IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

Forecasting the Oil Price with a Periodic Regression ARFIMA-GARCH Process

Daniel Ambach
Department for Data Science
smava GmbH
Berlin, Germany
daniel.ambach@smava.de

Oleksandra Ambach
FirmenCenter Gründung und Nachfolge
Berliner Sparkasse
Berlin, Germany
oleksandra.ambach@berliner-sparkasse.de

***Abstract*—This article provides a new periodic time series model to predict the oil price. Moreover, the approach discusses short-term forecasting of the oil price. Hence, we discuss the model fit and the out-of-sample performance. Finally, we derive further enhancements and improvements for further research.**

***Index Terms*—long-memory, forecasting, oil-price, ARFIMA, periodic model**

I. Introduction

Supply and demand are two major forces that drive the equilibrium on the market. Foreign exchange rates, interest rates, stock prices are the instruments that play an important role in this process. Though in the modern world there is one more important variable that determines the equilibrium on the market - the oil price. An oil price is a commodity traded on a global market. The crucial thing about it is that in contrast to the interest rates, for example, which have mostly an economic influence, the oil price affects us all being the principle source of the energy and pricing coal and natural gas.

Moreover, as we know the oil price is captured in the oil future contracts. An oil futures contract is described as a binding agreement which provides a the right to purchase a barrel oil at a predefined price on a predefined date in the future. Therefore, the buyer and the seller are obliged to make a deal according to the contract. It means that that the oil prices are manually agreed and accordingly predicted. Things become more difficult with the speculations on the oil market and cyclical trends. It turns out that regardless of how the price on the market is determined, based on its use in fuels and countless consumer goods, the oil price is inevitably in high demand for the future [18].

The aim of our paper is related to the successful application of a periodic regression model and residual process that follows a autoregressive fractional integrated moving average process with generalised autoregressive conditional heteroscedasticity (ARFIMA-GARCH) process to predict the oil price. In recent literature we find on the one-hand different ideas to model the oil price and on the other-hand we find different applications of the ARFIMA-GARCH model. Therefore, we use the properties of our proposed process to model and predict the oil price as good as possible. [3] analyse inflation by the fractionally integrated ARFIMA-GARCH model. They consider the application of long-memory processes to describing the inflation for ten countries. It is proved that for three high inflation economies there is evidence that the mean and the volatility of inflation interact in a way that is consistent with the Friedman hypothesis. [17] use the ARFIMA-GARCH Model and apply it to the realized volatility and the continuous sample path variations constructed from high-frequency Nikkei 225 data. [20] consider a periodic seasonal Reg-ARFIMA-GARCH model for daily electricity spot prices. This approach depicts periodic extensions of dynamic long-memory regression models with autoregressive conditional heteroscedastic errors. There model is accurate to analyse and predict the daily spot prices. [21] sufficiently explore the heart rate variability data with its stationary characteristics, long range correlations and instantiations volatility with the help of ARFIMA GARCH model. Another application is presented by [19]. They assess the persistence dependence of rainfall time series of Chui Chak, a station in Peninsular Malaysia that observed the highest rainfall event for the period 1975-01-01 to 2008-12-31.

The theory and the application of the periodic regression with ARFIMA-GARCH process is discussed in this paper. There is an evidence in the literature that there exist characteristics in the data which enables the use of the aforementioned model. For example we find cyclic behaviour throughout the trading year, high autocorrelation which is related to the seasonal behaviour and use of oil, conditional heteroscedasticity related to the volatility of the price process and heavy tailed residuals modelled by a t-distribution. Therefore, we will discuss the theoretical background and the modelling process. After that we will apply the model to the in-sample data, and if it fits well, we will try to forecast the out-of-sample results.

This article starts with a description of an ARFIMA-GARCH in Section II. Hereafter, the theory of long memory process, GARCH process, model fit and model diagnostics is presented. Section III discusses the application of the aforementioned model to the data set and presents in-sample results. Section IV provides the out-of sample results and discusses improvements and finally V concludes.

