Production Analysis of Oilfield Data

RaviKumar, Arjun[*] Podhoretz, Seth[†] Kou, Rui[‡] Feng, Gan[§] and Gupta, Vivek[¶]

Department of Statistics, Texas A & M University

July 8, 2018

Abstract

Our project aims to utilize the abundance of oilfield data to arrive at conclusions on the physical and economic aspects of the flow of hydrocarbon in the reservoir. Physical aspects that we aim to analyze include the prevailing flow regime (the character of flow with respect to geometry and pressure drop), interference between wells, and communication between wells. Economic aspects include forecasting of production rates into the future, and estimating the ultimate hydrocarbon recovery volumes.

The fluid that is of interest resides in pores of rock, several thousands of feet under the surface. As this fluid flows through the rock and into the well, the pressure in the rock drops, and the production rate drops. The nature of the production rate drop, and the backpressure held at the well together carry information on the physics of the process: the flow regime, possible interaction with another well, presence of boundaries in the reservoir. With these physical characteristics in mind, we can forecast the production rate into the future, thus calculating the economic life of the well, and the economics of production.

Our dataset at the moment consists of 60 wells from the Bakken. The Bakken is a shale play in North Dakota, and one of the largest oil developments in recent history. It helped start the shale boom in the United States, reversing decades of declining US oil production volumes.

[*] aravikumar@tamu.com

[†] sbpodhoretz@gmail.com

[‡] kourui.pete@tamu.edu

[§] fenggan@tamu.edu

[¶] vivek235@tamu.edu

1 Introduction

Some Introduction to be filled by subject matter experts on the problem (Arjun et al.?)

2 Data Description

Dataset presented consists of 60 distinct wells in the <> area. The dataset has information about Monthly averages of Oil (in bbl) , Gas (in mcf, Thousand Cubic Feet) and Water (in bbl , barrels). There is also information about Daily production of these quantities. We are interested in modeling the stochastic processes , Monthly Average of Gas , Oil and Water.

Density plots of these quantities as shown in Figure 1 indicates high variance and right skewed distribution in data. Similar are the observations from time series plot as shown in Figure 2.

Sample data summary statistics for a selected well is shown below as well. It is clear from the sample summary that the data has great deal of variance in it for all the three processes that are of interest to model and forecast.

Process	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Kurtosis	Skewness
Avg. Gas(bbl)	4.00	14.00	19.00	29.19	31.00	204.00	9.63	3.02
Avg. Oil(mcf)	9.32	26.00	37.00	75.74	76.50	642.00	10.01	3.03
Avg. Water(bbl)	2.00	6.00	9.00	11.37	12.75	60.00	10.2	1.96

