Time series analysis of electricity production in Australia

- Student : Margaux Mouyard – r0826714
- Course : Advanced Time Series Analysis- D0M63B
- Professor : Christophe Croux

Presentation of the dataset

Definition of the time series and the business goals

- Dataset from the « tsibbledata » package in R[1]
- Production statistics of multiple goods in Australia
- Quartely data from 1956 Q1 to 2004 Q4 (196 observations)

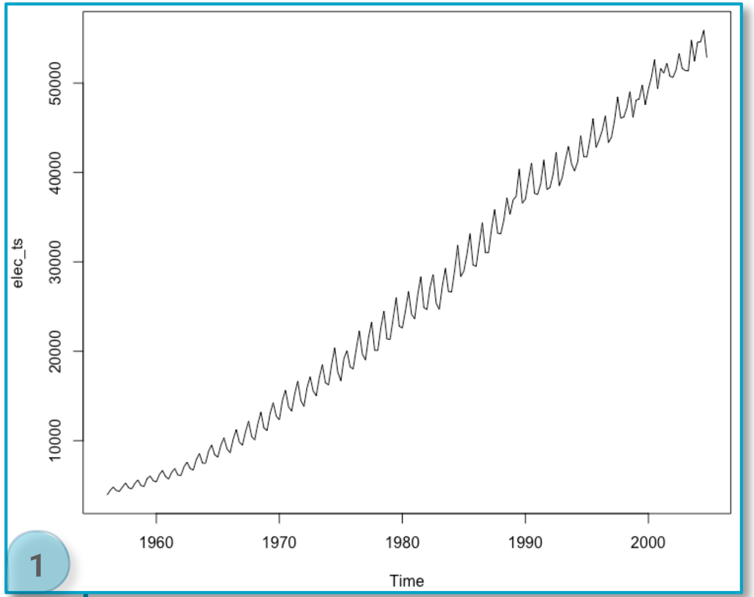
- **For this analysis, we will focus on two time series** :
  - the production of **electricity** (gigawatt hours)
  - the production of **gas** (petajoules)

- **Business questions to be study in this analysis** ;
  - Is there a particular quarter where more electricity is produced each year ?
  - What is the forecasted production in electricity for the next quarters ?
  - Does the production of gas at a time t-1 have an impact on the production of electricity at time t ?

[1] https://cran.r-project.org/web/packages/tsibbledata/tsibbledata.pdf

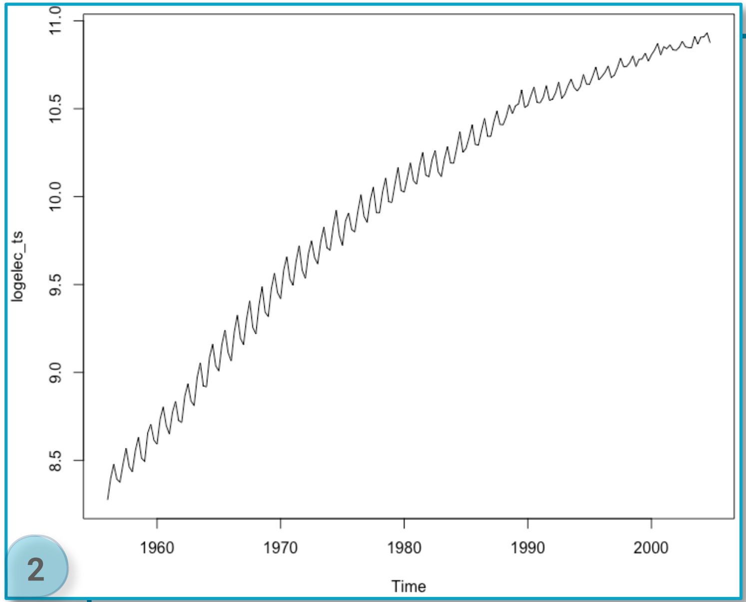
Univariate analysis

Definition and analysis of electricity time series : « elec_ts »



We can observe that :

- There is a positive trend
- The time series is not stationary
- The variance is not constant over time



Following the observations, the series is computed in logs

ADF test

```
data:  logelec_ts
ADF(12) = -0.96307, p-value = 0.9453
alternative hypothesis: true delta is less than 0
sample estimates:
      delta
-0.00889602
```

A unit root test is performed to check whether the trend in the log-transformed series is deterministic or stochastic.

