Estimation and model selection in the case of time series. Application on natural gas production

Victor Platon[1],

Andreea Constantinescu[2],

Sorina Jurist[3]

Abstract: *In 2021, natural gas price, extraction and distribution across EU became a key theme. The evolution of prices and quantities delivered are in the middle of a heated debate. On the one hand there is a clear descending trend and, on the other hand, demand is up, pushing prices upward. In the last 10 years was noticed a steady decline in natural gas extracted both at EU level and in Romania. This tendency is typical for the new paradigm concerning energy policy oriented towards a cleaner power output. In these circumstances of prices volatility, it is useful to elaborate an consolidative model for natural gas production in Romania. One important advantage of this model would be to have a tool for better estimate of Romania's energy output. The methodology used several econometric models adapted to non-stationar time series suitable for natural gas production. By comparing the models, we end up selecting the most significant model, the one that showed improved forecast statistics. In this paper were envisaged several results: analysis of the volume of natural gas production; developing regression models for natural gas production in Romania; conclusions on most suitable econometric model for natural gas production in Romania.*

Keywords: *Econometric modelling, time series, linear regression, autocorrelation of errors, AIC, Durbin – Watson statistic, natural gas production*

JEL Classification: *Q32, Q35, C30*

[1] PhD., CS I, Institute of National Economy, victor.platon54@gmail.com

[2] PhD., CS III, Institute of National Economy, andreea_constantinescu07@yahoo.com

[3] As. Institute of National Economy, sorinaj2005@yahoo.com

INTRODUCTION

The goal of the article is finding the best econometric model for specific time series and proposing a method for model selection. The model estimation and selection will be applied to a time series representing natural gas production in Romania in the period 1960-2016. The series considered is a time series, expressed in physical units.

Natural gas is a fuel with a high energy value, with the lowest emission level after combustion, compared to other fossil fuels. Methane gas, together with oil and coal, represent the most significant natural resources in Romania.

The evolution of natural gas production, the trends registered during the last 20 years in EU 28 and in Romania are considered. These developments are seen in conjunction with the main trends in EU Member States and global trends. The topic is part of the efforts to optimize the production model of fossil fuel resources in Romania. To analyse the way natural gas production evolved in Romania, statistical data available in the Eurostat database and data from the Statistical Yearbook of Romania were used. This way, we have constructed a continuous series for the period between 1960 and 2017.

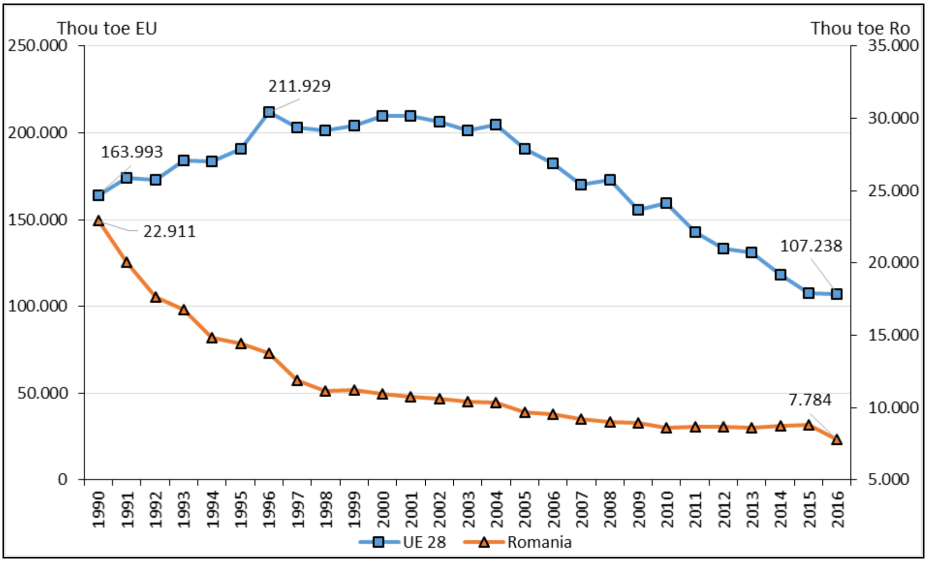
Evolution of natural gas production in Romania and EU

Prospecting activity decreased in intensity since 1990. The decline in Romanian economy, in gas industry, outsourcing companies specialized in natural gas drilling, decreased gas demand on the Romanian market due to disappearance of large industrial consumers, the high cost of drilling, the decrease of natural gas price due to the imported gas during 1997-1998, etc. determined a significant reduction of the drilling activity for natural gas.

The reduction of the prospection activity, of the drilling works, the natural depletion of the deposits and the technical accidents have led to the diminution of the number of active wells. From Figure 1 it can be observed that, at present, the quantity of gas extracted is less than one third of the value registered in 1990, respectively 7,784 thousand toe in 2016 compared to 22,911 thousand toe in 1990.

