets as independent variables. This method has many limitations e.g. one cannot be sure about a certain explanatory variable that accounts for variations in the crude oil price. It is a difficult task to determine the exact functional form of a variable even if the exact variable is identified. The second approach is the time series modelling. In this approach, we no longer depend on the nature of explanatory variables, rather

Advances in Industrial and Applied Mathematics
AIP Conf. Proc. 1750, 060015-1–060015-7; doi: 10.1063/1.4954620
Published by AIP Publishing. 978-0-7354-1407-5/$30.00

the predictions for the future values based on the past behaviour of the study variable. Several studies have been conducted on this approach for example.[3-6]. For forecasting the crude oil price, they used the well-known Box Jenkins methods. For forecasting the crude oil price, several researchers such as [7], [8] and [9] used the GARCH type of models. Moreover, it is to be observed that the time series data acted the way it did because we are not capable to report all the changes of ups and downs based on natural reasoning, economic theory or inventory levels in the crude oil price. For better forecasting, many different approaches are used. In the midst of competing models for obtaining the forecast, selecting an appropriate model is a problem. In such situations, the choice of a model is usually based on the past accuracy, but the problem arises when the differences are statistically significant. Diebold and Mariano [10] discussed the value of the econometric agents demanding to be more critical, when deciding on what forecasts to base their decisions.

The state space approach also known as dynamic linear modelling used for the univariate methods. These techniques are not new for forecasting, using them contribute to the model development of common mathematical problems [11]. The main advantage of putting the model in state space form is that, a lot of statistical powerful tools can be applied to the mathematical model. In other words, it is a simple way for putting all procedures in a common form. To obtain optimum forecasts the Kalman filter technique can be used, discussed in the next section and the best estimates of the unobserved components are attained by the Kalman smoothing [12]. The state variables are the unobserved variables and the model parameters are estimated as well as the unobserved expectations are calculated instantaneously [13].

First time Kalman filter was introduced by Kalman [14]. Due to the Kalman filter, the state space models modern work started. First time it was verified that the Kalman Filter approach is much useful in engineering and in space technology. First time [15] stated that there is no general technique available which can be used for applying the Kalman Filter in statistical forecasting. However, the work started at that time for obtaining a procedure for Kalman Filter in the field of statistics. Likewise the work of [16], [17] and [18] decided that the Kalman Filter have a great importance for statistical applications in econometrics, for example forecasting of time series data and other economic properties. The forecasts of the state space models are obtained by the Kalman filtering technique. The work of [19] highlighted the importance of this technique. Also [20] discussed that the Kalman Filter produces better forecast as compared to the SARIMA models forecast for the inflation data of four different countries. The main reason for the restricted use of the Kalman Filter approach on ARIMA models possibly be the difficulties as to how to find/specify the prior matrices, and there is no guarantee of the benefits by adopting this hectic procedure as well. Hence, it is viable to apply Kalman Filter technique in this study in order to achieve better forecasts.

Thus, the purpose of this work is to compare the performance of the three techniques, how they perform well for out of sample forecasting by using the estimated values from the training set. The in sample forecast accuracy will not be considered in this work, the reason being that the data are divided in to two parts and for the testing set the ARIMA model estimated coefficients are used as initial values for the Kalman Filtering. The other benefit of the Kalman Filter is that, it works with the missing data [21] but this case is not considered in this work. The Kalman smoother is also closely connected with the framework of Kalman Filter and it is used for the state vector inferences for example, for any of its past values. It is also observed that the state vector is valuable but not feasible for this work, rather be included in some future studies. The work of [22] consider the basic paper for this work. The remainder of this work is summarized, in section 2 is comprised of the methodology regarding the considered models, section 3 represents the analysis and numerical work while the last section 4 consists of the discussions, conclusion and future work recommendations.

METHODOLOGY

Methodology related to ARIMA and ARIMA GARCH model is presented first, secondly the state space models, thirdly Kalman filter and in the last the forecast accuracies measures are discussed.