978-1-5386-2874-4/18/$31.00 2018 European Union

II. THEORETICAL BACKGROUND

The oil price (P_t) is from the economic sense an important indicator. Figure 1 shows all observations from 1986-01-02 to 2017-01-09 of the West Texas Intermediate (WTI) Cushing oil price. The oil price incorporates different properties. It is reasonable to assume that the oil price has periodic as well as autoregressive disturbances. Moreover, the variance process will show conditional heteroscedasticity. Therefore, we will apply the following model to the oil price

$$P_t = \mu + Periodic^{Reg} + \epsilon_t, \quad (1)$$

where μ is an intercept and ϵ_t is the residual process that follows an autoregressive fractional integrated moving average process with generalised autoregressive conditional heteroscedasticity (ARFIMA-GARCH). Furthermore, we model the periodic or seasonal structure by periodic sine and cosine functions which are

$$Periodic^{Reg} = \left(a_{\cos}\cos\left(\frac{2\pi t}{P}\right) + a_{\sin}\sin\left(\frac{2\pi t}{P}\right)\right), \quad (2)$$

where P is the period which is the basis of the periodic functions. Here we suggest that the period could be related to an annual cycle.

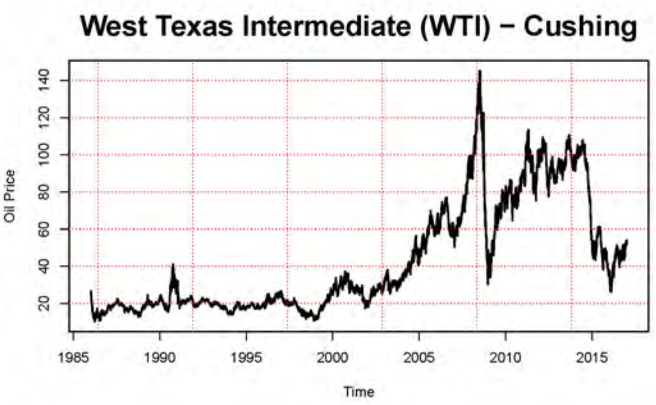

Fig. 1. Price of the West Texas Intermediate (WTI) Cushing, full sample (left)

A. the long-memory process

We introduce the ARFIMA-GARCH model to describe the autocorrelation and conditional heteroscedasticity. The residual process $\{\epsilon_t\}$ foremost as ARFIMA(p,d,q)-GARCH(P,Q) model in the following way

$$\epsilon_t \equiv \quad \phi(B)\nabla^d X_t = \theta(B)\eta_t \quad (3)$$

$$\eta_t = \sigma_t \tau_t, \quad (4)$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \eta_{t-i}^2 + \sum_{j=1}^{p} \gamma_j \sigma_{t-j}^2, \quad (5)$$

$$\tau_t \sim F \quad \text{here we take the t-distribution.} \quad (6)$$

Obviously equation (3) describes a typical ARFIMA(p,d,q) process and (4) depicts a GARCH(P,Q) process. Quite frequently long memory processes are considered while analyzing environmental time series [16]. In contrast an ARMA process denoted as U_t is considered as a short-memory process because the covariance of U_1 and U_{1+k} is decreasing fast as k converges to ∞. [7] explain that the autocorrelation function of U_t is geometrically bounded and therefore shows a short dependence structure. In contrast a long memory process has an autocorrelation function for which $\rho(k) \sim Ck^{2d-1}$ as $k \to \infty$, where C is a constant with $C \neq 0$ and $d < 0.5$ [7]. [16] use the ARFIMA model to predict the wind speed in Ireland.

Defining for any real number of d and $d > -1$, the difference operator $\nabla^d = (1-B)^d$ by the binomial expansion:

$$\nabla^d = (1-B)^d = \sum_{k=0}^{\infty} \frac{(k-d-1)!}{k!(-d-1)!} B^k,$$

where B is the backward shift operator $B^v X_t = X_{t-v}$.

