Analysis of spanish olive oil price

Università degli Studi di Padova

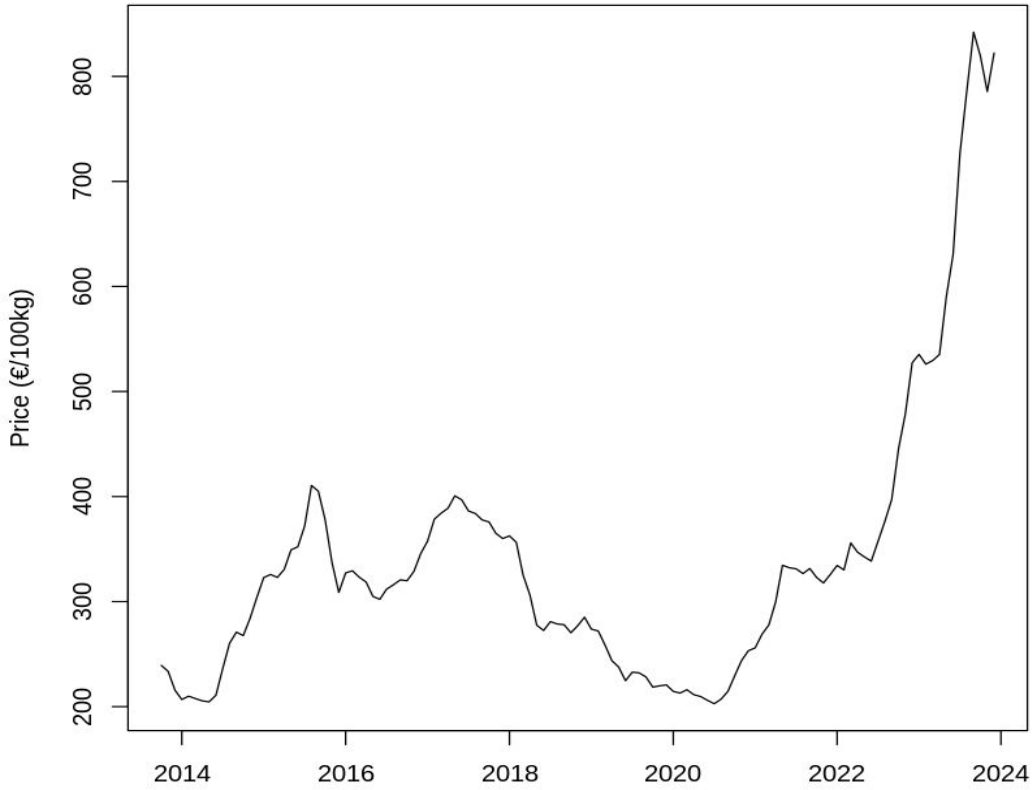
Mojsovska Marija
Orlando Filippo

Outlines

1. **Exploratory analysis**
2. **Linear model**
3. **GAM**
4. **ARIMA**
5. **Gradient boosting**
6. **Result and conclusion**

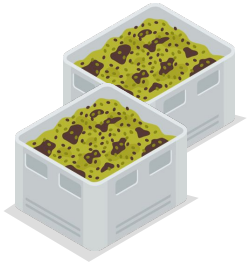
1.