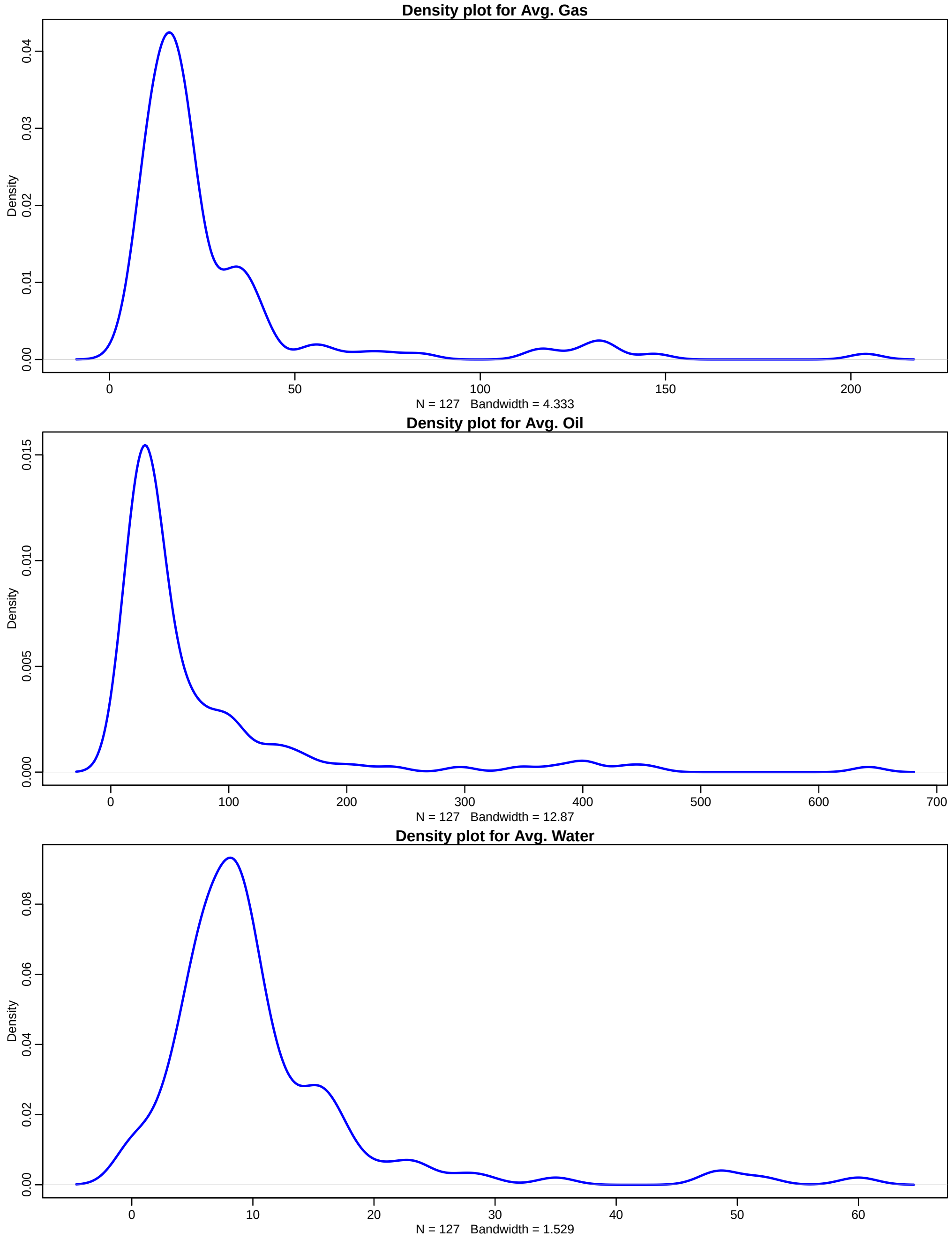


Figure 1: Density Plots of Avg. Gas , Oil and Water

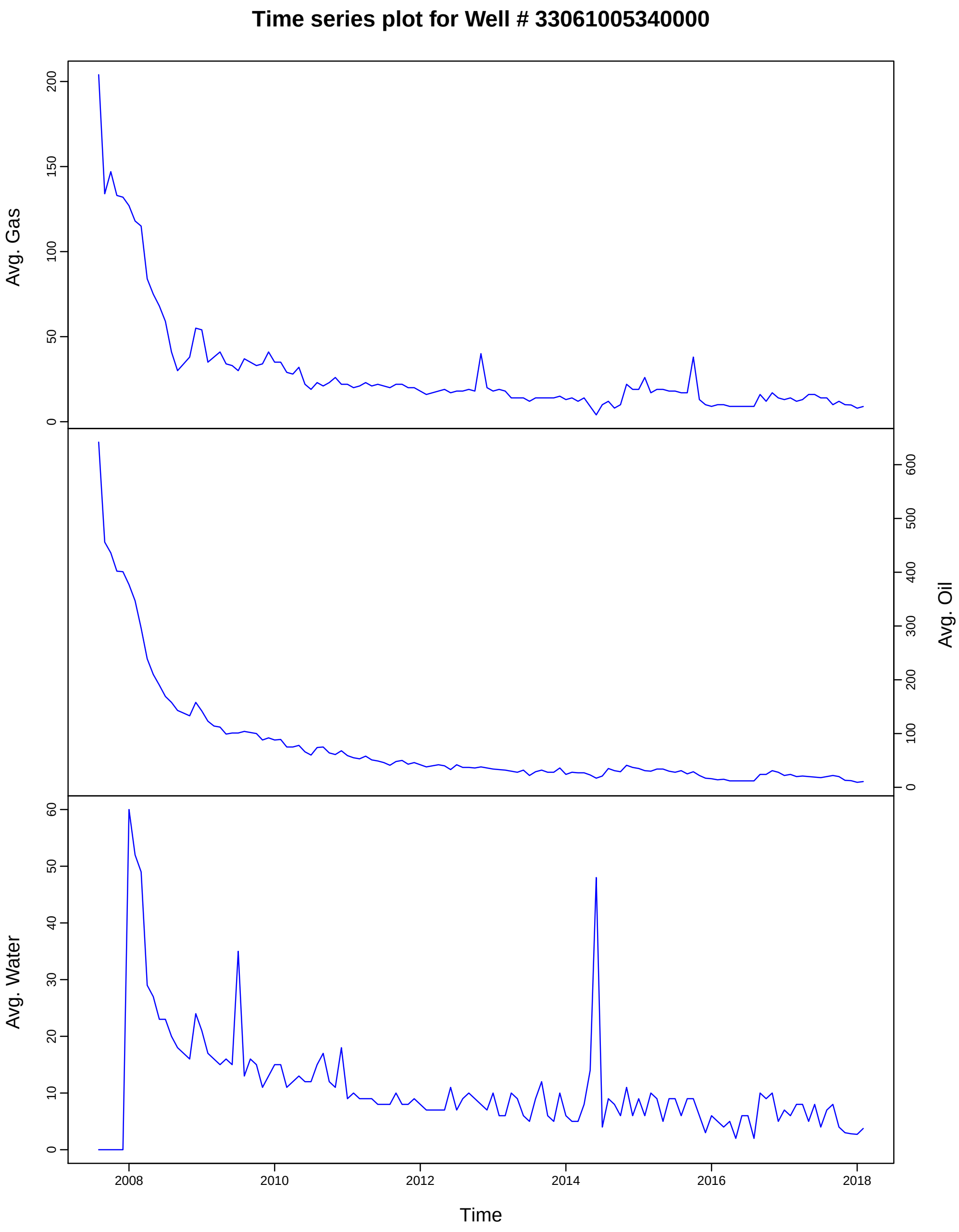


Figure 2: Time Series Plots of Avg. Gas , Oil and Water

3 Data Analysis

3.1 Model Formulation

We consider two broad categories of model formulation for modelling the stochastic processes of Avg. Oil, Avg. Gas and Avg. Water. These are

1. **Semi - parametric family of regression [1] [4]**

   Here at start, given that the function of the processes are non linear with respect to time as shown in Figure 2, we will use Bspline Basis Functions with a modification suggested by Finbarr O'Sullivan of University College Cork, in Ireland called cubic O'Sullivan splines. The O-splines approximate the underlying regression functions as

   $$f(x) = \beta_0 + \beta_1 x + \sum_{1}^{K} u_k z_k(x)$$

   where x is time , $z_k(.)$ are the O - Sullivan Splines. $f(x)$ is linear if $u_1 = u_2 \cdots = 0$.

   With a larger value of K and appropriate penalization, $f(x)$ can approximate very complex shapes.