ARIMA Models

The ARIMA models need stationary time series. Stationary means the autocorrelation structure, the mean and variance of the time series over the time remains constant because the data is smooth and it does not consist of trend and seasonality factors. If sometimes the time series is not stationary, convert the time series into a stationary series by taking the successive differences of the series to achieve the required goal. ARIMA models are commonly used for the measurement of the serial dependence in a time series, where the interdependency in time series (y) is measured by the AR terms, and how much the series (y) depends on the previous error terms is described by the MA terms [23]. An Autoregressive Moving Average ARMA (ARIMA) (m, r) model of a univariate time series has the following structure:

$$y_t = \gamma_0 + \gamma_1 y_{t-1} + \cdots + \gamma_m y_{t-m} + \epsilon_t + \beta_1 \epsilon_{t-1} + \cdots + \beta_r \epsilon_{t-r} \tag{1}$$

The sample autocorrelation function (ACF) and (PACF) partial autocorrelation functions can be used for selecting the appropriate order of the polynomials, and compared with the theoretical bounds of [24]. Rival ARIMA models can also be selected on the basis of AIC, BIC or AICc information criterion.

ARIMA GARCH Hybrid Model

The hybrid modelling procedure of ARIMA GARCH consists of two phases. The first phase fits the ARIMA best model for the stationary and linear time series while the residuals contain the non-linear part of the data. After completing the first phase, the second phase covers the non-linearity part of the residuals by fitting the GARCH model. So, this hybrid model of ARIMA GARCH is used to evaluate the forecast of the values of time series see [25], [26] and [27]. The first phase of this procedure uses the model (1) and the second phase uses the following GARCH model (2). . Hence, GARCH (*m, r*) model, where m is the order of GARCH terms while *r* is the order of ARCH terms. Also noted that when *m=0* the GARCH model equals ARCH model. The form of the σ_t are as follows: $\varepsilon_t = \sigma_t y_t, y_t$ white noise,

$$\sigma_t^2 = \omega + \sum_{i=1}^{m} \vartheta_i \varepsilon_{t-i}^2 + \sum_{i=1}^{r} \mu_i \sigma_{t-i}^2 \tag{2}$$

The State Space Model

The procedure of state space model consists of two steps. First, a state vector is generated for the significant components which are accumulated until the end. The smallest vector which is called the state vector summarizes the past behaviour of the full system playing the key role of state space modelling which determines the state vector [28]. The state space modelling approach explains the smoothing in the form of two linear equations. The first equation shows the relationships between the current observation and unobserved states which is also called the observation equation while the second equation shows the progress in the states over time called state equation. The state vector is continuously updated through the state equation. The state equation is a vector comprising of undetected components such as trend, seasonality and level or Autoregressive and Moving Averages factors of a time series. The general form of the state space model are as follows:

$$y_t = H'\alpha_t + \varepsilon_t$$
$$\alpha_t = F'\alpha_{t-1} + Tw_t$$

In the above equations, y_t represent the observed vector of variables, α_t display the vector for the unobserved variables. The matrices F and H are parametric matrices while ε_t and w_t are white noise terms with their covariance matrices R and Q respectively [29].

ARIMA State Space Representation

Initially [29] presented the ARMA models in state space form but later on it was modified by [21] concerning with the MA terms. The latest form of the ARMA model in state space are as follows:

$$y_t = H\alpha_t + \varepsilon_t$$
$$\alpha_t = F\alpha_{t-1} + Tw_t$$

The observation equation of the state space represents ARMA (*m, r*) are as under:

$$y_t = [1 \ 0 \ 0 \ \ldots 0] \begin{bmatrix} \alpha_t \\ \alpha_{t-1} \\ \vdots \\ \alpha_{t-i-1} \end{bmatrix} + \boldsymbol{\varepsilon_t}$$

The state equation of ARMA (*m, r*) is

$$\begin{bmatrix} \alpha_{t+1} \\ \alpha_t \\ \vdots \\ \alpha_{t-i} \end{bmatrix} = \begin{bmatrix} \gamma_0 & 1 & 0 & \ldots & 0 \\ \gamma_1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \gamma_{i-1} & 0 & 0 & \ldots & 1 \\ \gamma_i & 0 & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} \alpha_t \\ \alpha_{t-1} \\ \vdots \\ \alpha_{t-i-1} \end{bmatrix} + \begin{bmatrix} 1 \\ \beta_1 \\ \vdots \\ \beta_{i-1} \end{bmatrix} w_t$$