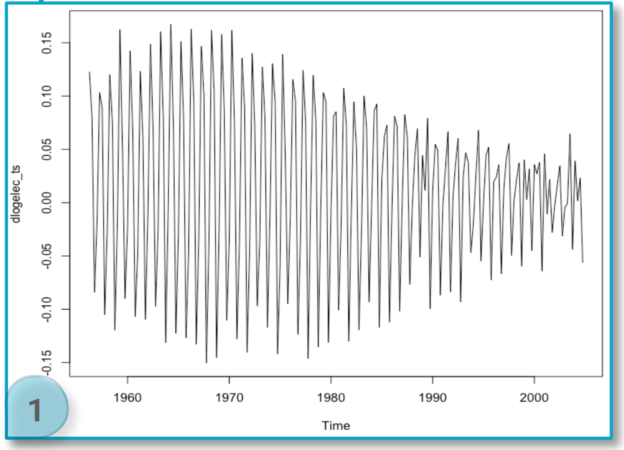
The trend is stochastic as the p-value > 5%

We go in (log) differences

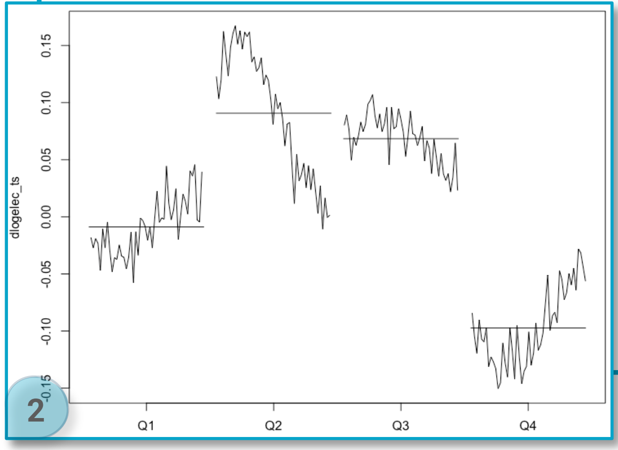
Univariate analysis

Definition and analysis of electricity time series in differences

elec_ts in log-differences



Seasonal plot



There is a clear seasonal pattern

We go in seasonal differences: « ddlogelec »

ADF test

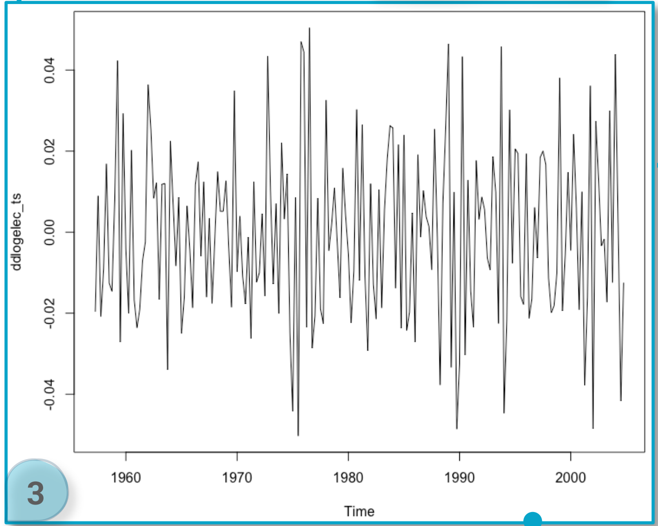
stationary (H0 rejected)