2. **ARIMA based models**

Standard model selection and estimation procedures as learnt under SARIMA modelling techniques , which is

$$\Phi_P(B^s)\phi_p(B){\nabla_s}^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t$$

where

$\Phi_P(B^s)$ and $\phi_p(B)$ are seasonal and non-seasonal Auto-Regressive Operators of order P and p respectively.
${\nabla_s}^D$ and∇^d are seasonal and non-seasonal difference operator respectively.
$\Theta_Q(B^s)$ and $\theta(B)$ are seasonal and non-seasonal Moving Average Operators of order Q and q respectively.

For model formulation we follow Exploratory Data Analysis techniques using difference to de-trend the series and plot ACF and PACF to get a set of plausible values of p, q, d, P, Q, D and S.

Given that the density of the observations are highly non normal with high variance as shown in Figure 1 and 2, we transform the series to logarithmic scale to contain both of these. The transformed and differenced series is plotted in Figure 3 and 1. We notice non-normality in data and while, better transforms are available to normalize and contain the variance we choose to stick to log transforms since we can easily back transform the forecasts to original scale along with standard errors.

We look individually at ACF and PACF plots for the three processes to get plausibal set of values of the orders of the model.

(a) **Avg. Gas**
Figure 5 shows involvement of seasonal components at lag 1 and 2. We note that its almost non significant at

$$\alpha = 0.05$$

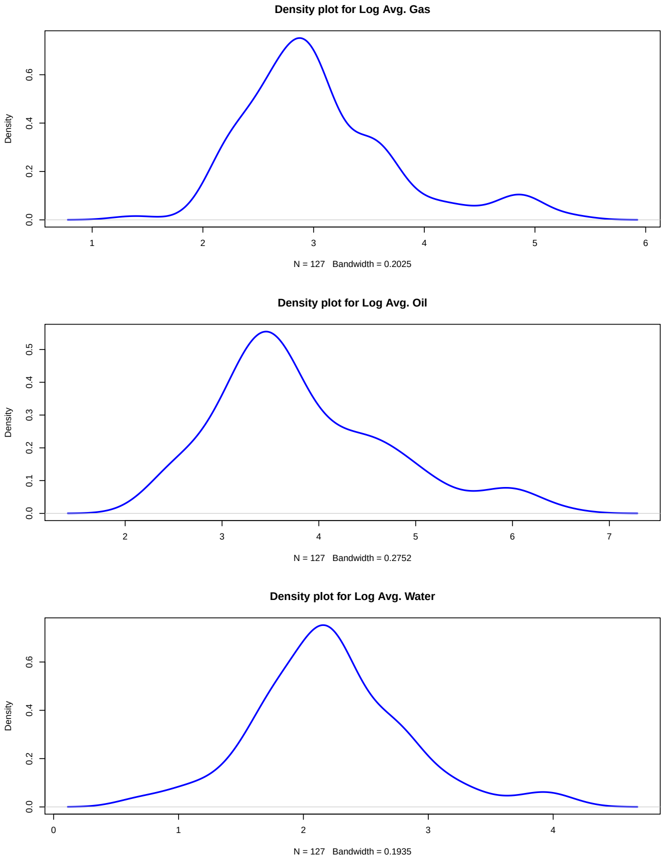
. We also see presence of effects of AR and MA non-seasonal components. Thus the plausible set of values of the orders for which we estimate and diagnose the model are for S = 12 , 0 , Q = 2 , 0 , P = 0 , D = 0 , d = 1 , q = 1 ,4 , p = 2.

(b) **Avg. Oil**
Figure 6 does not show any interesting affects of correlation. We notice weakly significant

$$\hat{\rho}(11)$$

so we diagnose and estimate the model with both q = 0 and 11, d=1.

(c) **Avg. Water**
Figure 7 shows involvement of non-seasonal components only. Thus the plausible set of values of the orders for which we estimate and diagnose the model are for S=0 , P = Q = 0 , p = 1 , 2 , q = 1 , d = 1.