At European level, the trend of natural gas exploitations is also decreasing, with the difference that the maximum recorded, for the period analyzed in this paper, was in 1996, respectively 211,929 thousand toe. Currently, the value of natural gas exploitations is below half the maximum registered in 1996, respectively 107,238 thousand toe at the level of 2016.

Figure 1. Evolution of natural gas production in Romania and the EU (1990-2016)



Source: Own processing according to Eurostat data.

Methodology used in econometric modelling and selection of time series

In order to find out the best model for a time series, a methodology to help us in this attempt was developed. Thus, the methodology that was used in this article in order to perform econometric modelling and selection of time series was the following:

1. Visual analysis and interpretation; decide on plausible alternative model specifications;
2. Test for stationarity (unit root test for stationarity), account for possible structural breaks (unit root rest for structural break);
3. Estimation & Model Selection; proposing three regression equations to be analysed; estimating models using Ordinary Least Square (OLS) method and ARMA Maximum Likelihood method if that would be the case;
4. Choose “best” model, based on several indicators as: R-squared, significance of coefficients with less than 5% probability, the AIC (Akaike Info Criterion) and RMSE (root mean square error) indicators, white noise residuals (Durbin-Watson statistic -

DWS); the main goal using AIC is to minimize the estimated information loss and to differentiate between models.

5. Autocorrelation of errors would be eliminated by introducing autoregressive terms.

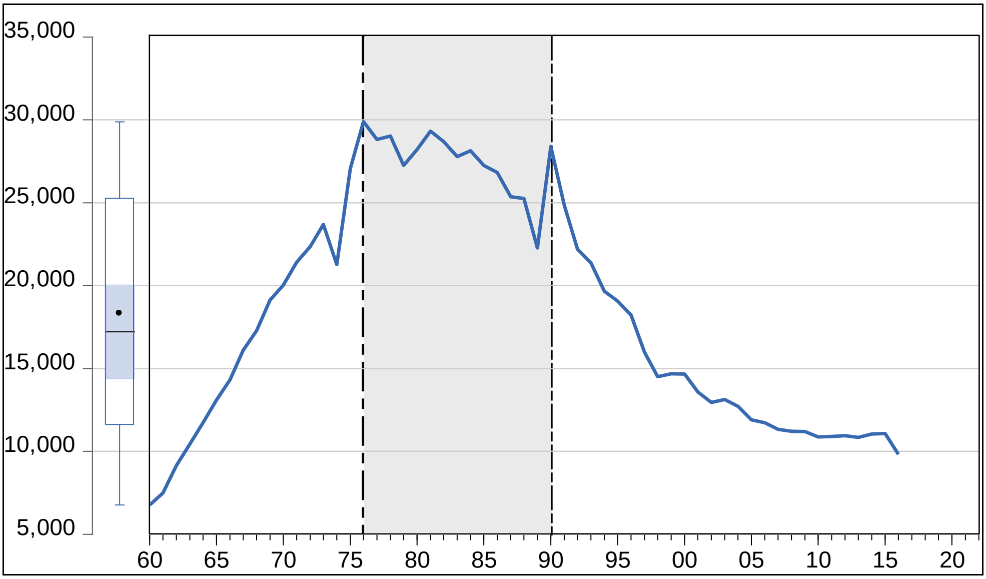
Econometric modelling of time series

In this section we are going to apply the methodology exposed to a time series: natural gas production from 1960 to 2016 (Annex 1).

Visual analysis of the time series (natural gas production)

Analyzing the evolution of the time series natural gas production in Romania (Figure 2) we can see a rapid increase, up to a maximum in 1976 (29,83 mil. m3), followed by a an almost constant plateau until 1990 (28,33 mil. m3). After 1990, natural gas production decreased up to 9.76 mil. m3 in 2016.

Figure 2. Natural Gas production in Romania (mil. m3)



Source: Own processing according to Eurostat data.

Due to the fact that the series has a shape with at least three trends we may conclude that the series is not stationary. In order to confirm this, we perform two tests: one for stationarity and another for structural break.

a) Stationarity: from the shape of the series, we may notice a trend that outline the fact that the time series is not a stationary one. Unit root test confirm this: Null Hypothesis: Nat.gas.prod has a unit root that cannot be rejected as Augmented Dickey-Fuller test statistic has a probability of 0.33 (Table 1), higher than the accepted limit of 5%.