In the continuous case [22] introduce the fractional Brownian motion. In the discrete setting the fractional integrated noise are the following difference equations $\nabla^d X_t = Z_t$, where $\{Z_t\} \sim WN(0, \sigma^2)$ and $d \in (-0.5, 0.5)$ (see [7]).

The ARFIMA(p,d,q) and GARCH(P,Q) process is a combination of an ARFIMA and a GARCH model. The contribution of a process which has a fractionally integrated conditional mean and generalized autoregressive conditional heteroscedasticity is to cover the effects that are not probably modelled by short time memory and a constant variance. An initial approach is given in [15] wherein ARMA-ARCH model explains macroeconomic time series. [2] used an ARFIMA-GARCH model to describe inflation rates. Moreover, they develop an approximate maximum likelihood estimate of an ARFIMA-GARCH process. The ARCH model [10] and GARCH model [4] are invented to model financial time series. The combined ARFIMA-GARCH model is given by

After discussing different time series models for the residual process we have to apply them to the data set and proceed with the parameter estimation and diagnostic checking. The next Section provides such results as well as the estimation of the complete univariate time series model.

B. Parameter estimation and model selection

The estimation of the complete model could be done in two different ways. On the one-hand it is possible to use a two step approach, where first the mean and hereafter the variance is estimated. Hence, on the other-hand the whole model is estimated in a single maximum likelihood approach. However, the exact likelihood is unknown, therefore we consider the conditional (quasi) maximum likelihood. For deriving the quasi maximum likelihood we can use the explanations and derivations of [7] and [26].

Finally, we assume that the residual process $\{\tau_t\}$ has to be independent and equally distributed with $\{\tau_t\} \sim t$. The t-distribution provides better results related to the tail behaviour of the residuals. The estimated periodic regression model with

ARFIMA-GARCH residual process is evaluated by means of the in-sample and the out-of-sample performance. [11] describes the R-square R^2 as a possible criterion for the in-sample performance of a regression model. Unfortunately, while using correlated data, we do not obtain an unbiased value of R^2. Hence, we can not interpret the R^2 so easily. One alternative is related to the goodness of the model. Therefore, we calculate the autocorrelation function ACF and the partial autocorrelation function.

The autocorrelation function can be used for two purposes. The first reason is to detect non-randomness in data. Moreover, we identify an appropriate time series model if the data are not random. If we want to identify the correct model order of an autoregressive model and figure out whether we observe further autocorrelation, we have to consider the partial autocorrelation function as well. These functions are also useful for the model diagnostic, if they provide evidence for further autocorrelation, we observe that the goodness of our model is not sufficient.

[1] proposes to measure the goodness of fit for a certain model by balancing the error of the fitted model against the number of parameters in the estimated model. The Akaike information criterion (AIC) is given by [1]

$$AIC(k) = \ln \widehat{\sigma}(k)^2 \, + \, k \, \frac{2}{n}, \tag{7}$$

where $\widehat{\sigma}(k)^2$ is the estimated variance, k is the number of all parameters in the complete model and n is the sample size. Another information criterion is the Bayesian information criterion (BIC). [24] propose the BIC which has a larger punishing term. The BIC is given by

$$BIC(k) = \ln \widehat{\sigma}(k)^2 \, + \, k \, \frac{\ln n}{n} \tag{8}$$

Both information criteria are appropriate to choose the best model order of our ARFIMA-GARCH model, but they could select a different model order in the end. The reason for such a result is related to the punishing terms $k\, 2/n$ and $k\, \ln n/n$, which are different. We calculate the information criterion for each model order of the ARFIMA-GARCH process. Hence, the AIC and the BIC can be derived by means of the likelihood function instead of $\widehat{\sigma}(k)^2$. The optimal model order minimizes the AIC and BIC.