Density plot for Log Avg. Gas

Density plot for Log Avg. Oil

Density plot for Log Avg. Water

Figure 3: Density of the Time Series after Log

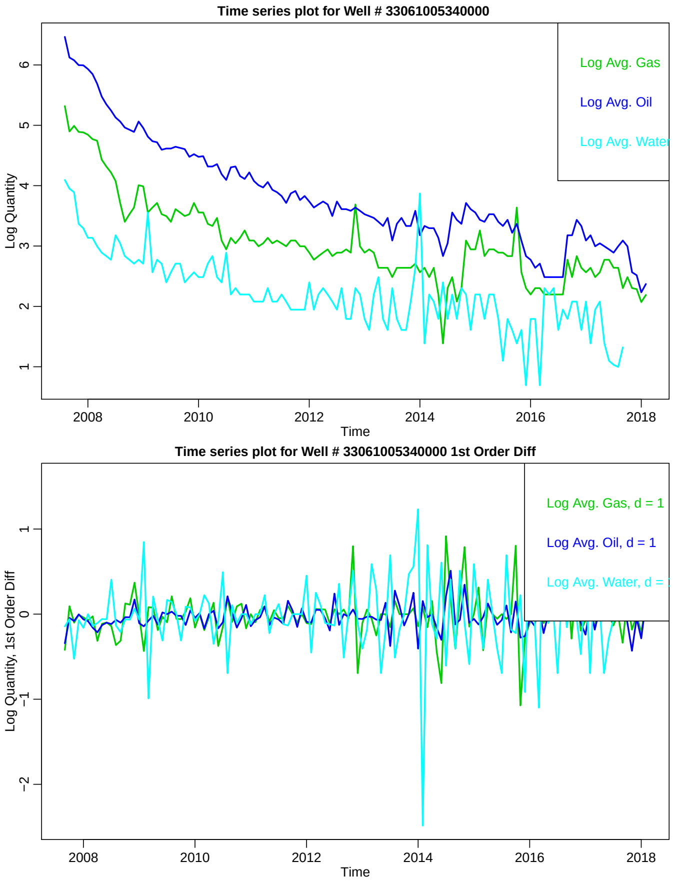


Figure 4: Time Series after transformation and Difference

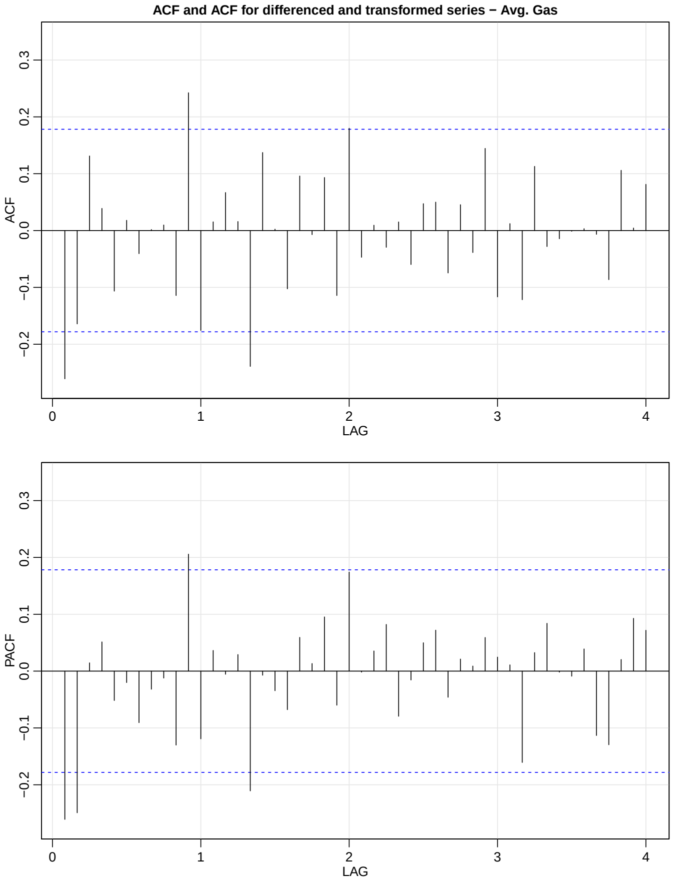


Figure 5: ACF and PACF plot for Avg. Gas

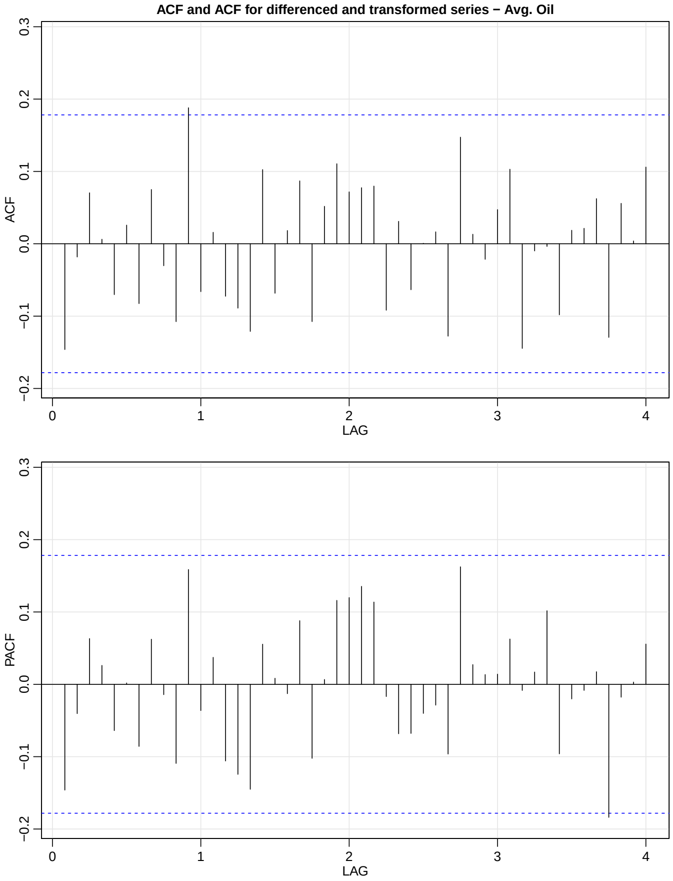


Figure 6: ACF and PACF plot for Avg. Oil

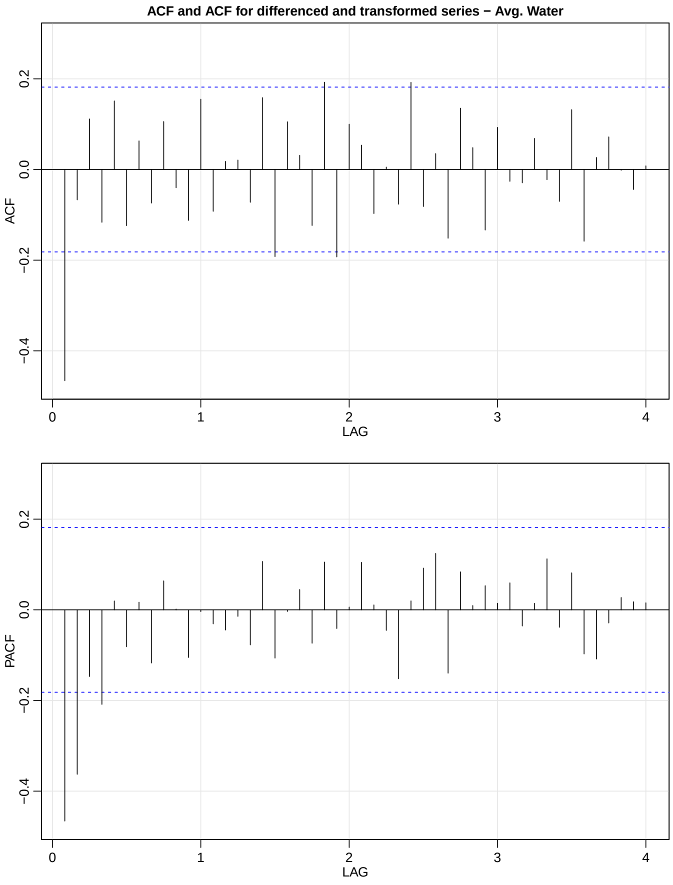
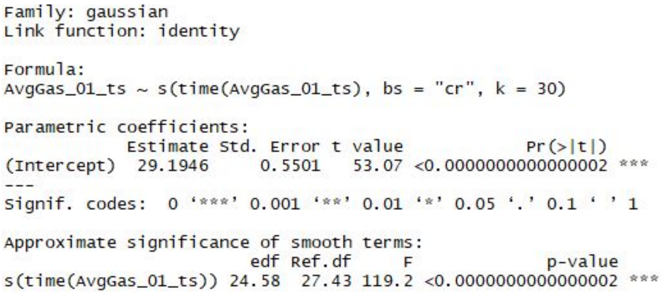


Figure 7: ACF and PACF plot for Avg. Water



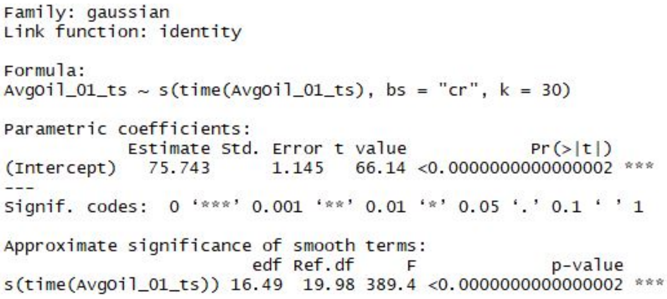
```
Family: gaussian
Link function: identity

Formula:
AvgGas_01_ts ~ s(time(AvgGas_01_ts), bs = "cr", k = 30)

Parametric coefficients:
            Estimate Std. Error t value           Pr(>|t|)
(Intercept)  29.1946     0.5501   53.07 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                         edf Ref.df     F             p-value
s(time(AvgGas_01_ts)) 24.58  27.43 119.2 <0.0000000000000002 ***
```

Figure 8: Semi-par regression of Avg. Gas with Time



```
Family: gaussian
Link function: identity

Formula:
AvgOil_01_ts ~ s(time(AvgOil_01_ts), bs = "cr", k = 30)

Parametric coefficients:
            Estimate Std. Error t value           Pr(>|t|)
(Intercept)   75.743      1.145   66.14 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                         edf Ref.df     F             p-value
s(time(AvgOil_01_ts)) 16.49  19.98 389.4 <0.0000000000000002 ***
```

Figure 9: Semi-par regression of Avg. Oil with Time

3.2 Model Estimation

1. **Semi - parametric family of regression [1] [4]**

   We look at modeling the data via package mgcv [4]. The mathematical details, derivations and estimations are given referenced in [1] , [3] and [2]