Exploratory analysis



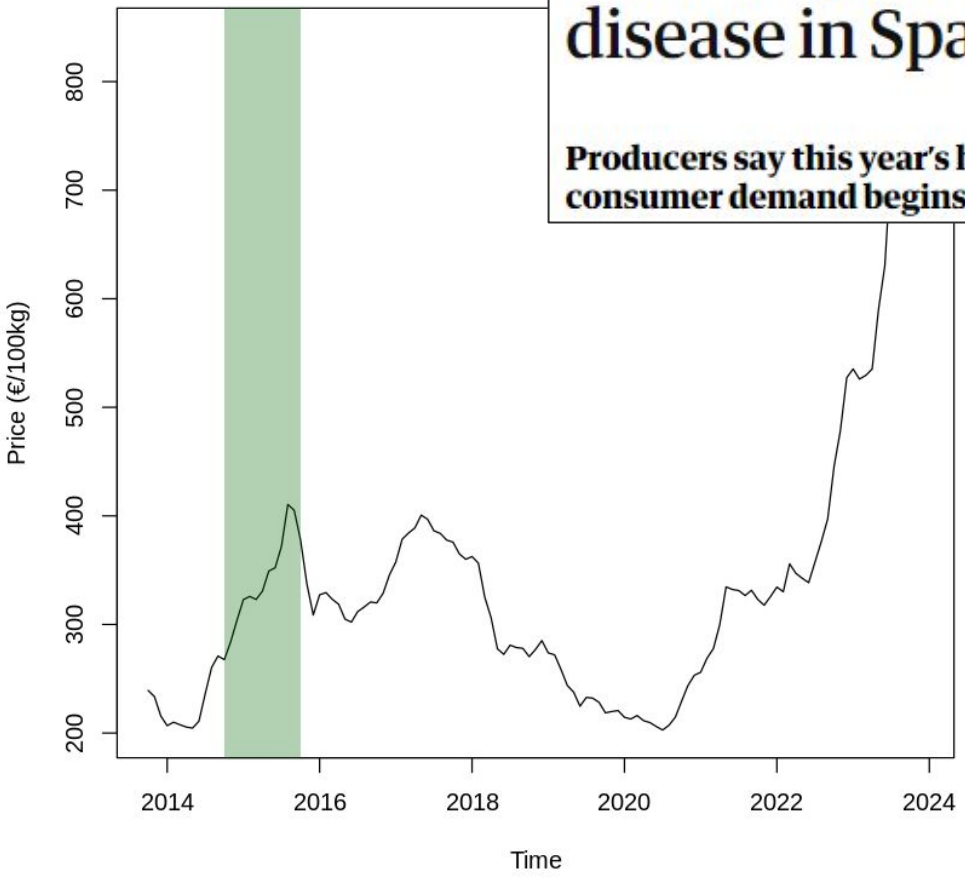
European Commission

Data from
European Commission
Agriculture and rural development

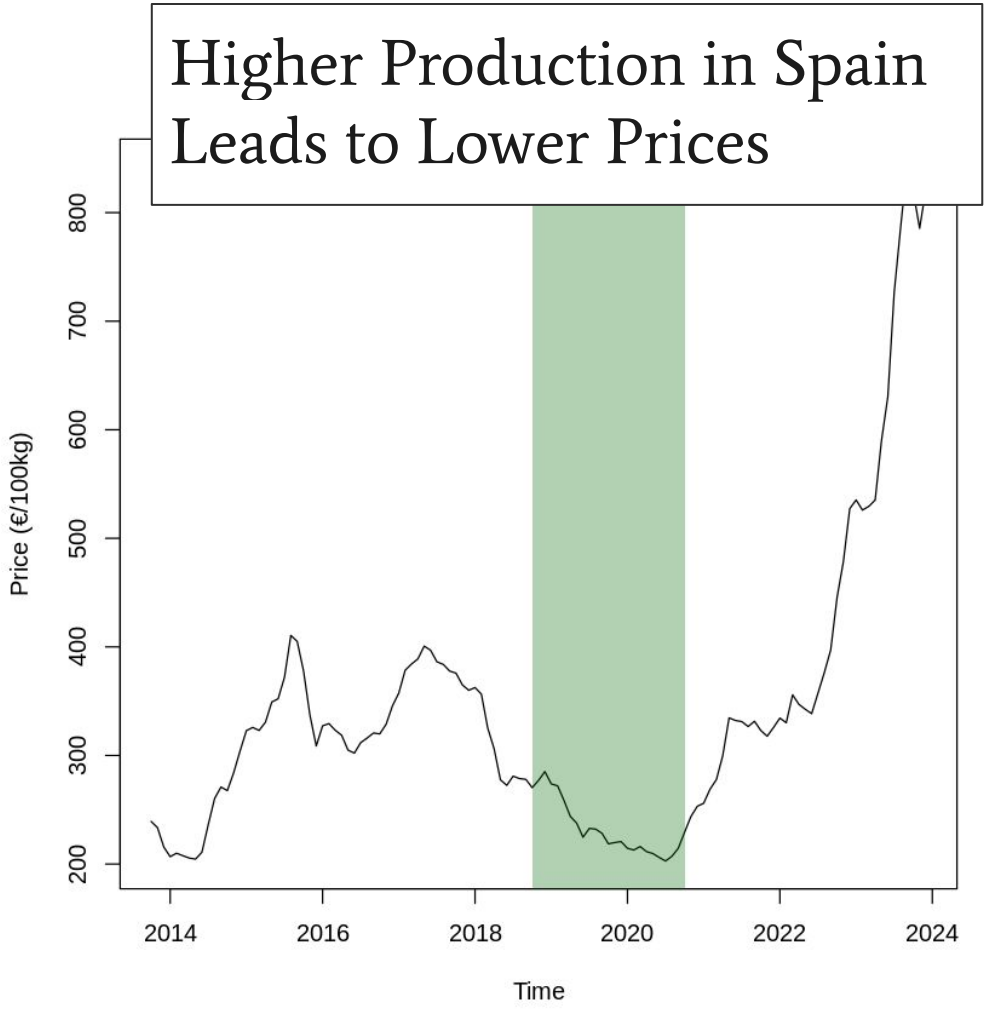


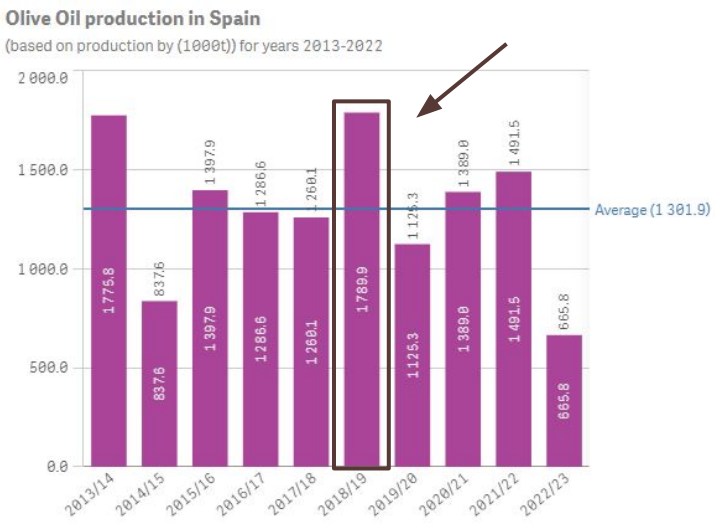
Olive oil prices surge due to drought and disease in Spain and Italy

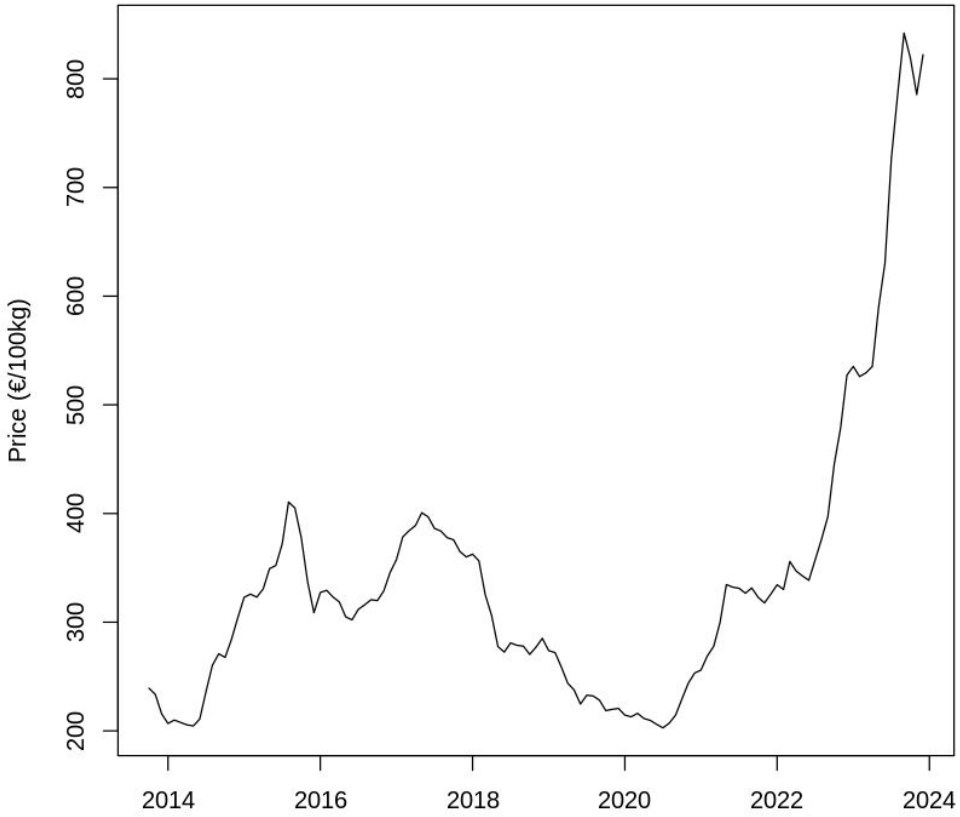
Producers say this year's harvests are worst they have seen, as consumer demand begins to outstrip supply



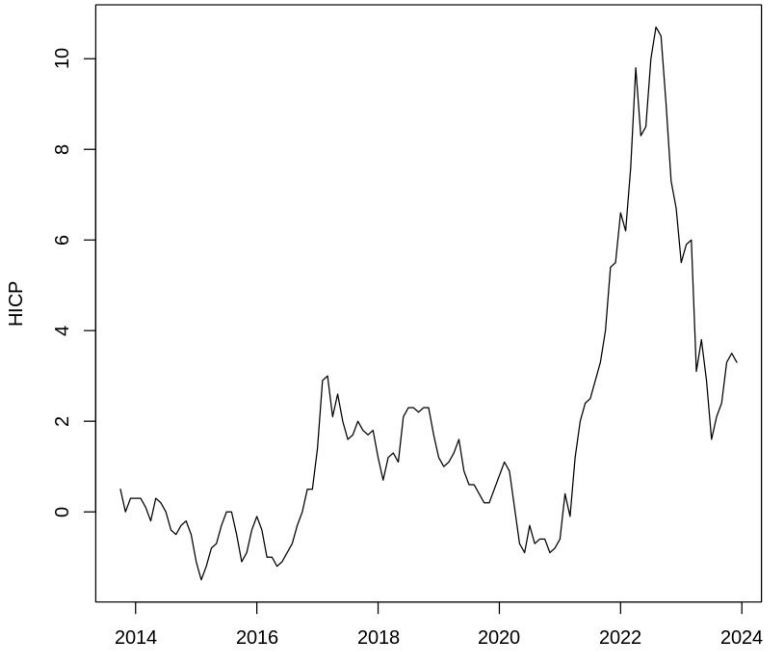
Higher Production in Spain Leads to Lower Prices