The optimal model provides necessary information for the in-sample performance, but in addition, we are able to predict different unobserved values. [7] discusses the prediction of ARMA, ARFIMA and GARCH model.

III. Analysing the Oil Price and the Goodness of Fit

The regression model with ARFIMA-GARCH residual process with t-distribution given by (1) - (3) is applied to our oil price data set of the West Texas Intermediate (WTI). Figure 2 shows a part of the data set which is modelled with the aforementioned approach

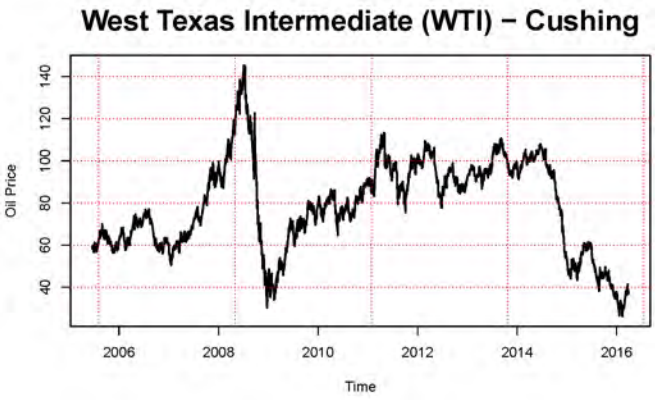

Fig. 2. Price of the West Texas Intermediate (WTI) Cushing

The data set is provided by “U.S. Energy Information Administration” and the investigated time horizon reaches from 2005-06-17 till 2016-03-29. The regressors of the model are given by periodic variables.

$$P_t = \mu + Periodic^{Reg} + \epsilon_t, \tag{9}$$

$$Periodic^{Reg} = \left(a_{\cos} \cos\left(\frac{2\pi t}{252}\right) + a_{\sin} \sin\left(\frac{2\pi t}{252}\right) \right) \tag{10}$$

where the period is 252, which is related to an average trading year of 252 days. The price process P_t has to be analysed according to the autocorrelation structure and to find heteroscedastic effects. Figure 3 provides a huge autocorrelation structure and the PACF selects a positive autocorrelation for the first lag and some more autocorrelation for the following lags.

Furthermore, Figure 5 depicts a high presence of conditional heteroscedasticity. From the aforementioned findings, we may assume that the proposed periodic regression model with ARFIMA-GARCH residuals are appropriate to capture the main characteristics of the WTI oil price. Subsequently, we fit the model to our data set. Table III shows the model estimation results. The derived model order is related to the smallest AIC and BIC and significance of the parameters. The majority of the parameters are significant, but some GARCH parameters are not. Moreover, the periodic regressor which is modelled with the sine function is only significant to a significance level of 0.1, which is acceptable in practice.

The obtained estimation results seem to be sufficient, but we have to discuss the goodness of fit. The data set has to be uncorrelated, homoscedastic and the remaining residuals should follow the t-distribution. The ACF of the residuals does not provide any correlation structure. Furthermore, we detect only for the first lag of the squared standardized residuals a remaining presence of autocorrelation. In addition we can observe that the Ljung and Box test points in the same direction. The test is applied to the residuals as well as to the squared standardised residuals. Moreover, a weighted ARCH-LM test for detecting further conditional heteroscedasticity is