   A semi-parametric model was thus fit to a particular well number 33061005340000 and estimates were obtained for all the three processes. The model summaries are as below.

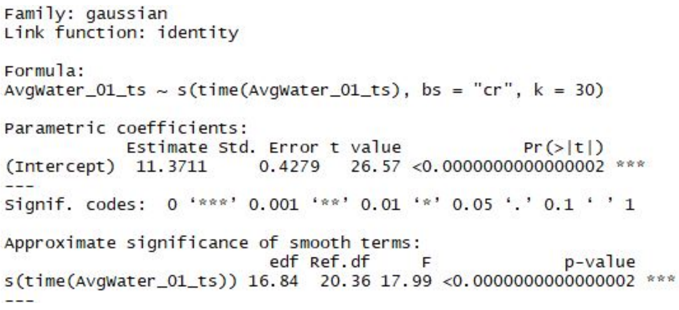
   (a) Avg. Gas : Figure 8 shows the model formulation and estimation of parameters via R . We see a significant effect of time as we anticipate.

   (b) Avg. Oil : Figure 9 shows the model formulation and estimation of parameters via R . We see a significant effect of time as we would anticipate.

   (c) Avg. Water : Figure 10 shows the model formulation and estimation of parameters via R . We see a significant effect of time as we would anticipate.

   Note that in all of the above fits we have chosen the number of basis functions to be used as 30. This was arrived at after running few diagnostic procedures during the model formulation. Since the number was very high, so was the number of estimated parameters. The estimated values of these are thus not shown in this document

   The fitted values for all of the above fits are plotted in Figure 11. We notice that the fitted values follow closely the observations in all the three processes. There are, however, observations where fits perform poorly and may be contributing to the large value of cross validation error as reported by mgcv. This occurs specially for Water dataset. We also notice observations that does not follow the usual pattern in the data set.



```
Family: gaussian
Link function: identity

Formula:
AvgWater_01_ts ~ s(time(AvgWater_01_ts), bs = "cr", k = 30)

Parametric coefficients:
            Estimate Std. Error t value              Pr(>|t|)
(Intercept)  11.3711     0.4279   26.57 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                          edf Ref.df     F             p-value
s(time(AvgWater_01_ts)) 16.84  20.36 17.99 <0.0000000000000002 ***
---
```