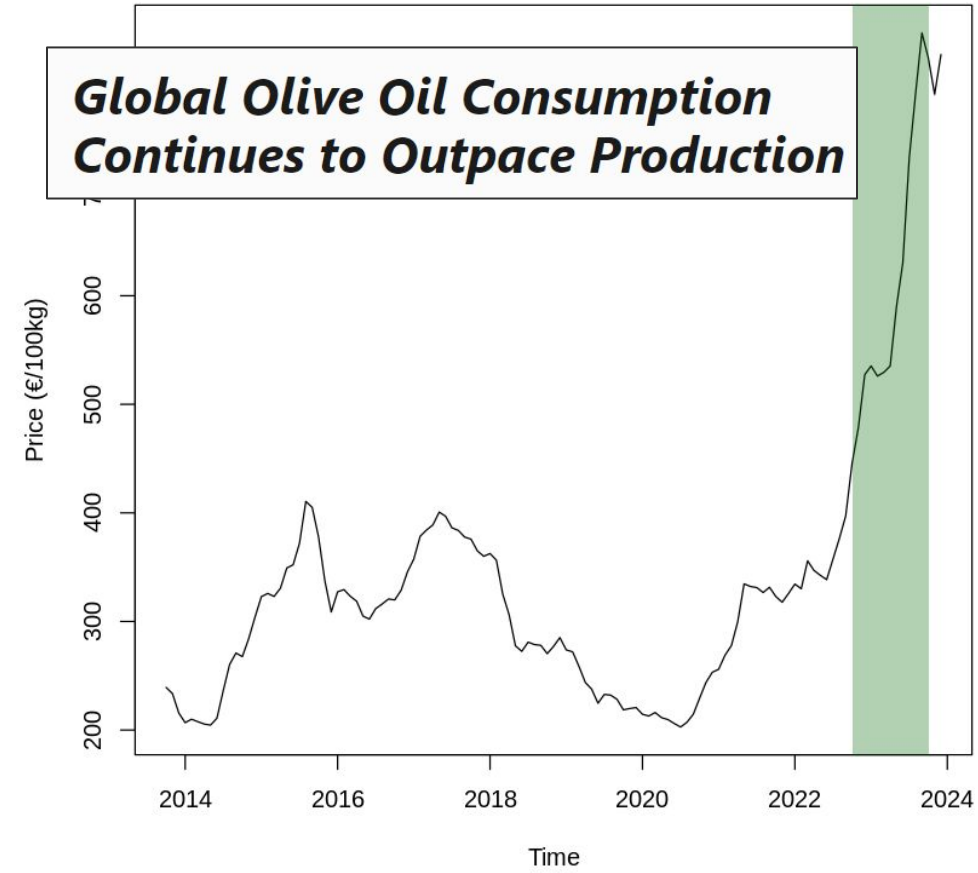


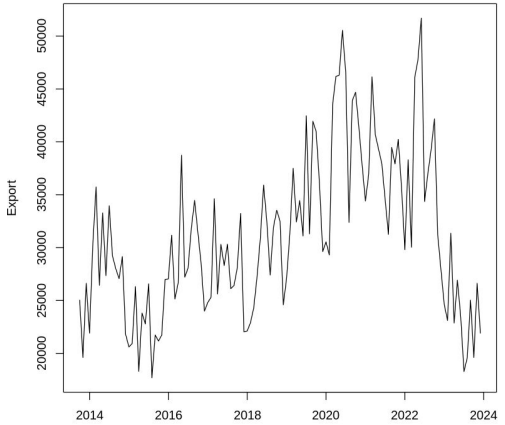


Inflation rate

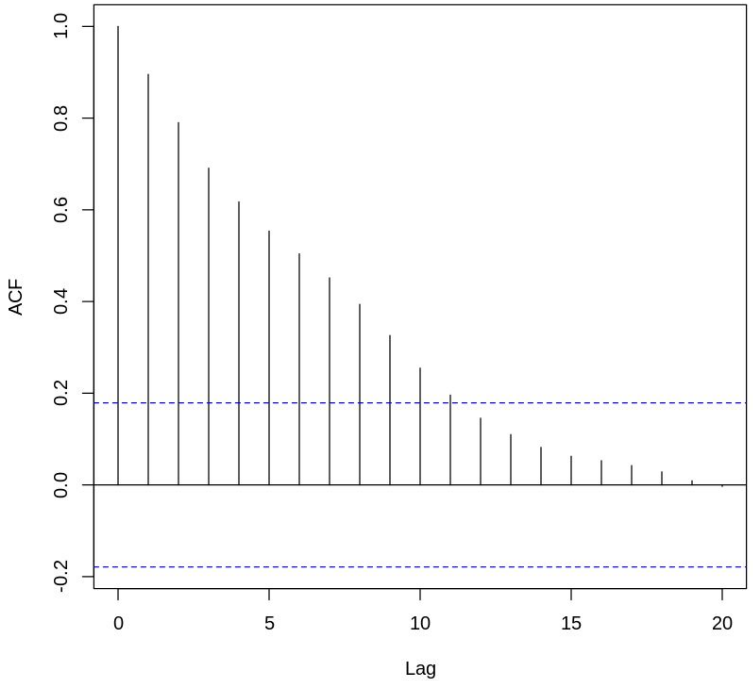


Drought and extreme heatwaves have halved Spanish olive oil production.
The price at origin has increased by 112 per cent since last year.





Time series analysis



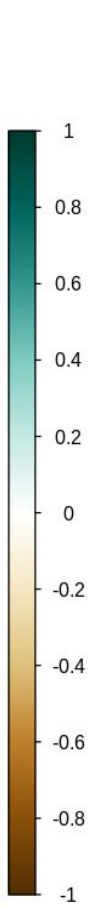
- Trend
- No evident seasonality

2.