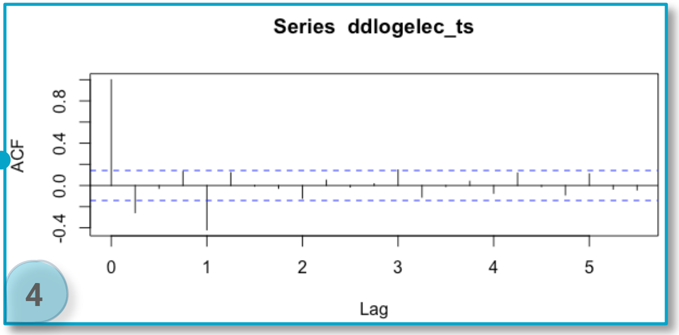
```
data:  ddlogelec_ts
ADF(7) = -8.6672, p-value = 7.016e-13
alternative hypothesis: true delta is less than 0
sample estimates:
    delta
-2.677696
```

Box-Ljung test

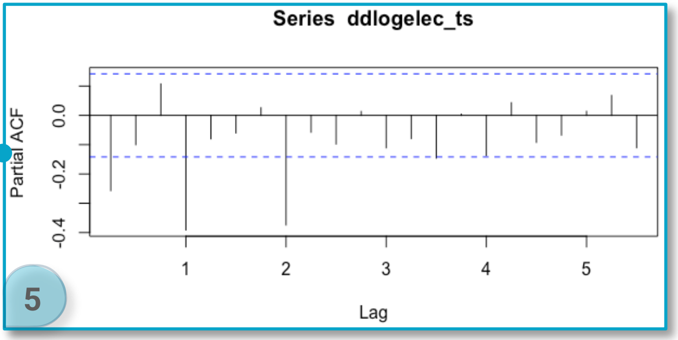
not white noise (H0 rejected)

```
data:  ddlogelec_ts
X-squared = 64.775, df = 14, p-value = 1.678e-08
```





MA(1) structure and indication of order 1 seasonality



AR(1) structure and indication of order 2 seasonality

Univariate analysis

Summary of model exploration and comparison

Model ID	Model	Valid? Ljung-Box	AIC	BIC	MAE	MSE	Comments
fit_sar1	SARIMA(1,1,0)(1,1,0)	No	-969.3	N/A	N/A	N/A	
fit_sar2	SARIMA(1,1,0)(2,1,0)	Yes	-995.8	-982.7	0.019	0.001	
fit_sar3	SARIMA(2,1,0)(2,1,0)	Yes	-998.5	-982.1	0.061	0.004	
fit_sma1	SARIMA(0,1,1)(0,1,1)	Yes	-996.1	-986.2	0.027	0.001	
fit_sma2	SARIMA(0,1,2)(0,1,1)	Yes	-998.2	-985.1	0.034	0.002	
fit_sarima1	SARIMA(1,1,1)(2,1,1)	Yes	-1002.1	-982.5	0.018	0.001	SMA(1) term not significant
fit_sarima2	SARIMA(1,1,1)(2,1,0)	Yes	-1002.1	-985.7	0.018	0.001	Selected as final model
fit_sarima3	SARIMA(2,1,2)(2,1,1)	Yes	-1000.5	-974.3	0.028	0.001	AR(1), AR(2) & SMA(1) terms not significant

- *fit_sarima2* is chosen as final model because it has the best combination AIC/BIC (smallest AIC and second smallest BIC) and because it performs better in MAE and MSE than all other candidates (smallest for both).