Figure 10: Semi-par regression of Avg. Water with Time

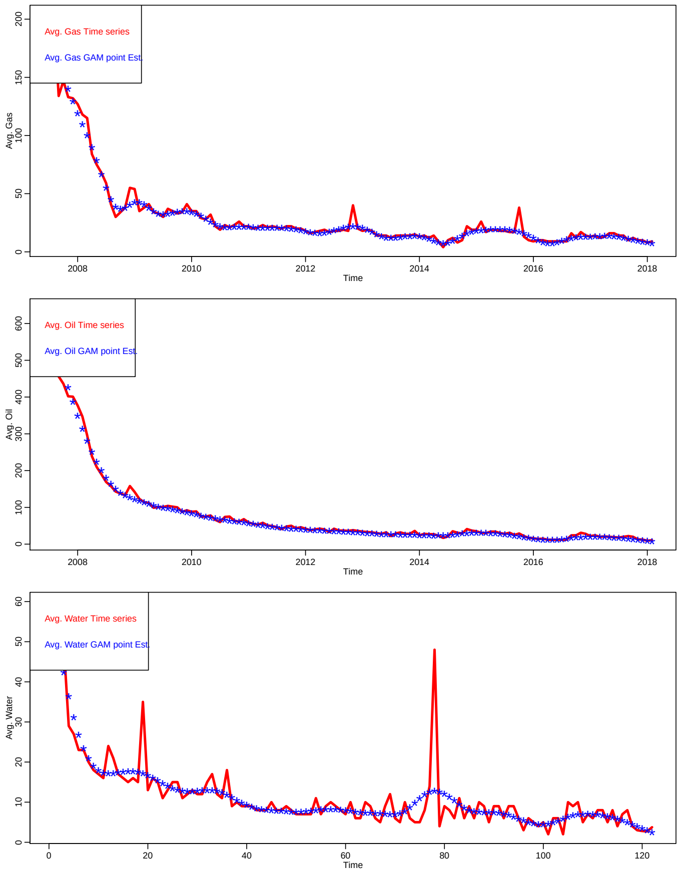


Figure 11: Fiited vs Observed via GAM model

2. **ARIMA based models**

We estimate the parameters of the model using `astsa` package in `R`.

For Avg. Gas process we settle with a simpler model with q = 1 and d = 1 and rest all 0 based on lowest value of AIC, BIC and AICc. For brevity, the model estimated parameters is omitted in this document. We proceed with this model for forecast Avg. Gas but we caution that a more robust procedure multi- fold cross validation might be necessary to further validate the model.

For Avg. Oil process we settle with a simpler model with q = 1 and d = 1 and rest all 0 based on lowest value of AIC, BIC and AICc. For brevity, the model estimated parameters is omitted in this document. We proceed with this model for forecast Avg. Gas but we caution that a more robust procedure multi- fold cross validation might be necessary to further validate the model.

For Avg. Water process we settle with a simpler model with q = 1 and d = 1 and rest all 0 based on lowest value of AIC, BIC and AICc. For brevity, the model estimated parameters is omitted in this document. We proceed with this model for forecast Avg. Gas but we caution that a more robust procedure multi- fold cross validation might be necessary to further validate the model.

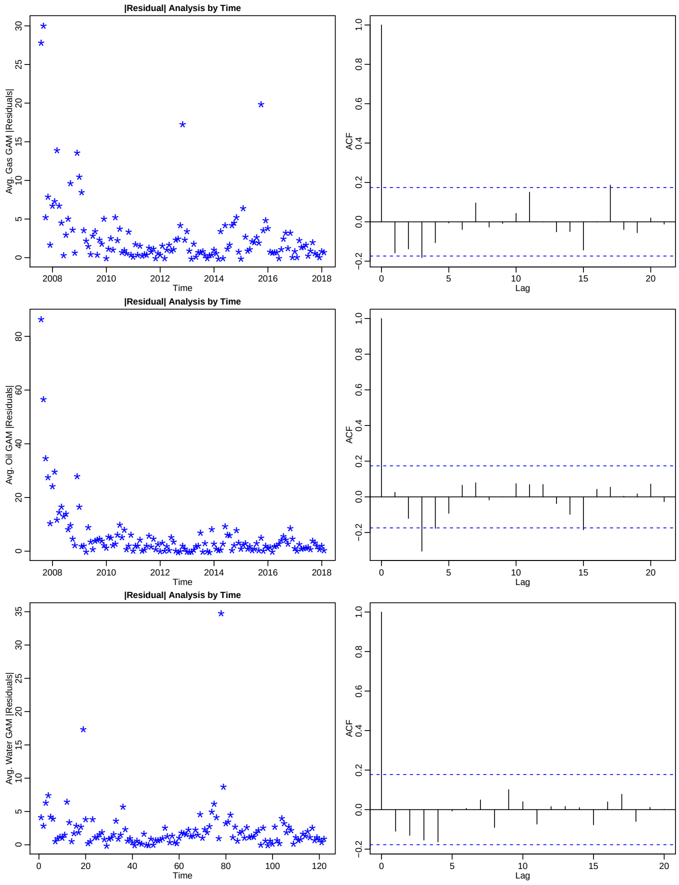


Figure 12: |Residuals| in all the three fits

3.3 Model Diagnostics

1. **Semi - parametric family of regression [1] [4]**

The residuals from the fits in the Model 8 , 9 and 10 were extracted and plotted with respect to time to find any visual evidence of patterns in the residuals. Their ACF plots was also obtained to provide support to the assumptions of randomness in the residuals. Both these plots are shown in the Figure 12 . The plots are for all the three fits. We notice that plots indicate a significant autocorrelation of the order 3 in Avg. Oil model indicating a need for modeling serially correlated errors as well. We will analyze this again in ARIMA model to confirm if we have a better model. We also notice the presence of influential points in all the three timeseries further indicating that we may need to understand the behavior of the processes around these points. At the moment we will ignore and accept these anomalies in the dataset with the caveat that we may be missing to account for the scenario when these anomalies occur.