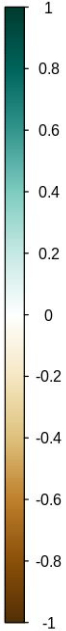
LINEAR MODEL

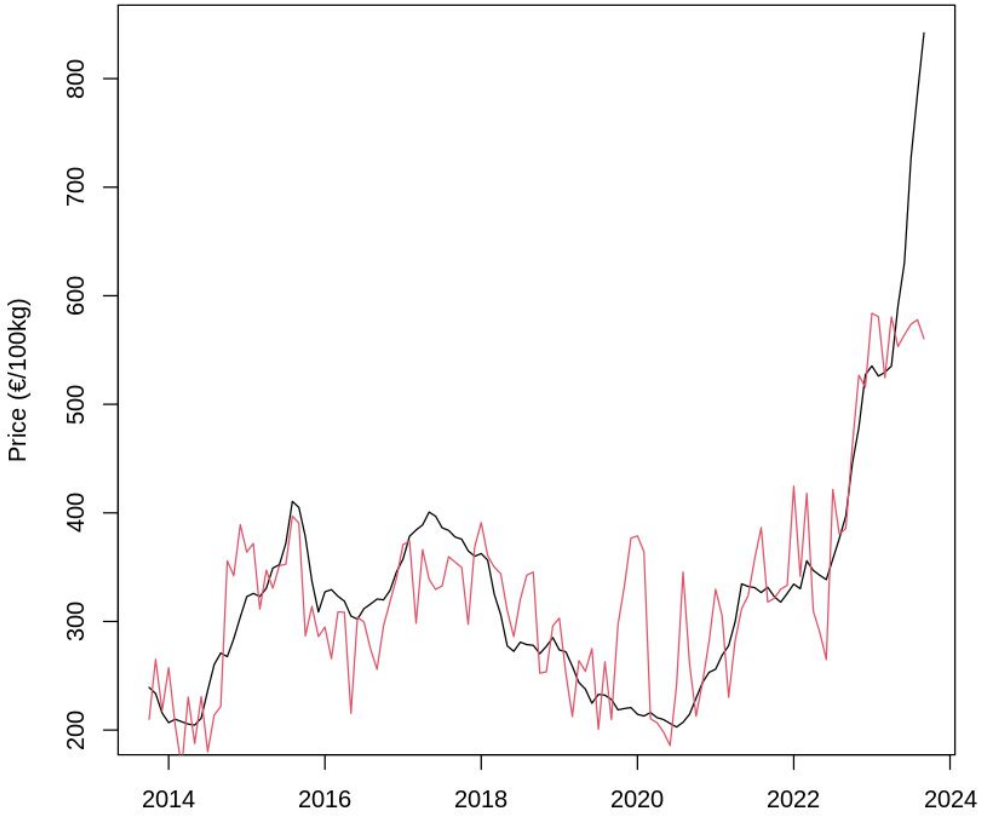
CORRELATION AND COLLINEARITY

	Price	Inflation	Production	Stock	Export	Imports	Urea	Potassium	DAP
Price	1.00	0.39	-0.63	-0.52	-0.38		0.26	0.26	0.42
Inflation	0.39	1.00			0.25		0.83	0.79	0.86
Production	-0.63		1.00	0.74	0.27	-0.33			
Stock	-0.52		0.74	1.00	0.44				
Export	-0.38	0.25	0.27	0.44	1.00	0.25	0.29	0.22	0.22
Imports			-0.33		0.25	1.00			
Urea	0.26	0.83			0.29		1.00	0.77	0.92
Potassium	0.26	0.79			0.22		0.77	1.00	0.80
DAP	0.42	0.86			0.22		0.92	0.80	1.00

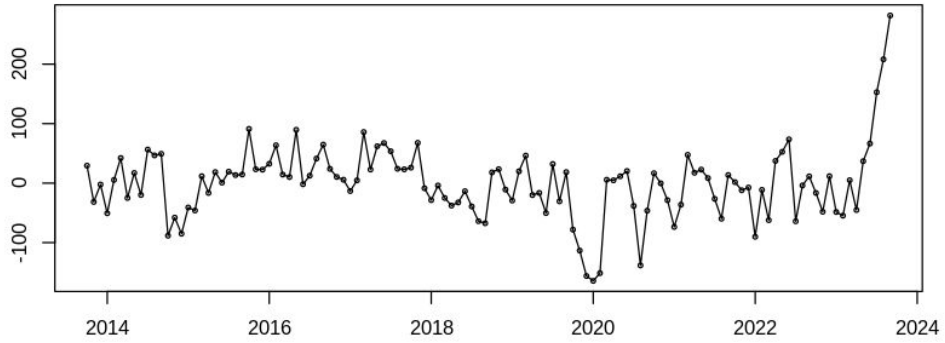


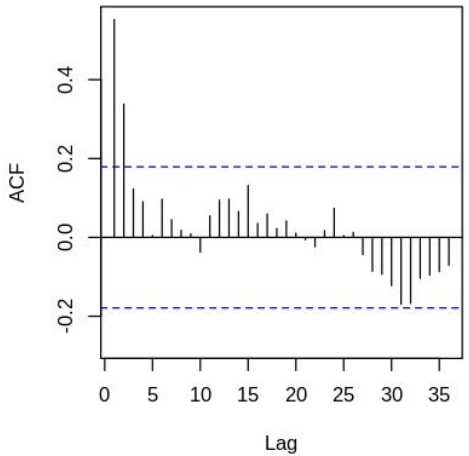
	Price	Inflation	Production	Export	Imports
Price	1.00	0.39	-0.63	-0.38	
Inflation	0.39	1.00		0.25	
Production	-0.63		1.00	0.27	-0.33
Export	-0.38	0.25	0.27	1.00	0.25
Imports			-0.33	0.25	1.00

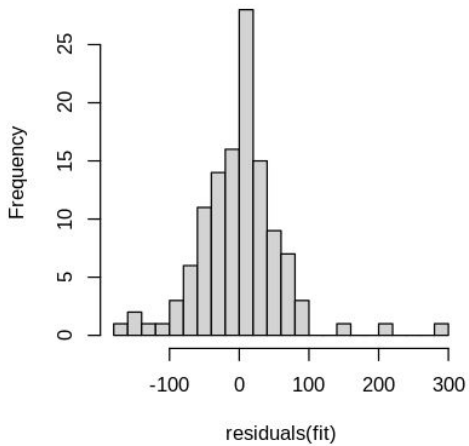





```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.500e+02  3.152e+01  20.620  < 2e-16 ***
production  -1.344e-01  2.057e-02  -6.537 1.84e-09 ***
inflation    6.537e+00  2.841e+00   2.301  0.02322 *
export      -7.162e-03  9.585e-04  -7.472 1.73e-11 ***
import      -4.454e-03  1.493e-03  -2.983  0.00349 **
trend        1.587e+00  2.708e-01   5.860 4.58e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
’ 1

Residual standard error: 62.03 on 114 degrees of freedom
Multiple R-squared:  0.7115,
Adjusted R-squared:  0.6989
F-statistic: 56.23 on 5 and 114 DF,  p-value: < 2.2e-16

AIC =  1339.028
```





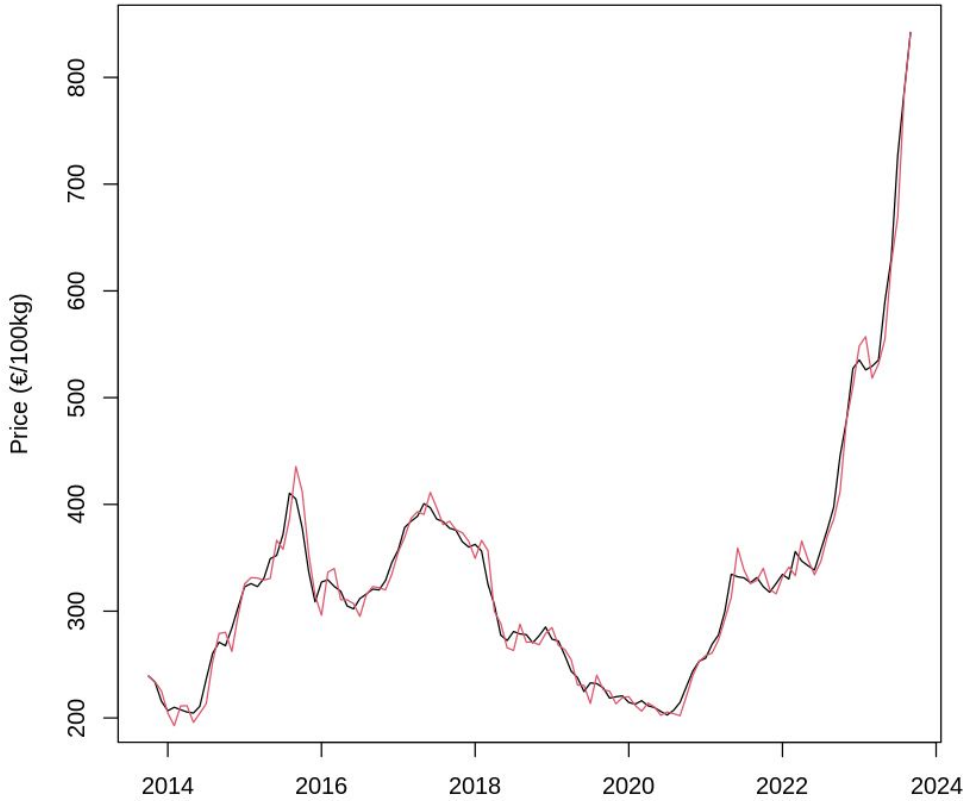


Durbin-Watson test

DW = 0.71294, p-value = 2.773e-14
alternative hypothesis: true
autocorrelation is greater than 0