Univariate analysis

Analysis of the selected model SARIMA(1,1,1)(2,1,0)

Coefficients analysis

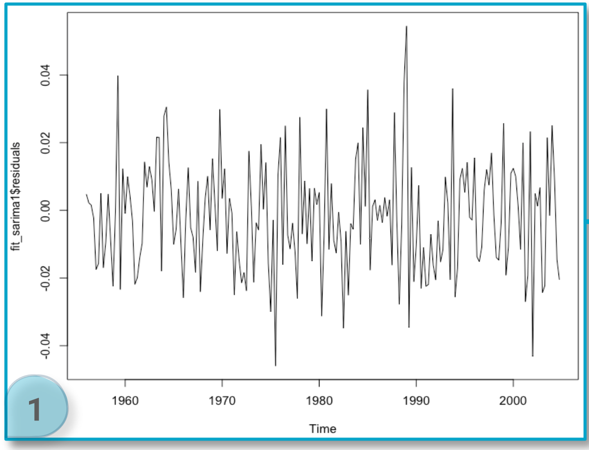
```
Call:
arima(x = logelec_ts, order = c(1, 1, 1), seasonal = list(order = c(2, 1, 0)))

Coefficients:
         ar1      ma1     sar1     sar2
      0.6047  -0.8690  -0.5432  -0.3568
s.e.  0.1138   0.0753   0.0722   0.0708

sigma^2 estimated as 0.0002885:  log likelihood = 506.03,  aic = -1002.07
```

- Verification of the coefficients significance (coef/sd) :

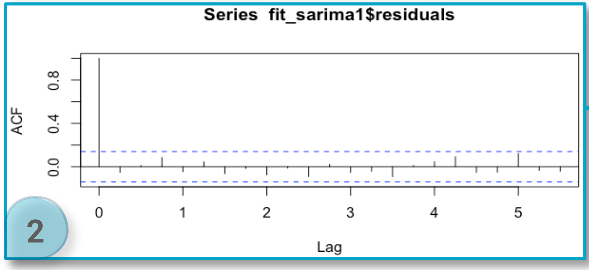
```
     ar1       ma1      sar1      sar2
5.311902 11.545463  7.525119  5.040142
```

- Every coefficients are significant (> 2)

Residuals analysis



- The residuals seem white noise



- No significant autocorrelation

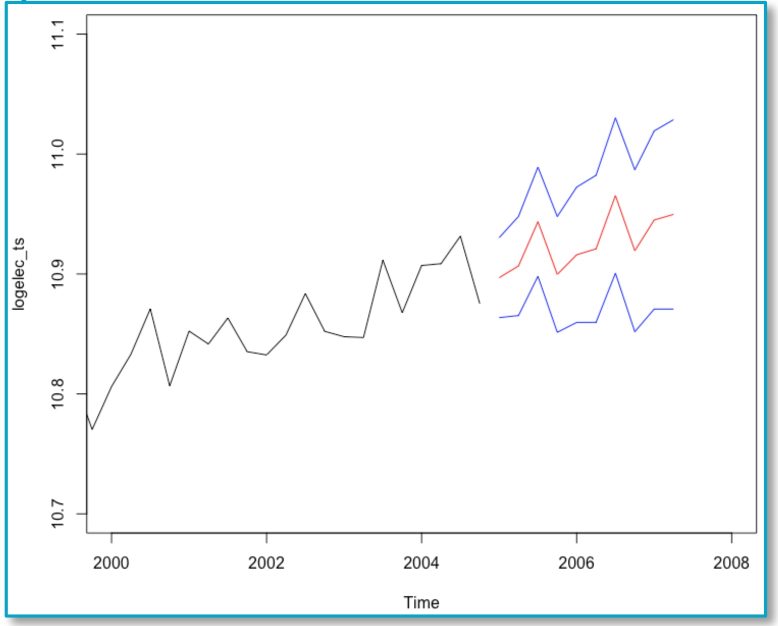
```
        Box-Ljung test

data:  fit_ar5$residuals
X-squared = 11.565, df = 14, p-value = 0.6412
```

- H0 not rejected : residuals are indeed white noise and the model is valid

Electricity production forecast

Projection for the next 10 quarters (2005 Q1 – 2007 Q2) and comparison

Forecast based on model *fit_sarima2*



- The forecast shows an upward trend and peaks in the Q3 of 2005 and 2006, which corresponds to the winter period in Australia. This respects the seasonal pattern.
- The prediction interval is close to the forecasted values and it widens over time because the further away we forecast, the more uncertain our forecast will be.