`Ljung-Box` test for whiteness in the residuals at values of lag 30 , 50 and 100 provide evidence, at critical value of $\alpha = 0.05$, to support a fairly good fit to whiteness.However, we note that test of whiteness fails at lag of order 10 for Avg. Oil process when modeled with this approach. Plot 13 shows the p-values obtained at various lags.

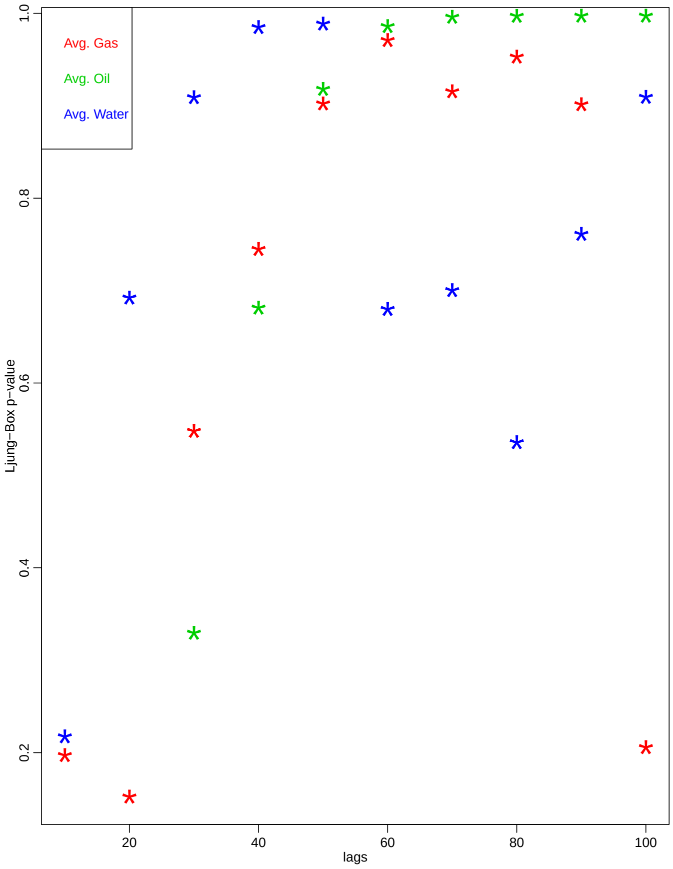


Figure 13: Ljung-Box test p.value for the three series,$\alpha = 0.05$

2. **ARIMA based models**

We individually take a look at model diagnostics of the three processes of interest.

(a) **Avg. Gas**

Figure 14 shows various plots detailing the analysis of the time series AVg. Gas. The test of Whiteness for the model looks positive indicating a good fit to the white noise. Normality in residuals is questionable and hence prediction intervals obtained from the point estimates of the forecasts might not be reliable. We will have to address this issue before we can forecast using this model. We also notice some significant auto correlation in residuals which we will have to address. We think that having some explaining variables accounting information on anomalies will help us to address this issue.

Overall, while the fit of the model looks great the normality assumptions of the residuals does not look encouraging.

(b) **Avg. Oil**

Figure 15 displays the check on residuals of the fitted values by the ARIMA model. We notice that we have great fit to white noise from the `Ljung-Box` test. The normality assumptions on the residuals is acceptable and hence indicating a good and simple potential model.

(c) **Avg. Water**

Figure 16 displays the check on residuals of the fitted values by the ARIMA model. We notice that we have great fit to white noise from the `Ljung-Box` test. The normality assumptions on the residuals is acceptable and hence indicating a good and simple potential model.