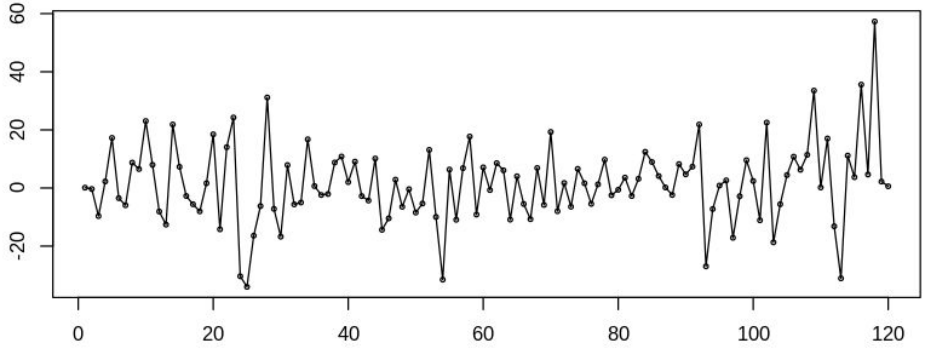
Can we improve the previous model?

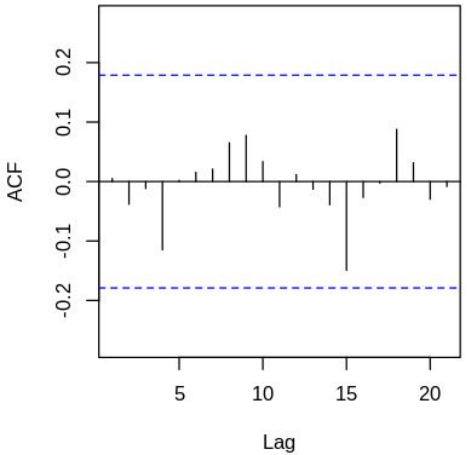
Dynamic regression model with ARIMA errors

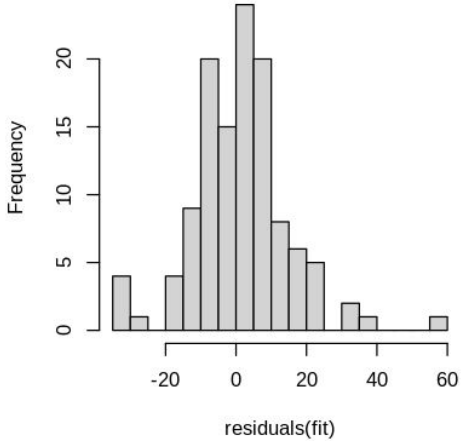



```
Regression with ARIMA(0,2,3) errors

Coefficients:
          ma1     ma2      ma3  production  inflation  export  import
      -0.4263  0.0203  -0.3634     -0.0033     1.7078  -3e-04   4e-04
s.e.   0.0878  0.1100   0.0966      0.0071     1.5557   2e-04   3e-04
Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE
Training set 1.456792 13.55568 9.902572 0.4133479 3.021596 0.7701124
```

```
sigma^2 = 198.7:
log likelihood = -476.46
AIC=968.93   AICc=970.25
BIC=991.09
```







Ljung-Box test

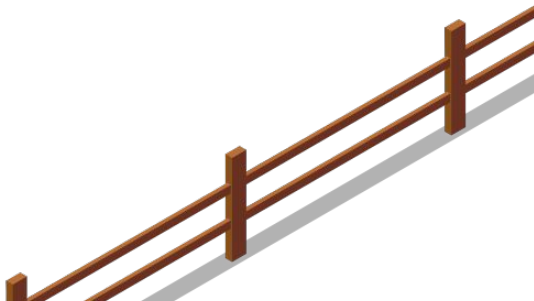
data: Residuals from Regression with ARIMA(0,2,3) errors
Q* = 3.4469, df = 7, p-value = 0.8408

Model df: 3. Total lags used: 10

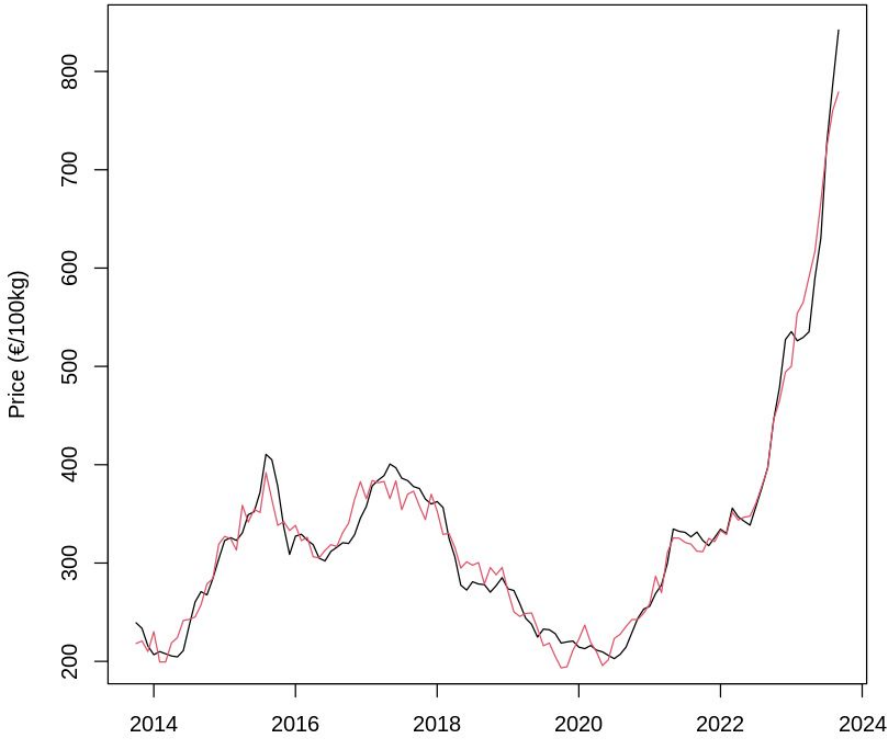
3.

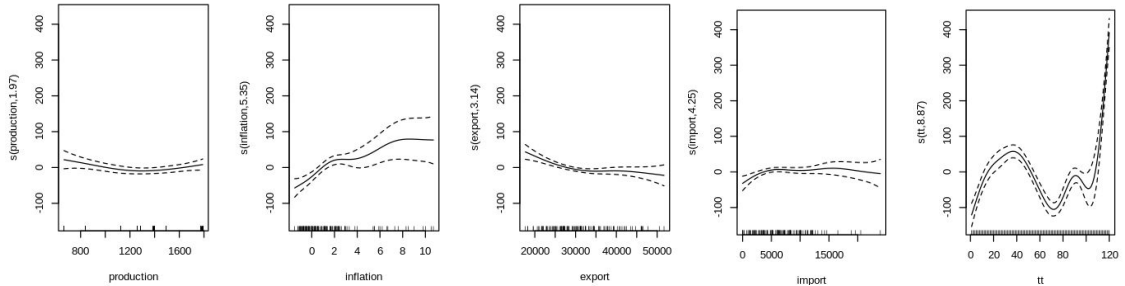
GAM

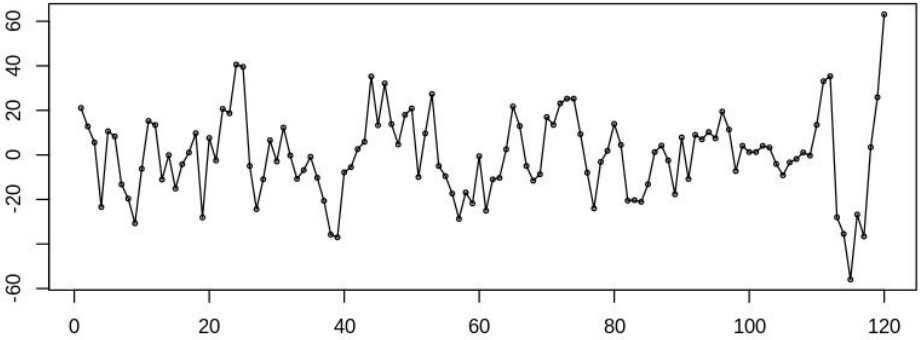
model	AIC
price ~ lo(production) + lo(inflation) + lo(export) + lo(import) + lo(tt)	1216.74
price ~ production + s(inflation) + export + import + s(tt)	1121.72
price ~ s(production) + s(inflation) + export + import + s(tt)	1109.29
price ~ s(production) + s(inflation) + s(export) + s(import) + s(tt)	1094.02
price ~ production + s(inflation) + s(export) + s(import) + s(tt)	1100.01