Comparison example

Comparison of MSE between *fit_sma2* and *fit_sarima2* :

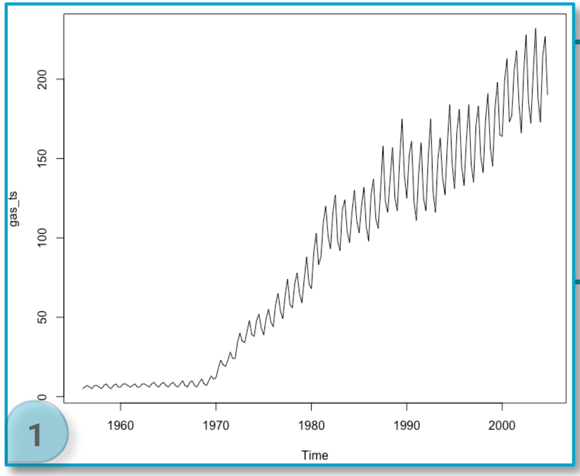
```
        Diebold-Mariano Test

data:  error1.herror2.h
DM = 3.8299, Forecast horizon = 1, Loss function power = 2, p-value = 0.0003716
alternative hypothesis: two.sided
```

Model ID	Model	MSE
fit_sma2	SARIMA(0,1,2)(0,1,1)	0.002
fit_sarima2	SARIMA(1,1,1)(2,1,0)	0.001

- p-value < 5% = H0 is rejected : the forecasting performances of the two models are significantly different
- The forecasting performance of *fit_sarima2* is better as its MSE is lower than the one of *fit_sma2*

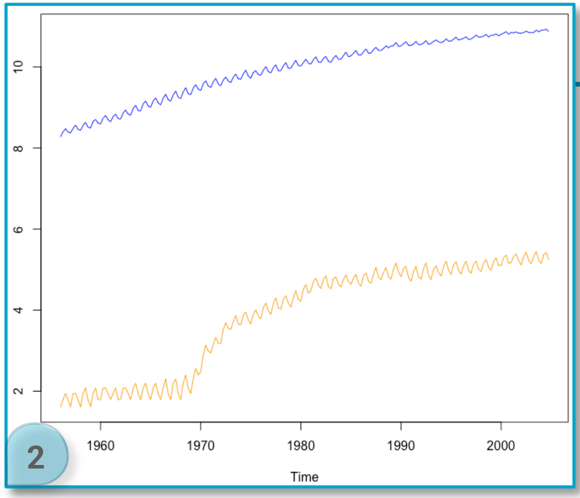
Multivariate analysis

Introducing a new time series : « gas_ts »



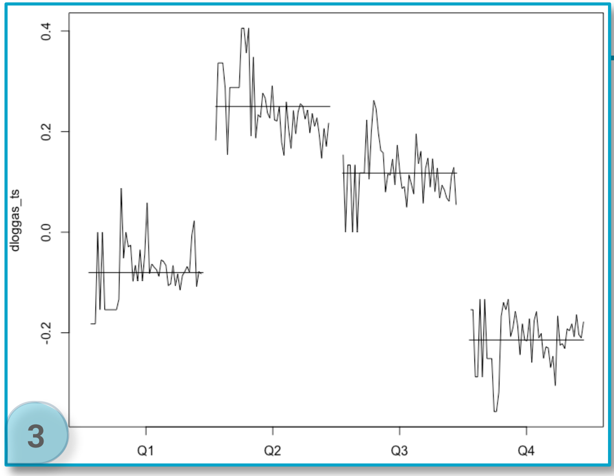
gas_ts represents the production of gas (petajoules) in Australia by quarter