Where $\varepsilon_t \sim N(0, Z_t)$, $w_t \sim N(0, Q)$ and $Q = T\sigma^2 T'$, $H = [1\ 0 \ldots 0]$, $Z_t = 0$

$$F = \begin{bmatrix} \gamma_0 & 1 & 0 & \ldots & 0 \\ \gamma_1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \gamma_{i-1} & 0 & 0 & \ldots & 1 \\ \gamma_i & 0 & 0 & \ldots & 0 \end{bmatrix}, T = \begin{bmatrix} 1 \\ \beta_1 \\ \vdots \\ \beta_{i-1} \end{bmatrix} \text{and } Q = \sigma^2 \begin{bmatrix} 1 & \beta_1 & \ldots & \beta_{i-1} \\ \beta_1 & \beta_1^2 & \ldots & \beta_1 \beta_{i-1} \\ \vdots & \vdots & \ldots & \vdots \\ \beta_{i-1} & \beta_1 \beta_{i-1} & \ldots & \beta_{i-1}^2 \end{bmatrix}$$

This leads to an ARMA (m, r) model, for which $i = \max(m, r+1)$, where $\gamma_i = 0,\ for\ i > m$ and$\beta_i = 0, for\ i > r+1$. Thus, from the above matrices it is clear that when using Kalman Filter, it first needs to estimate *F*, *T*, *Z* and *Q* matrices.

Kalman Filter

Kalman filter is an optimal linear estimator. It infers the model parameters of interest from indirect, uncertain and inaccurate observations. The interesting feature of Kalman filter is, it updates the system knowledge after receiving the new observation, and due to this factor the error terms are minimized as well, which yields the minimum mean square error of the state vector of the system for the given information available at time *t*. By filtering out the noise term the best estimates are found, and the measurements onto the state estimate is projected, which leads to update the state vector. After updating the state vector, the moments of the state vector distribution revises progressively and unobserved values of the time series. The filter is an algorithm used for solving the linear state space models. The prediction and updating the system equations called Kalman Filter. The new part of the time series y_t which is not explained from the past values are called the innovation terms and equal to the Kalman filter residuals. Kalman filter maintains the estimates of the state vector and error covariance matrix of the state. The Kalman Filter recursively calculates the states predictions of y_t conditional on the historical information and also on the variance of their prediction error. The term $\boldsymbol{\varepsilon_t}$ is the innovation at time *t*, that is the new information in y_t which could not be predicted from the past knowledge, is referred to as the one step ahead prediction error [30]. The recursion of the simple Kalman Filter is comprised of three steps are as follows:

Step 1. Start with the guess values for the state vector $\alpha_{0/0}$ and for the covariance $P_{0/0}$ at time $t = 0$.
Step 2. Prediction. At time $t = 1$ form an optimal prediction value for $y_{1/0}$ using the estimated value for$\alpha_{1/0}$.
Step 3. Updating. Use the observed value of *y* at time $t = 1$ to find the prediction error.

i.e. $\omega_{1/0} = y_1 - y_{1/0}$ this prediction error contains the information to use that information we can refine our guess about state estimator α. $\alpha_{1/1} = \alpha_{1/0} + K_t\omega_{1/0}$ Where K_t is called the Kalman gain (weight assign to the new information) and $\alpha_{1/1}$ is conditional on time 1. The overall view of the Kalman Filter are as follows:

Starting values at time $t = 0$	$\alpha_{0/0}, P_{0/0}$
Predict state vector at time $t = 1$	$\alpha_{t/t-1} = F\alpha_{t-1/t-1}$,
	$P_{t/t-1} = FP_{t-1/t-1}F' + Q$
Calculation of prediction error	$\omega_{t/t-1} = y_t - y_{t/t-1} = y_t - H_t\alpha_{t/t-1}$
	$f_{t/t-1} = H_tP_{t/t-1}H' + Z$
Update States	$K_t = P_{t/t-1}H'f_{t/t-1}^{-1}$
	$\alpha_{t/t} = \alpha_{t/t-1} + K_t\omega_{t/t-1}$
	$P_{t/t} = P_{t/t-1} - K_tH_tP_{t/t-1}$

MODELLING AND ANALYSIS OF PAKISTAN CRUDE OIL PRICES

Data Information

In this piece of work, the average monthly prices of Pakistan crude oil are used as an experimental sample (all prices are per barrel in Pakistani rupees). This study covers the period from February, 1986 to March, 2015 and a total of 350 observations. The total data is divided into two parts one is training set consisted of the first 300 observations and the rest 50 observations are used as a testing set. FIGURE 1 shows the plot of the whole data set and a vertical line on a plot separate the training and testing data sets.