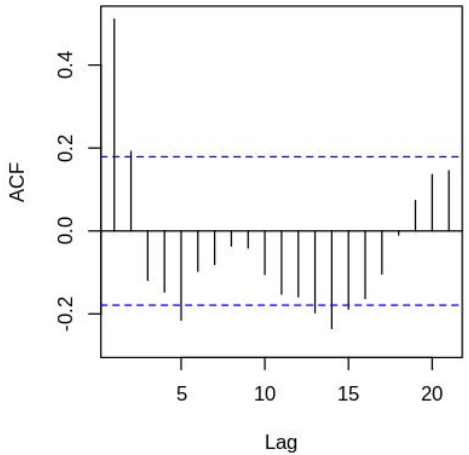


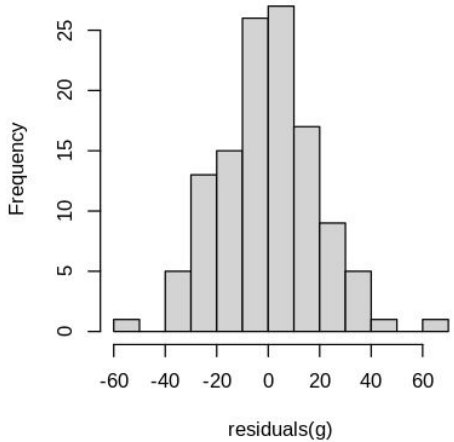
```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   327.41       1.91   171.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Approximate significance of smooth terms:
                edf Ref.df      F  p-value
s(production) 1.967  2.308  2.629 0.072504 .
s(inflation)  5.348  6.483  4.851 0.000171 ***
s(export)     3.142  3.963  5.632 0.000480 ***
s(import)     4.248  5.265  2.211 0.057389 .
s(tt)         8.871  8.983 69.377  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

R-sq.(adj) =  0.966
Deviance explained = 97.3%
AIC =  1094.02
```











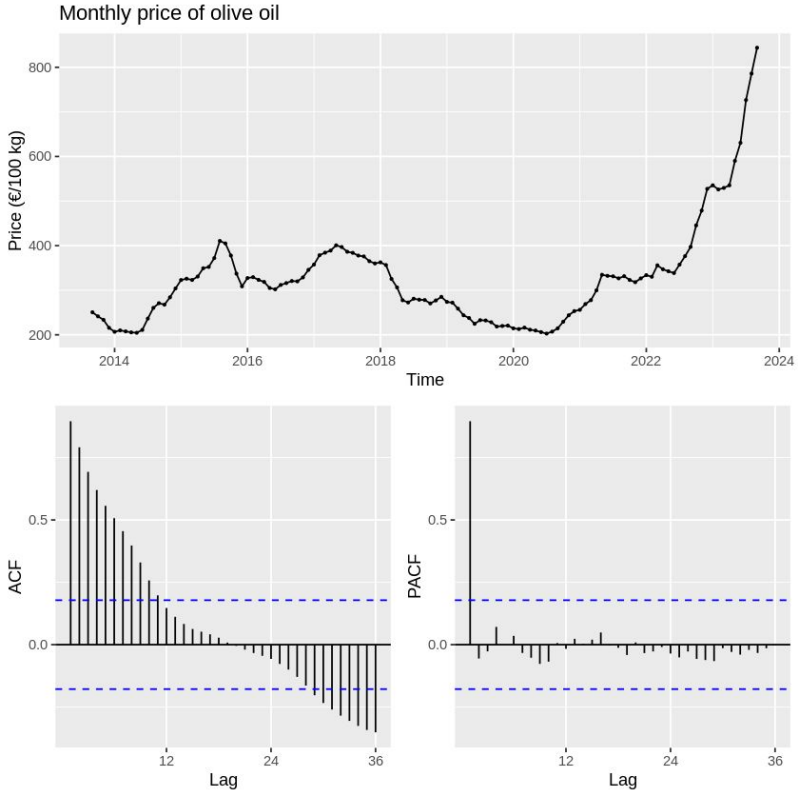
Durbin-Watson test

DW = 0.71294, p-value = 2.773e-14
alternative hypothesis: true
autocorrelation is greater than 0

4.

ARIMA

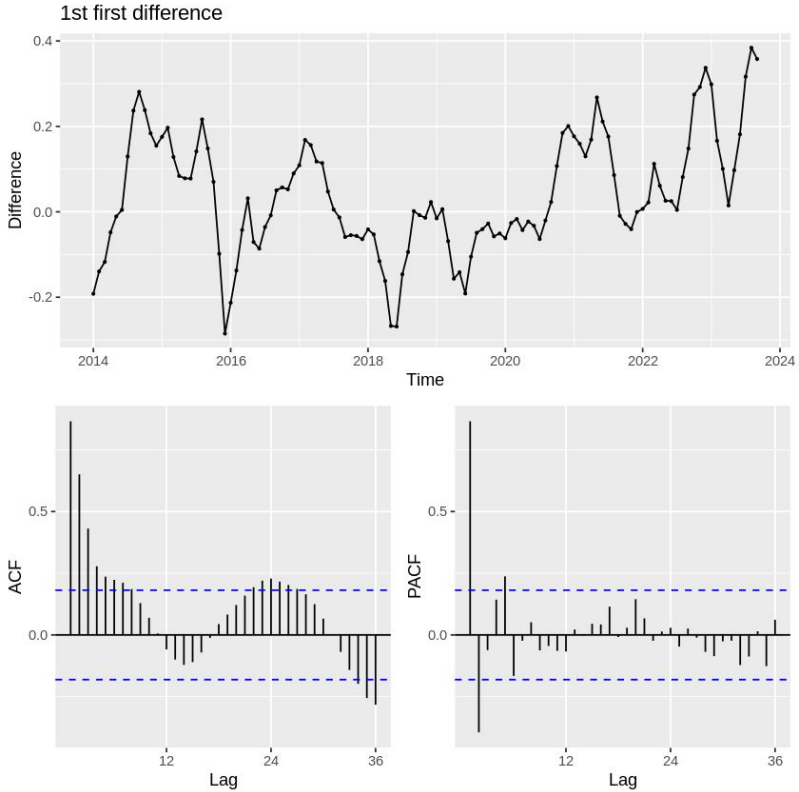
ARIMA - First analysis



Looking at the time-serie:

- It is non - stationary
- Need to difference

ARIMA - Differencing

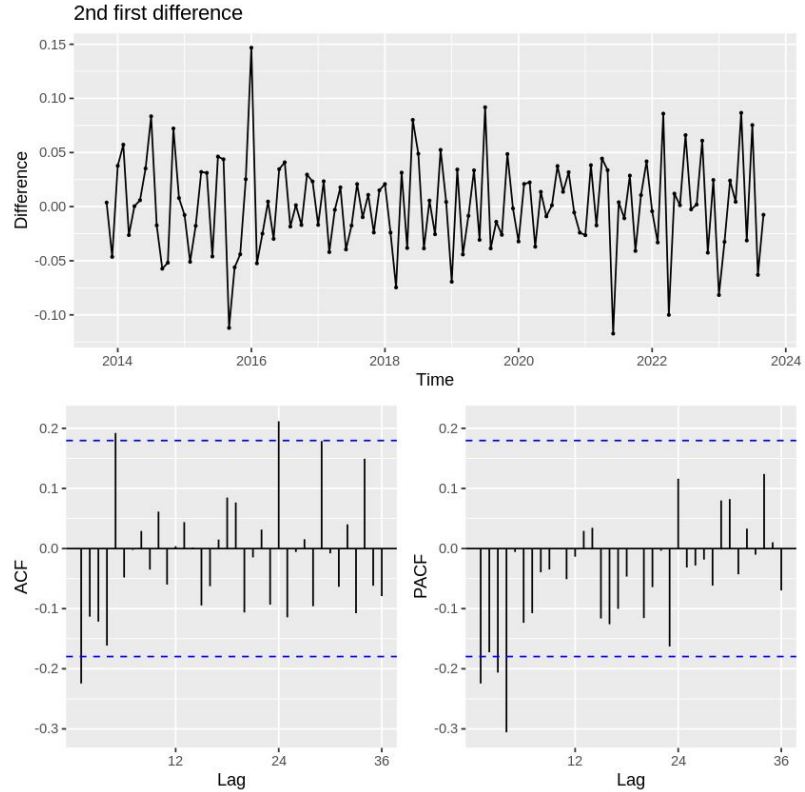


Differencing:

- Try to eliminate the trend
- First difference
- Need to difference: there is still a trend

```
        Ljung-Box test

data:  Residuals
Q* = 225.21, df = 23, p-value < 2.2e-16
```

ARIMA - Differencing



Differencing:

- Another first difference
- Trend is eliminated
- Infer in p and q values for AR and MA
- p : 0, 1, 2, 3, 4
- q : 1, 2, 3, 5

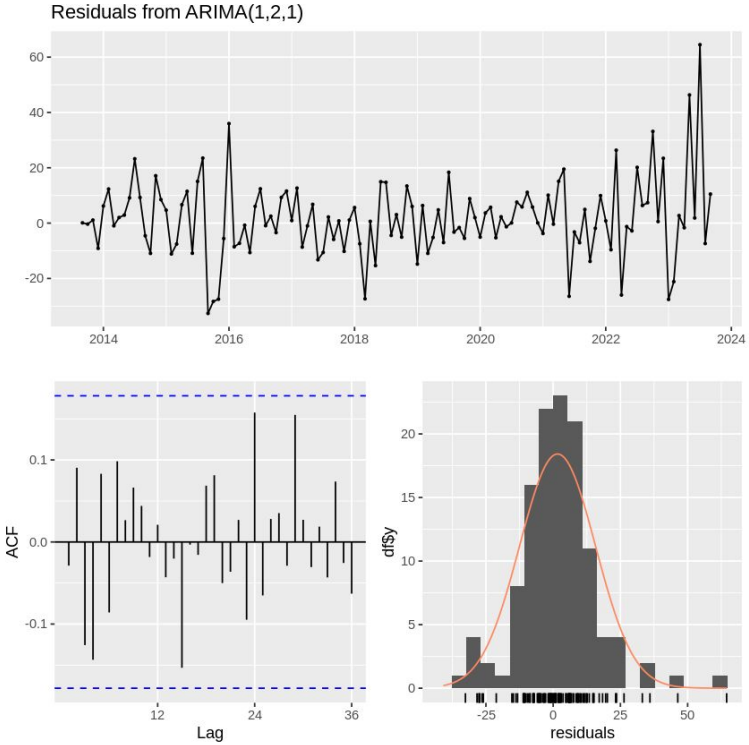
```
        Ljung-Box test

data:  Residuals
Q* = 32.916, df = 24, p-value = 0.1058
```

ARIMA "Manual" Results

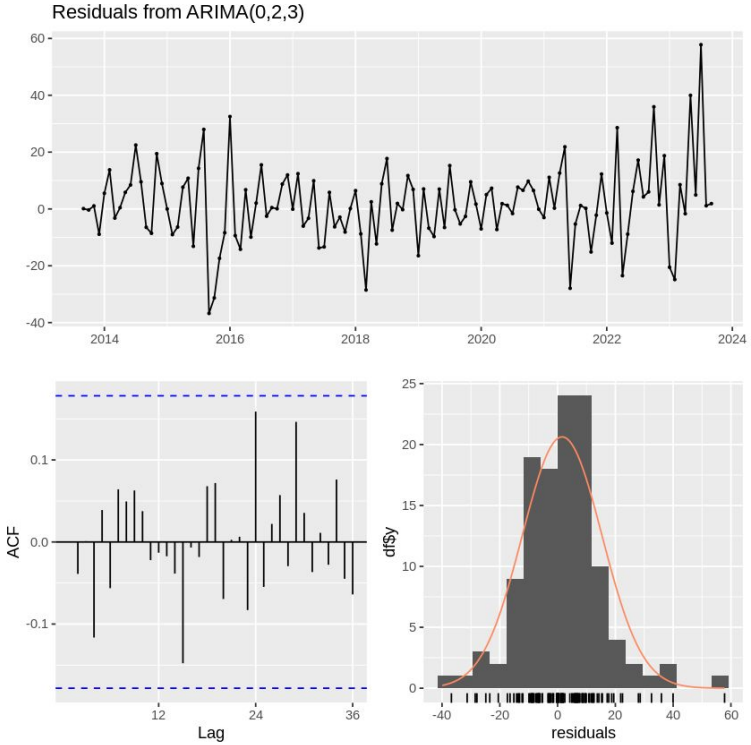
Top 4	AIC
ARIMA(0, 2, 3)	**974,32**
ARIMA(3, 2, 1)	976,36
ARIMA(3, 2, 2)	976,41
ARIMA(4, 2, 1)	977,03

Other	AIC
ARIMA(1, 2, 1)	977,39
ARIMA(1, 2, 5)	977,5
ARIMA(2, 2, 1)	979,29
ARIMA(0, 2, 2)	982,2

ARIMA "AUTO" RESULTS



AIC = 977,39 RMSE = 14.17 €/100 kg



Search on all possible models AIC = 974,32
RMSE = 13.86 €/100 kg

5.

Gradient Boosting

Parameters - Grid Search

Parameter	Value
n.trees	2000
interaction_depth (max node per tree)	4
Shrinkage (learning rate)	0.1

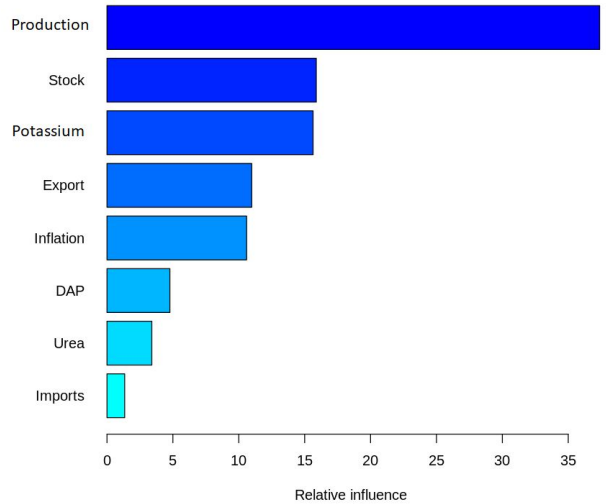
Parameters:

- Using Grid Search we can make a fine selection of parameters based on the accuracy of the model
- Variation of these parameters may cause a different evaluation of relevance variables

Parameters and result

Variable	Relative Influence
Production	37.38
Stock	15.88
Potassium	15.63
Export	10.98
Inflation	10.59
DAP	4.77
Urea	3.40
Imports	1.34

Initially we considered all variables.