Table 1. Unit root test for stationarity

Null Hypothesis: Nat.gas.prod has a unit root				
Exogenous: Constant				
Lag Length: 6 (Automatic - based on AIC, maxlag=10)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-1.899205	0.3301
Test critical values:	1% level		-3.568308	
	5% level		-2.921175	
	10% level		-2.598551	

*MacKinnon (1996) one-sided p-values

Source: Own processing according to Eurostat data.

b) Structural break: the shape of the series indicates that there are structural breaks. In order to determine these breaks, we perform unit root test for structural break (Augmented Dickey-Fuller (ADF) test statistic). In Table 2 we may notice that ADF has a probability of 0,256 so we cannot reject the Null Hypothesis that the time series has one structural break in 1990.

The year 1990 is the break point from when the natural gas production has started to decline until the present.

Table 2. Unit Root Test for structural break

Null Hypothesis: Nat.gas.prod has a unit root				
Trend Specification: Trend and intercept				
Break Specification: Intercept only				
Break Type: Innovational outlier				
Break Date: 1990				
Break Selection: Minimize Dickey-Fuller t-statistic				
Lag Length: 0 (Automatic - based on Schwarz information criterion, maxlag=10)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-4.204962	0.2565
Test critical values:	1% level		-5.347598	
	5% level		-4.859812	
	10% level		-4.607324	
*Vogelsang (1993) asymptotic one-sided p-values.				

Source: Own processing according to Eurostat data.

Model Estimation

After the first stage of our analysis we may conclude that the time series considered, in which the independent variable is time, is not stationary and has a structural break. We start modelling this time series by using a linear regression that describes a linear relationship between the forecast variable y (Nat.gas.prod expressed in mil. m3) and a single predictor variable x (time) plus the random effect:

$$yt = \beta0 + \beta1xt + \varepsilon t$$

This basic model will be fine-tuned by adding some dummy variable when needed.

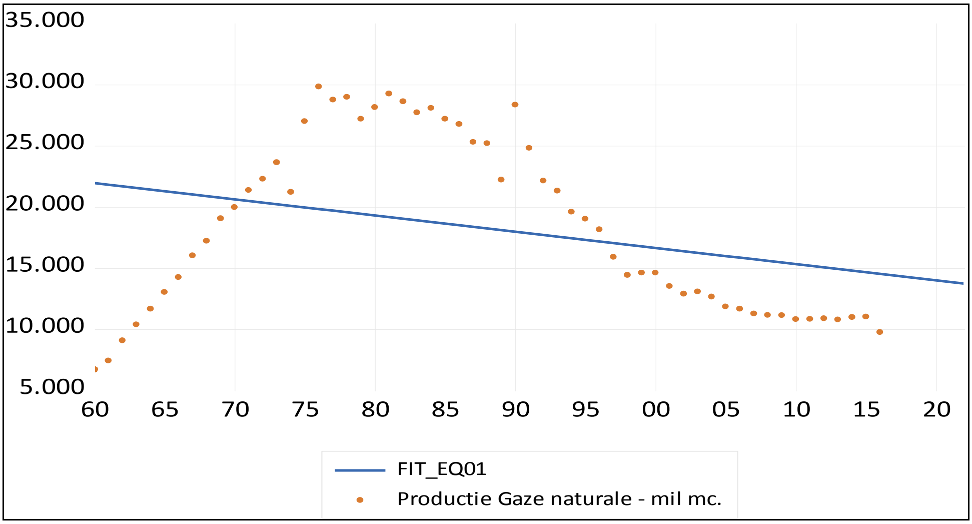
A. Model 1. Simple linear regression

The simplest regression model is a straight line that crosses the data as it is in the next figure. The basic equation describing this model is:

Eq1: Nat.gas.prod = C(1) + C(2)*TIMP

Analysing Figure 3 we can notice that a straight line does not perform a good job in modelling the Nat.gas.prod series. From Annex 2 we may notice that the coefficients are statistically significant but R-squared is low (0.096) showing a poor fit of the regression line. The value for AIC statistic is 20.52, while Durbin-Watson statistic has a low value (0.0652) showing a strong auto-correlation of residuals.

Figure 3. Linear regression for Model 1



Source: Own processing according to Eurostat data.

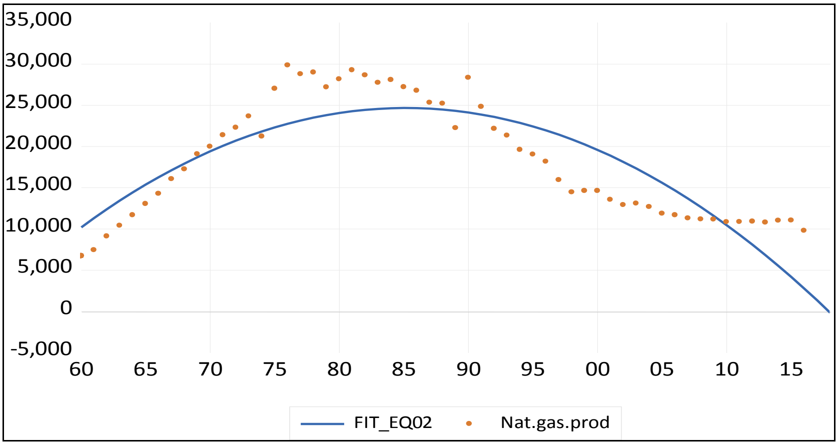
B. Model 2. Parabolic regression

The second model is a parabolic relation according to the next equation:

Eq02: Nat.gas.prod = C(1) + C(2)*TIMP + C(3)*TIMP^2

The graphic representation of Model 2 is displayed in the next figure.