- positive trend
- not stationary
- variance not constant

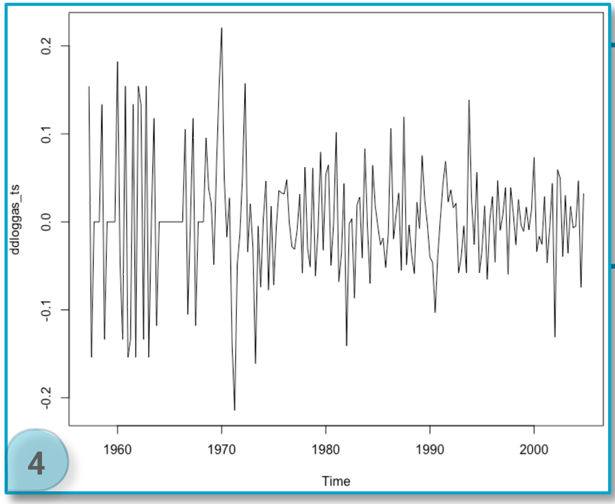


log-transformed electricty production

log-transformed gas production



Seasonal pattern similar to elec_ts



log-transformed gas_ts in both differences and seasonal differences: « ddloggas »

```
ADF test

data:  ddloggas_ts
ADF(3) = -10.194, p-value = 3.698e-16
alternative hypothesis: true delta is less than 0
sample estimates:
   delta
-1.32107
```

stationary (H0 rejected)

Multivariate analysis

Test for cointegration and model construction

Engle-Granger ADF test

```
        ADF test

data:  res_fit_ci
ADF(8) = -2.7649, p-value = 0.06543
alternative hypothesis: true delta is less than 0
sample estimates:
     delta
-0.0415203
```

- The test statistic is bigger that the Engle-Granger ADF test statitic for one explanatory variable (-3.41). Therefore, we do not reject H0 of no cointegration and conclude that log(elec) and log(gas) are not cointegrated.

Construction of a VAR model

- The Schwarz information criterion (SC) has been chosen to determine the order of the VAR model :

```
AIC(n)  HQ(n)  SC(n) FPE(n)
     8      4      4      8
```

- A VAR(4) model is constructed.

```
Estimation results for equation ddlogelec_ts:
==============================================
ddlogelec_ts = ddlogelec_ts.l1 + ddloggas_ts.l1 + ddlogelec_ts.l2
+ ddloggas_ts.l2 + ddlogelec_ts.l3 + ddloggas_ts.l3 + ddlogelec_t
s.l4 + ddloggas_ts.l4 + const

                  Estimate Std. Error t value Pr(>|t|)
ddlogelec_ts.l1 -0.2377110  0.0713873  -3.330  0.00106 **
ddloggas_ts.l1   0.0083911  0.0224175   0.374  0.70862
ddlogelec_ts.l2 -0.1185441  0.0748169  -1.584  0.11486
ddloggas_ts.l2   0.0213353  0.0220703   0.967  0.33501
ddlogelec_ts.l3 -0.0339713  0.0749420  -0.453  0.65088
ddloggas_ts.l3   0.0202129  0.0219537   0.921  0.35845
ddlogelec_ts.l4 -0.4137658  0.0725315  -5.705 4.78e-08 ***
ddloggas_ts.l4  -0.0022261  0.0216647  -0.103  0.91828
const           -0.0005436  0.0014076  -0.386  0.69983
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.01924 on 178 degrees of freedom
Multiple R-Squared: 0.2401,     Adjusted R-squared: 0.2059
F-statistic: 7.029 on 8 and 178 DF,  p-value: 4.736e-08
```