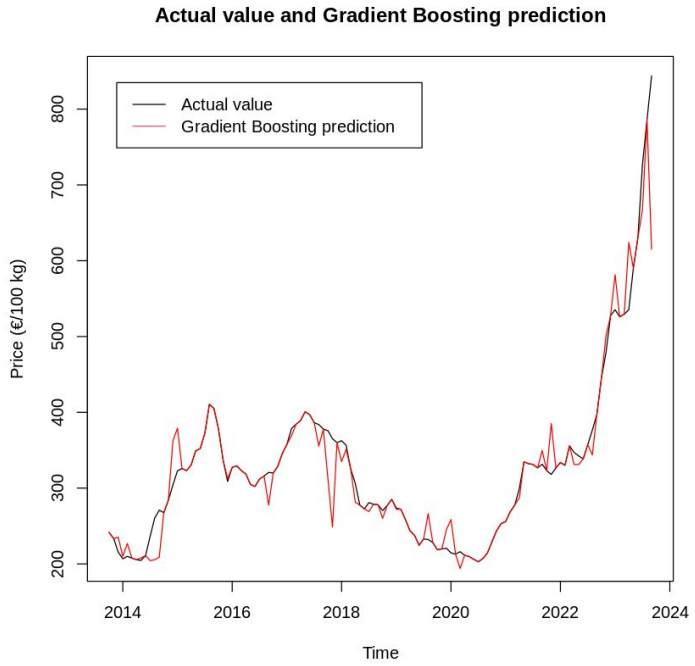
Parameters and result

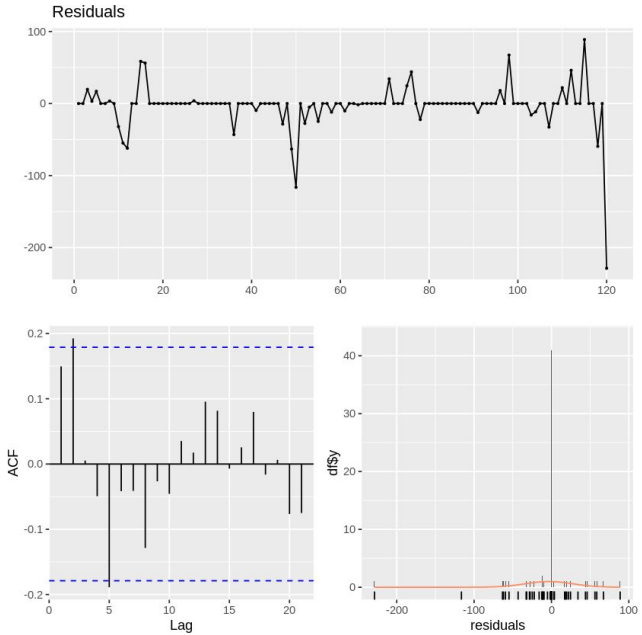
Variable	Relative Influence
Production	**37.38**
Stock	15.88
Potassium	15.63
Export	**10.98**
Inflation	**10.59**
DAP	4.77
Urea	3.40
Imports	**1.34**

After multicollinearity considerations

Variable	Relative Influence
Production	54.28
Inflation	20.6
Export	18.69
Imports	6.42

FIT





RMSE = 30.84 €/100 kg

6.

RESULT AND CONCLUSION

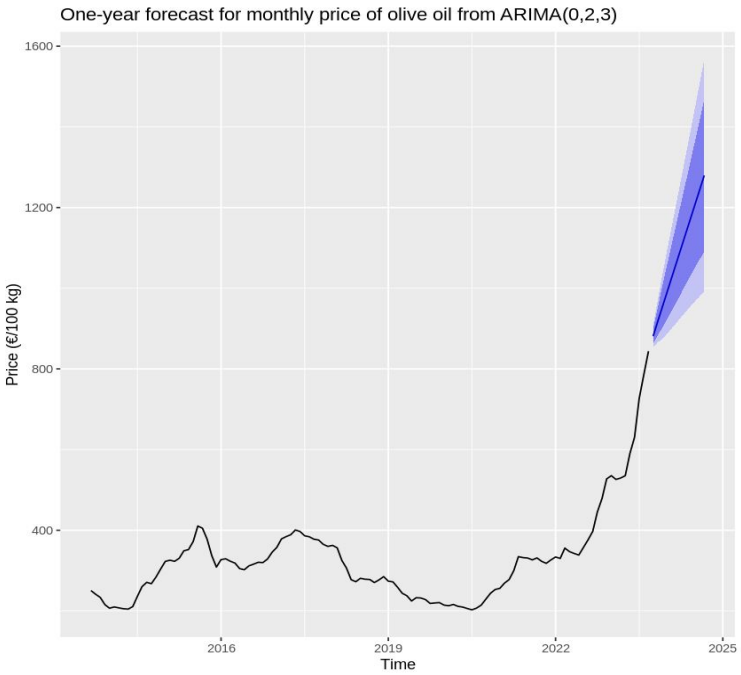
Results

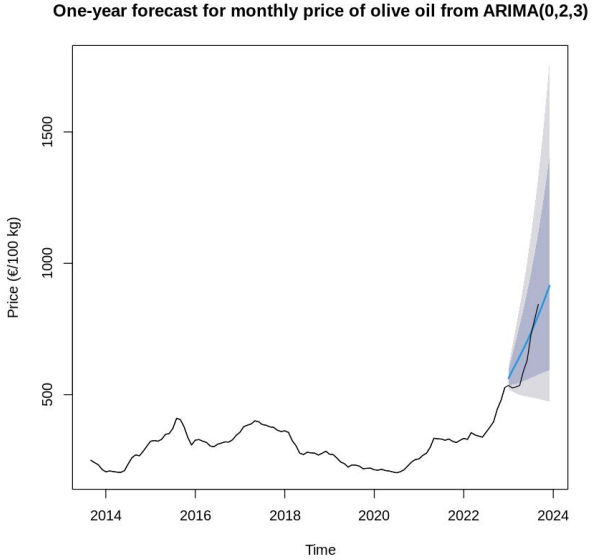
Model	AIC
ARIMA (0, 2, 3)	974,32
Regression + ARIMA errors (1, 1, 0)	983.04
GAM	1094.02
Linear Model	1339.03

```
ARIMA(0,2,3)

Coefficients:
          ma1      ma2      ma3
      -0.4114  -0.0213  -0.3442
s.e.   0.0848   0.1000   0.0974

sigma^2 = 200.5:  log likelihood = -483.16
AIC=974.33   AICc=974.68   BIC=985.44
```

```
        Ljung-Box test

data:  Residuals from ARIMA(0,2,3)
Q* = 14.607, df = 21, p-value = 0.8421
```

The best model - ARIMA(0, 2, 3)





	MAPE
Train	2.96 %
Test	2.65 %

Conclusion

- The best model in term of AIC is the **ARIMA(0, 2, 3)**
- The variables that are significant: **production**, **inflation**, **export** and **import**
- **Future work:** include production cost and consumption

Thank you for your attention