Analysis and Modelling of Data

First of all, the whole data set is plotted against the time, and modelling was done. For checking stationary of the data the unit root test ADF is used suggested by [31], which suggests that the data is stationary after the first difference. Next proceeded for the selection of proper AR and MA orders ACF, and PACF plots are used [32] from which an ARMA (3, 1) model is selected and the other criterion AIC suggested by [33], preferred the same model as the best candidate model for the said data set. Thus our final selected model for the training data set is

ARIMA (3, 1, 1) without trend factor. The coefficient of the selected candidate model is presented as in Table 1. From the Table 1, it is clear that all the AR and MA coefficient are highly significant except AR (2). So, the selected model is applied to the training data set and the fitted values are observed. Next, for the comparison purpose a hybrid model of ARIMA GARCH were also fitted for the training data set. The selected model in this case is ARIMA (3,1,1) + GARCH (1,1) according to the AIC criteria. The GARCH (1,1) estimated coefficient values are also presented in Table 1.

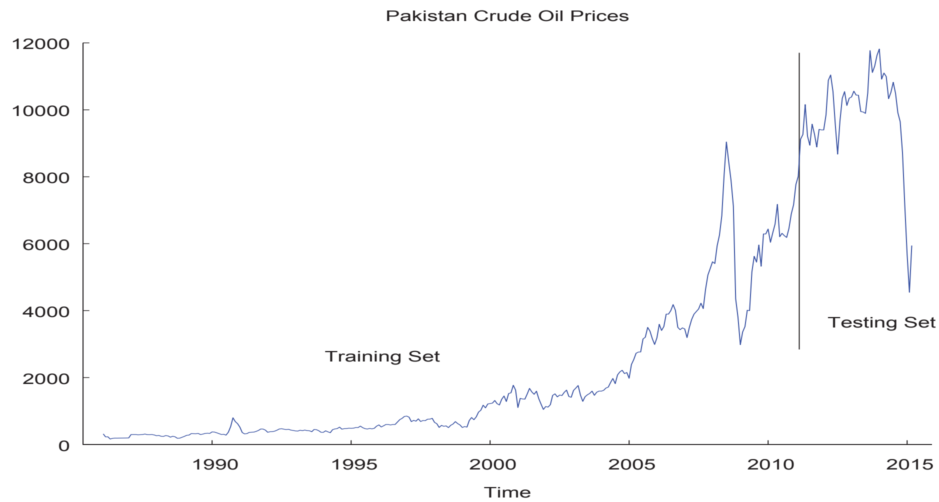

FIGURE 1. Plot of the Complete Data set also Showing the Testing and Training sets

TABLE 1. Parameters estimation results of ARIMA and GARCH models

Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
AR(1)	1.0132	0.0727	13.93	0.0000***
AR(2)	0.0886	0.0815	1.090	0.2771
AR(3)	-0.3279	0.0555	-5.909	0.0000***
MA(1)	-0.8338	0.0565	-14.76	0.0000***
Omega(ω)	62.135	13.293	4.674	0.0000***
Phi(ϑ)	0.4124	0.0649	6.355	0.0000***
Mu(μ)	0.7006	0.0303	23.11	0.0000***

Significance Codes: 0 '***', 0.001'**', 0.01 '*', $\hat{\sigma}^2 = 63300$

Analysis of Kalman Filtering and Forecasting

To obtain the values for the Kalman filter, Kalman smoothed series and Kalman forecasting of the training data set we use the dlm package of R proposed by [34]. For Kalman filter recursion, initially we need to assume the values of the fitted ARIMA model shown in Table 1. The order of m are three while r equal to one, therefore, the order of the matrices are $i = \max(3,2)$. Thus the matrices F and Q are of order (3×3).The assumed F and Q starting matrices for Kalman Filter recursion in dlm package are as follows:

$$F = \begin{bmatrix} 1.0132 & 1 & 0 \\ 0.0886 & 0 & 1 \\ -0.3279 & 0 & 0 \end{bmatrix} \text{ and } Q = \begin{bmatrix} 1.0000 & -0.8338 & 0 \\ -0.8338 & 0.6952 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Forecast Accuracy Comparison