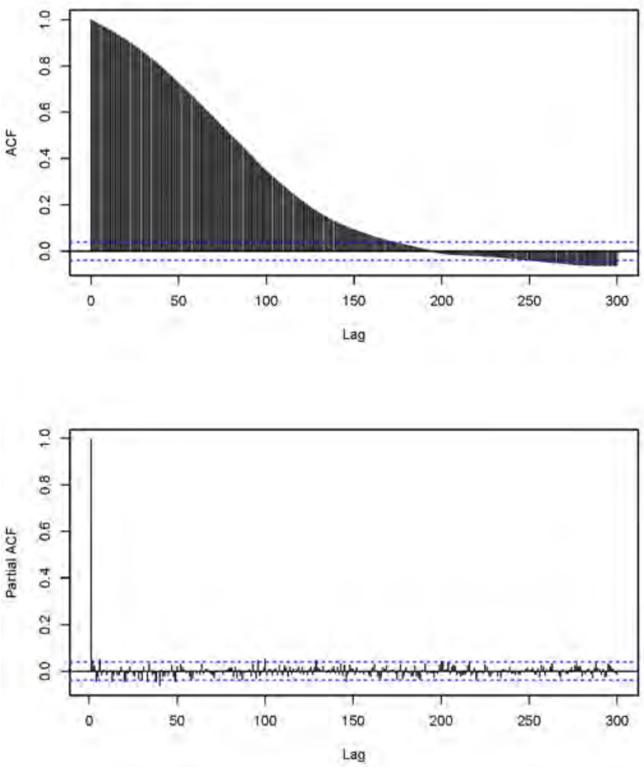

Fig. 3. Autocorrelation function and partial autocorrelation function of the data set.

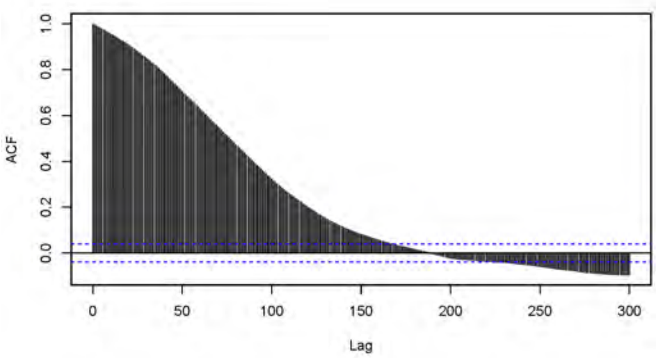

Fig. 4. Autocorrelation function of the squared observations.

applied. [13] propose the test which is much better for the distribution of the statistics of the values from the estimated models. The ARCH-LM test is a weighted portmanteau test. Under the null hypothesis it is assumed that the ARCH process is fitted accurately. The weighted ARCH-LM test does not reject the null hypothesis and thus we are able to conclude, that there is no further improvement of the ARFIMA-GARCH model order.

Figure 6 depicts the Quantile-quantile (Q-Q) plot, which

TABLE I
PARAMETER ESTIMATION OF THE OIL PRICE, WHERE THE BOLDED VALUES PROVIDE SIGNIFICANCE GIVEN A SIGNIFICANCE LIMIT OF $\alpha = 0.05$

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
Regression coefficients				
μ	55.301296	1.768863	31.26376	**0.000000**
a_{cos}	3.342449	1.544597	2.16396	**0.030467**
a_{sin}	2.566762	1.509088	1.70087	0.088967**
ARFIMA parameters				
d	0.099415	0.031782	3.12802	**0.001760**
ϕ_1	1.741240	0.000271	6419.64884	**0.000000**
ϕ_2	-0.672959	0.000128	-5247.05016	**0.000000**
ϕ_3	-0.053276	0.003060	-17.41137	**0.000000**
ϕ_4	-0.015083	0.003355	-4.49510	**0.000007**
θ_1	-0.876162	0.017276	-50.71414	**0.000000**
GARCH parameters				
α_0	0.043254	0.016544	2.61447	**0.008937**
α_1	0.081730	0.017577	4.64975	**0.000003**
γ_1	0.311167	0.269917	1.15282	0.248983
γ_2	0.217017	0.285986	0.75884	0.447949
γ_3	0.372788	0.174081	2.14146	**0.032237**
df	9.221080	1.410838	6.53589	**0.000000**
Diagnostic checking			Statistics	
AIC			3.6526	
BIC			3.6851	

shows that our considered t-distribution is a good choice for the heavy tailed residuals. This argument is supported by the Pearson chi-squared goodness of fit test.