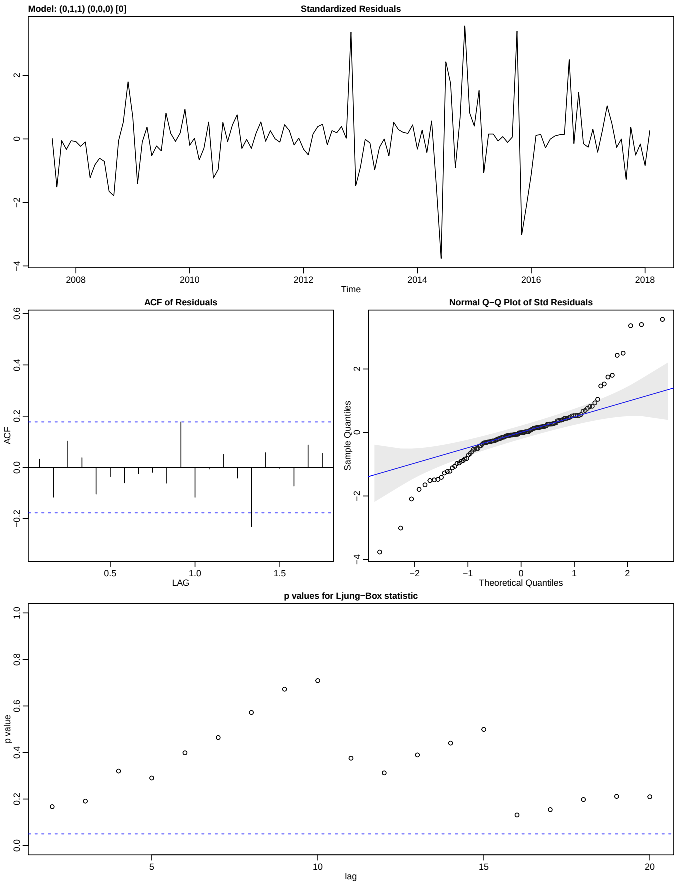


Figure 14: Model Diagnostics , Avg. Gas

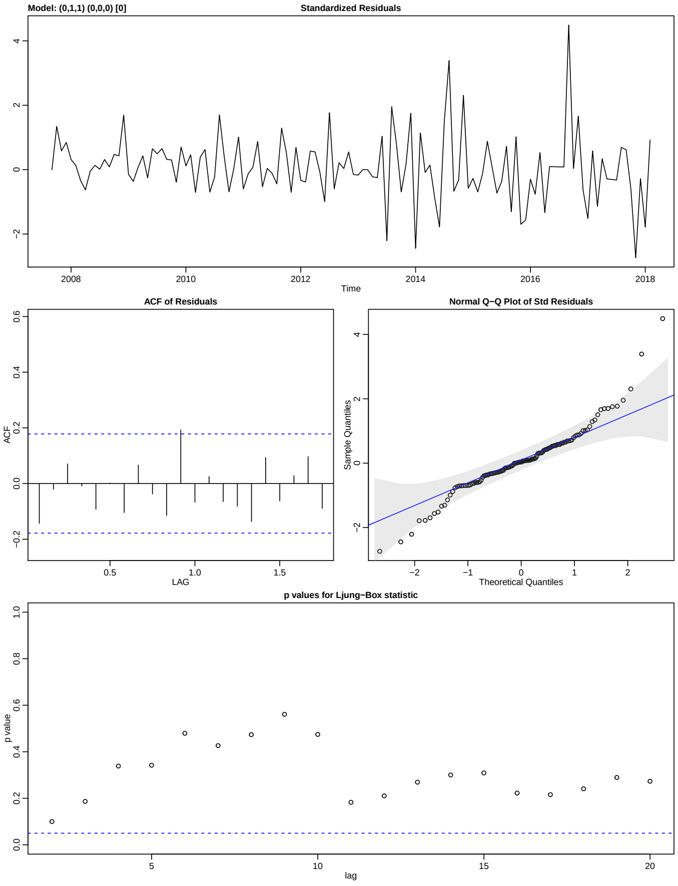


Figure 15: Model Diagnostics , Avg. Oil

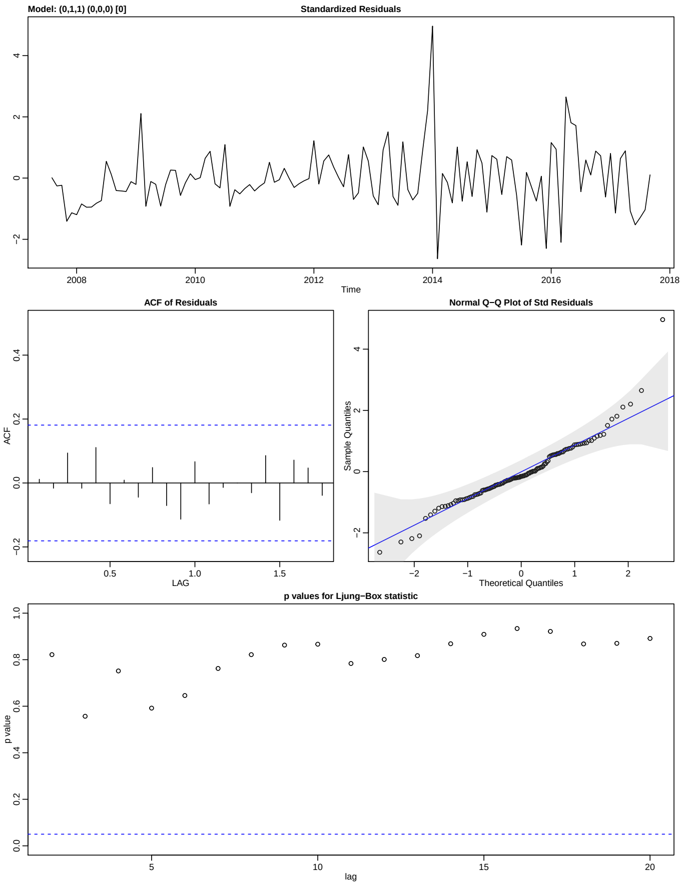


Figure 16: Model Diagnostics , Avg. Watter

4 Conclusion

We arrive at following learning while determining and estimating a model for forecasting time series.

1. Semi-Parametric family of regression worked very well for Avg. Gas and Avg. Water modeling. We have to crossvalidate (1 step or m step ahead) to determine the model accuracy.

2. ARIMA based modeling works very well when modeling for serially correlated errors as was with Avg. Oil where semi-parametric regression did not do so well.

3. ARIMA based approach for modeling all of these processes led to valid and reliable models. We will still need to very forecasts.

4. There are elements in the dataset which appear anomalous and not random which gives an intuition of existence of external factors influencing the production of Gas, Water and Oil.

5. We had to strip leading 0 measurements from Avg. Oil dataset for estimating and fitting the model.

6. We found that running power transform on the data yielded the best results on reduction to white noise with the simplest models on majority of the wells. However, because of issues related to back transformation to original scale we prefer to use ARMA or GAM based models.

References

[1] David Ruppert, M.P. Wand, and Raymond J. Carroll. Semiparametric regression during 2003–2007. *Electron. J. Statist.*, 3:1193–1256.

[2] S. N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686, 2004.

[3] S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

[4] S.N Wood. *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC, 2 edition, 2017.