In this section, we compare the performance of the two techniques in forecasting that which technique predict more accurately the future prices of the Pakistan crude oil prices. The study of [22] suggested that the hybrid model of ARIMA and GARCH is the best candidate model for the said data. Here, we observed that the forecasting power of ARIMA with Kalman filter is much higher than the ARIMA and ARIMA+GARCH models. The forecast accuracy is compared only for the testing set. Table 2 represents the values of MAE and RMSE values for the different techniques.

TABLE 2. The MAE and RMSE comparisons for different methods

Model	MAE	RMSE
ARIMA	542.28	679.32
ARIMA+GARCH	525.78	671.25
ARIMA+Kalman Filter	**421.91**	**522.54**

From Table 2, it is clear that the performance of ARIMA and Kalman filter is much better than the other two techniques. So, Kalman technique is better approach to use for forecasting of Pakistan monthly crude oil price. The smallest value in each case is highlighted showing the best model in terms of forecasting power accuracy. FIGURE 2 shows the graph of the observed values, ARIMA and Kalman Filter one step ahead forecasted values for the testing data set.

DISCUSSION AND CONCLUSION

In this study a different approach towards forecasting of Pakistan monthly crude oil price is used. Firstly, the whole data is divided into two parts one is training data set and the other is testing data set. For training data set, the best ARIMA model is fitted and their respective coefficients were calculated. The ARIMA fitted model coefficients for the training data set are used as initial values for Kalman Filtering and forecast the values for the testing data set. The Kalman forecast is a one step ahead forecast. For comparison purpose, the ARIMA and hybrid ARIMA GARCH models one step ahead forecasted values were calculated for testing data set. The two measures for the comparison of forecast accuracy were used, which decided that the ARIMA Kalman Filter technique is the best approach to be used for forecasting the Pakistan monthly crude oil price. Thus this study suggests that in future, the ARIMA Kalman Filter approach be applied for forecasting the monthly crude oil price of Pakistan in order to obtain the best results. This study focuses only on Kalman Filter while in future the Kalman smoother approach will be used for inferencing the states values.