Finally, we are able to sum up the in-sample goodness of fit. It turns out, that the model provides good in-sample performance. The model assumptions are fulfilled, Only the first lag of our squared standardized residuals shows a minor presence of autocorrelation. The next step is related to the out-of-sample forecast.

IV. FORECASTING PERFORMANCE OF THE MODEL

The Figure 7 below shows the oil prices in 2016 and our forecast up to 21 trading day. The quality of the forecast is not completely satisfying as the forecast does not reflect the cycle behaviour. Even the in-sample results are not completely satisfying. We observe that the 99%-confidence intervals are too narrow. Moreover, the realised values in future and our forecasts are different. Thereby the mean changes, but the periodic or cyclic behaviour is not captured accurately. The 99%-prediction interval is relatively narrow, which is a good result, but the observations lie outside the bounds. Nevertheless, the shortcomings of the forecast do not provide a good model result. It is obvious, that we need further improvements of the applied model. One possibility is related to the periodic behaviour which has to be captured by different and more periodic regressors. Furthermore, we can conclude that the process includes different changes which are suitable captured by a non-linear model.

V. CONCLUSION

The oil price is a very important indicator for the economy in the world. We need oil for everything so the development of the price is very important. We can find a direct correlation between the cost of gasoline or airplane fuel to the price

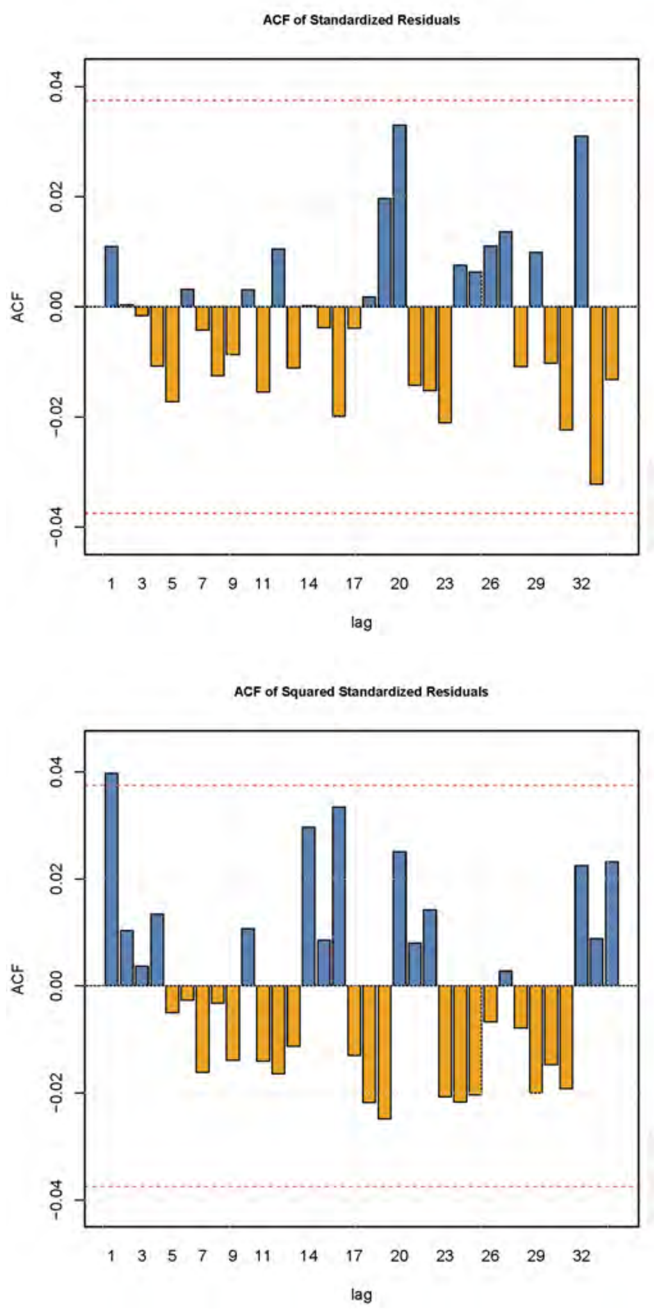

Fig. 5. Autocorrelation function of the standardised residuals and of the squared residuals.