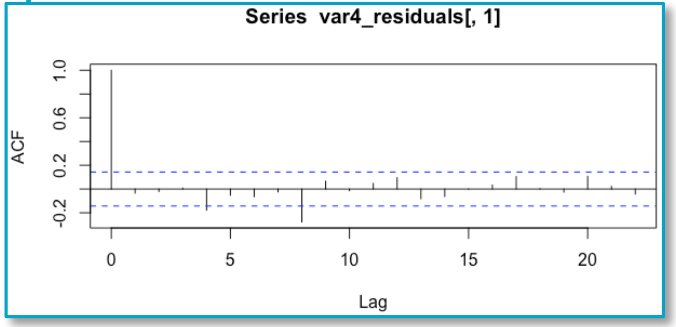
```
Estimation results for equation ddloggas_ts:
=============================================
ddloggas_ts = ddlogelec_ts.l1 + ddloggas_ts.l1 + ddlogelec_ts.l2
+ ddloggas_ts.l2 + ddlogelec_ts.l3 + ddloggas_ts.l3 + ddlogelec_t
s.l4 + ddloggas_ts.l4 + const

                  Estimate Std. Error t value Pr(>|t|)
ddlogelec_ts.l1  0.2047620  0.2148933   0.953   0.3420
ddloggas_ts.l1  -0.0585517  0.0674821  -0.868   0.3867
ddlogelec_ts.l2 -0.1020423  0.2252172  -0.453   0.6510
ddloggas_ts.l2  -0.0183443  0.0664370  -0.276   0.7828
ddlogelec_ts.l3  0.1280909  0.2255937   0.568   0.5709
ddloggas_ts.l3   0.1522764  0.0660861   2.304   0.0224 *
ddlogelec_ts.l4  0.0953905  0.2183374   0.437   0.6627
ddloggas_ts.l4  -0.4875379  0.0652161  -7.476 3.35e-12 ***
const            0.0003005  0.0042371   0.071   0.9435
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.05791 on 178 degrees of freedom
Multiple R-Squared: 0.3372,     Adjusted R-squared: 0.3074
F-statistic: 11.32 on 8 and 178 DF,  p-value: 6.531e-13
```

- 24% (33,7%) of the variance of ddlogelec (ddloggas) is explained by the lag observations of ddloggas and ddlogelec in the first four lags.
- The regressors are jointly significant (rejection of F-statistic H0 in both cases).
- Only a few regressors are significant, the others are not (see the stars).

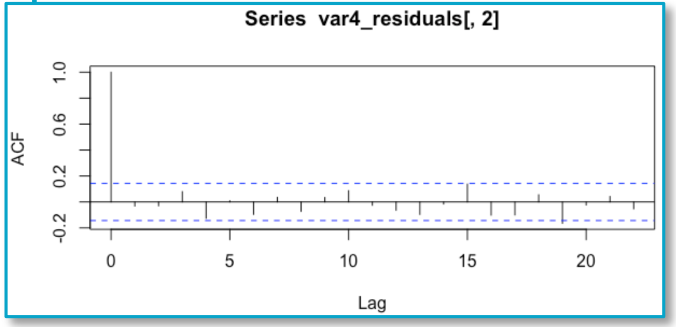
Multivariate analysis

Validation of the VAR(4) model

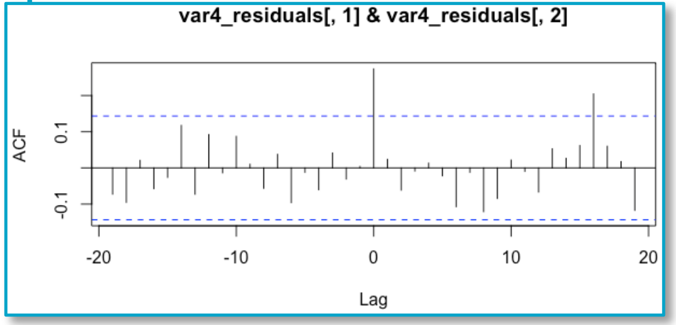
- ACF for ddlogelec residuals



- ACF for ddloggas residuals



- Cross-correlogram



- Since there is little or no significant autocorrelation, the residuals look multivariate white noise and the model is thus valid.

Conclusion

This analysis helped us to get more insights about electricity production in Australia. The quantity of electricity produced is in constant augmentation and we know there is a peak in production which is repeated each year in Q3. The forecast of our univariate model predict this same trend and seasonality. However, we are aware that other variables may influence this series. This is what we tried to find out by analyzing the gas production and creating a multivariate model. One area of improvement would be to explore other variables that may influence electricity production in Australia such as weather, etc.