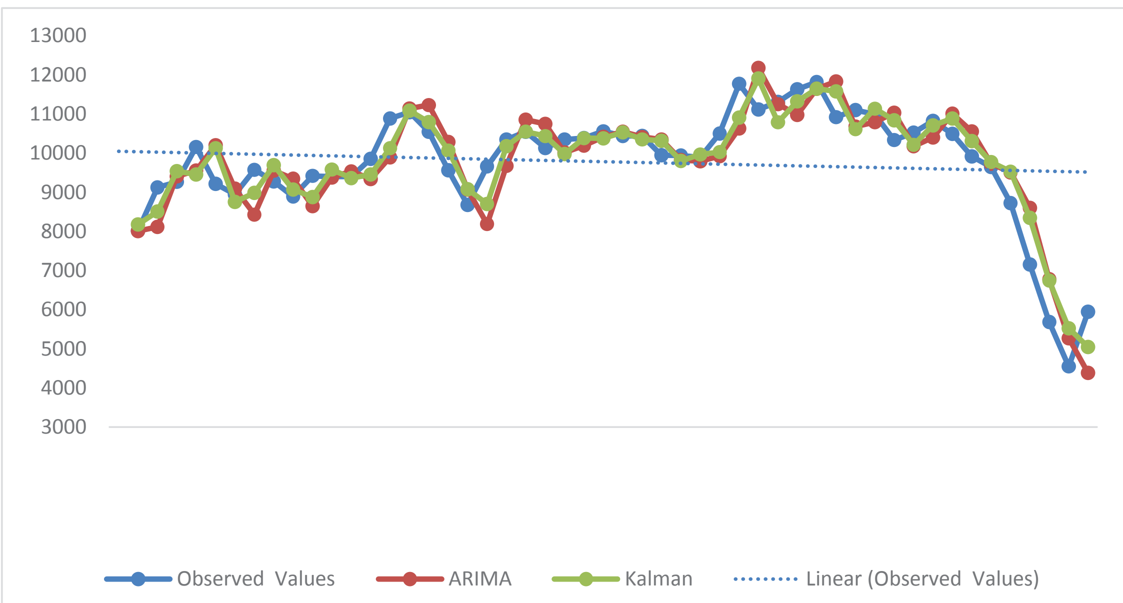

FIGURE 2. Testing set Observed and One Step Forecasted Values

REFERENCES

1. M. Ye, J. Zyren and J. Shore, International Journal of Forecasting **21** (3), 491-501 (2005).
2. S. Dées, A. Gasteuil, R. Kaufmann and M. Mann, European Central Bank Working Paper, 4-36 (2008).
3. L. M. Liu, Journal of Forecasting **10** (5), 521-547 (1991).
4. P. Agnolucci, Energy Economics **31** (2), 316-321 (2009).
5. M. Ahmad, International Journal of Trade and Global Markets **5** (1), 24-30 (2012).
6. M. Tariq, Z. U. Haq, S. Jan, M. Jehangir and M. Aamir, Life Science Journal **10** (11) (2013).
7. P. Sadorsky, Energy Economics **28** (4), 467-488 (2006).

8. A. Hou and S. Suardi, Energy Economics **34** (2), 618-626 (2012).
9. R. A. Ahmed and A. B. Shabri, Life Science Journal **10** (4) (2013).
10. F. X. Diebold and R. S. Mariano, Journal of Business & economic statistics (2012).
11. R. Hyndman, A. B. Koehler, J. K. Ord and R. D. Snyder, *Forecasting with exponential smoothing: the state space approach*. (Springer Science & Business Media, 2008).
12. A. C. Harvey, Journal of Forecasting **3** (3), 245-275 (1984).
13. E. Burmeister and K. D. Wall, Journal of Econometrics **20** (2), 255-284 (1982).
14. R. E. Kalman, Journal of Fluids Engineering **82** (1), 35-45 (1960).
15. G. W. Morrison and D. H. Pike, Management Science **23** (7), 768-774 (1977).
16. B. Rosenberg, in *Annals of Economic and Social Measurement, Volume 2, number 4* (NBER, 1973), pp. 399-428.
17. R. F. Engle, in *Seasonal analysis of economic time series* (NBER, 1979), pp. 281-308.
18. A. C. Harvey and G. D. Phillips, Biometrika **66** (1), 49-58 (1979).
19. S. Ravichandran and J. Prajneshu, Journal of Indian Society of Agricultural Statistics **54** (1), 43-51 (2001).
20. F. Nikolaisen Sävås, Master Thesis, Uppsala Universitet (2013).
21. G. Petris, S. Petrone and P. Campagnoli, *Dynamic linear models with R*. (Springer Science & Business Media, 2009).
22. M. Aamir and A. B. Shabri, Science International **27** (3), 2365-2371 (2015).
23. D. C. Montgomery, C. L. Jennings and M. Kulahci, *Introduction to time series analysis and forecasting*. (John Wiley & Sons, 2015).
24. G. E. Box, G. M. Jenkins and G. C. Reinsel, *Time series analysis: forecasting and control*. (John Wiley & Sons, 2011).
25. C. Chen, J. Hu, Q. Meng and Y. Zhang, presented at the Intelligent Vehicles Symposium (IV), 2011 (unpublished).
26. H. Liu and J. Shi, Energy Economics **37**, 152-166 (2013).
27. Z. Tan, J. Zhang, J. Wang and J. Xu, Applied Energy **87** (11), 3606-3610 (2010).
28. J. Brocklebank and D. A. Dickey, *SAS for forecasting time series*. (SAS institute, 2003).
29. J. D. Hamilton, *Time series analysis*. (Princeton university press Princeton, 1994).
30. N. Saini and A. K. Mittal, Journal of Statistical and Econometric Methods **3** (1), 115-136 (2014).
31. D. A. Dickey and W. A. Fuller, Journal of the American statistical association **74** (366a), 427-431 (1979).
32. G. E. Box and G. M. Jenkins, *Time series analysis: forecasting and control, revised ed*. (Holden-Day, 1976).
33. H. Akaike, Automatic Control, IEEE Transactions on **19** (6), 716-723 (1974).
34. G. Petris, Journal of Statistical Software **36** (12), 1-16 (2010).