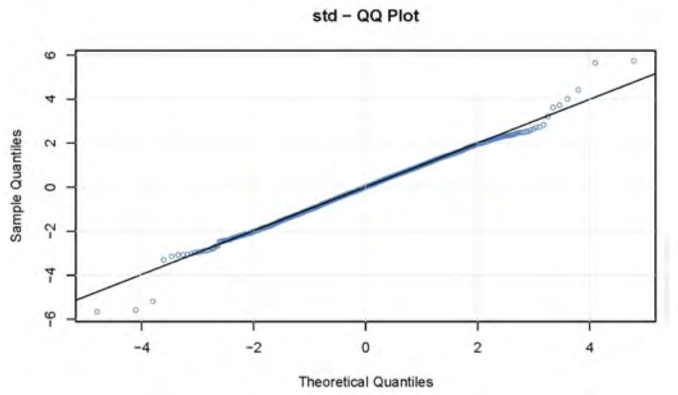

Fig. 6. Quantile-quantile plot of the residuals and the theoretical t-distribution.

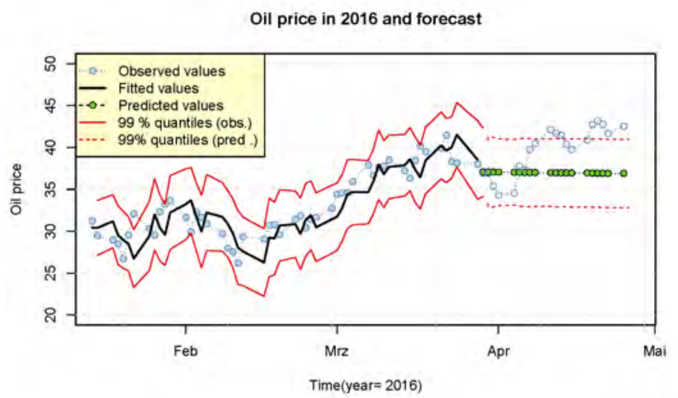

Fig. 7. Oil price forecasting for 21 future observations.

of transporting goods and people. A drop in fuel prices means lower transport costs and cheaper airline tickets, lower transportation costs for apples from Italy or furniture from China. Therefore, we discuss the application of a periodic regression model with ARFIMA-GARCH residual process to model and predict the oil price.

The aforementioned model captures the long-memory and the conditional heteroscedasticity. In addition we include two periodic regressors. The model provides some advantages, but it fails to capture the periodicity in a good way. Moreover, the model shows for the first lag of the squared standardised residuals a remaining presence of correlation, which is not satisfying at all.

The model should be extended. The volatility process could be modelled by an asymmetric or threshold GARCH model. Moreover, it might be useful that the power of the volatility is changed, so that extreme values in the volatility do not have such a big impact. The periodic regression part of the model might be extended by a stacked regularised model that captures the non-linear disturbances. The use of different periodic functions and different periods as basis has a huge influence on the structure of the model. Finally, we detect a suitable model for the oil price which provides much space for further research.

REFERENCES

[1] Akaike, H. (1974), A new look at the statistical model identification, *Automatic Control, IEEE Transactions on*, 19(6), pp. 716–723.

[2] Baillie, R.T. (1996), Long memory processes and fractional integration in econometrics, *Journal of econometrics*, 73(1), pp. 5–59.

[3] Baillie, R.T., Chung, C.F., and Tieslau, M.A. (1996), Analysing inflation by the fractionally integrated ARFIMA–GARCH model, *Journal of applied econometrics*, pp. 23–40.

[4] Bollerslev, T. (1986), Generalized autoregressive conditional heteroskedasticity, *Journal of econometrics*, 31(3), pp. 307–327.
[5] Box, G.E. and Pierce, D.A. (1970), Distribution of residual autocorrelations in autoregressive-integrated moving average time series models, *Journal of the American statistical Association*, 65(332), pp. 1509–1526.
[6] Breusch, T.S. and Pagan, A.R. (1979), A simple test for heteroscedasticity and random coefficient variation, *Econometrica: Journal of the Econometric Society*, pp. 1287–1294.
[7] Brockwell, P.J. and Davis, R.A. (2009), *Time series: theory and methods*, Springer, New York.
[8] Brockwell, P.J. and Davis, R.A. (2013), *Time series: theory and methods*, Springer Science & Business Media.
[9] Durbin, J. and Watson, G.S. (1951), Testing for serial correlation in least squares regression. II, *Biometrika*, 38(1/2), pp. 159–177.
[10] Engle, R.F. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica: Journal of the Econometric Society*, pp. 987–1007.
[11] Fahrmeir, L., Kneib, T., and Lang, S. (2007a), *Regression: Modelle, Methoden und Anwendungen*, Springer-Verlag.
[12] Fahrmeir, L., Künstler, R., Pigeot, I., and Tutz, G. (2007b), *Statistik: Der Weg zur Datenanalyse*, Springer-Verlag.
[13] Fisher, T.J. and Gallagher, C.M. (2012), New weighted portmanteau statistics for time series goodness of fit testing, *Journal of the American Statistical Association*, 107(498), pp. 777–787.
[14] Goldfeld, S.M. and Quandt, R.E. (1965), Some tests for homoscedasticity, *Journal of the American statistical Association*, 60(310), pp. 539–547.
[15] Greenwald, B.C., Stiglitz, J.E., and Weiss, A. (1984), Informational imperfections in the capital market and macro-economic fluctuations.
[16] Haslett, J. and Raftery, A.E. (1989), Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource, *Applied Statistics*, 30(1), pp. 1–50.
[17] Ishida, I., Watanabe, T., et al. (2009), Modeling and Forecasting the Volatility of the Nikkei 225 realized Volatility using the ARFIMA-GARCH model, *Global COE Hi-Stat Discussion Paper*, 32.
[18] Kalkman, J., Pfeiffer, W., and Pereira, S. (2013), Are we running out of oil?
[19] Kane, I.L. and Yusof, F. (2013), Assessment of Risk of Rainfall Events with a Hybrid of ARFIMA-GARCH, *Modern Applied Science*, 7(12), p. 78.
[20] Koopman, S.J., Ooms, M., and Carnero, M.A. (2007), Periodic seasonal Reg-ARFIMA–GARCH models for daily electricity spot prices, *Journal of the American Statistical Association*, 102(477), pp. 16–27.
[21] Leite, A., Rocha, A., and Silva, M. (2009), Long memory and volatility in HRV: an ARFIMA-GARCH approach, *Computers in Cardiology, 2009*, IEEE, pp. 165–168.
[22] Mandelbrot, B.B. and Van Ness, J.W. (1968), Fractional Brownian motions, fractional noises and applications, *SIAM review*, 10(4), pp. 422–437.
[23] Palm, F.C. (1996), 7 GARCH models of volatility, *Handbook of statistics*, 14, pp. 209–240.
[24] Schwarz, G. et al. (1978), Estimating the dimension of a model, *The annals of statistics*, 6(2), pp. 461–464.
[25] Shapiro, S.S. and Wilk, M.B. (1965), An analysis of variance test for normality (complete samples), *Biometrika*, 52(3/4), pp. 591–611.
[26] Shumway, R.H. and Stoffer, D.S. (2010), *Time series analysis and its applications: with R examples*, Springer Science & Business Media.
[27] White, H. (1980), A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica: Journal of the Econometric Society*, pp. 817–838.