Figure 4. Graphic representation of Model 2 (parabolic)



Source: Own processing according to Eurostat data.

Annex 3 gives us details concerning the performances of Model 2. Compared with Model 1 we may see some improvement in estimated indicators:

- All coefficients are statistically significant;
- R-squared is 0.722603, much higher than the value of Model 1; AIC is lower (19.37) indicating a quality improvement of the model;
- Durbin-Watson statistic has a higher value (0.184); unfortunately, this value is showing as well a strong auto-correlation of residuals.

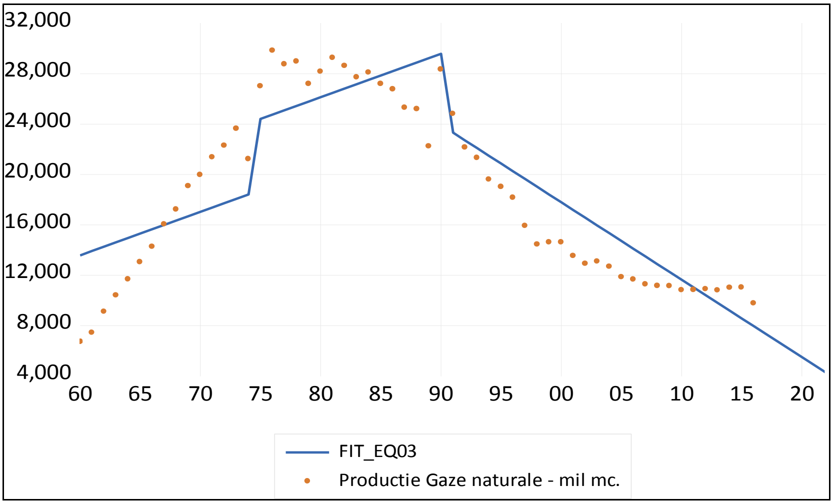
C. Model 3. Linear regression and dummy variables

The third model is based on Model 1 to which two dummy variables are added in order to simulate the shape of the series (the plateau between 1976 and 1990) and to take into account the structural break that took place in 1990. The dummy variables are described in the Annex 4. The equation for Model 3 is as follows:

Eq03: Nat.gas.prod = C(1) + C(2)*TIMP + C(3)*LEVEL3 + C(4)*TREND3

As we may see from the next figure, the new model follows more closely the raw data, including the plateau, between 1976 and 1990 and the structural break from 1990.

Figure 5. Graphic representation of Model 3 (time series with structural break)



Source: Own processing according to Eurostat data.

As regarding the estimated indicators of the Model 3, we may see an improvement compared with the Model 1 and Model 2 (Annex 5):

- All coefficients are statistically significant with probabilities near zero;
- R-squared is 0.831052, much higher than the value of Model 1 and Model 2; AIC is lower (18.913) indicating a tendency to minimize the estimated information loss, compared with previous models;
- Durbin-Watson statistic has a better value of 0.286598 but still far away of the white noise request.

Model Selection

As we have explained in the methodology, we will select the best model among the three models detailed earlier. The selection will be made based on the next criteria: significance of indicators, R-squared, AIC, DWS. Due to the fact that our model would be used for forecasting it will be added a statistic that expresses the forecast accuracy. This statistic is Root Mean Square Error (RMSE). Using RMSE is of a great help when we are using competing forecasts of a single variable due to the fact that it can be difficult to decide which single or composite forecast is “the best”. This is useful as well if we decide which solution would be more appropriate: either to use single forecasting or whether constructing a composite forecast by averaging.

Thus, performing a forecast evaluation we will get the result from the Annex 6. We notice that the Null hypothesis: “Forecast i includes all information contained in others”, is rejected for the Model 1 and cannot be rejected for model 2 and 3. The Evaluation statistics for the three equations shows a clear preference for Model 3 which includes all information contained in Model 1 and Model 2. All statistics considered (RMSE, MAE, MAPE, SMAPE, Theil U1, Theil U2 are better for the Model 3.

The indicators of the three models considered so far are presented in the next table.

Table 3: Comparison of all models

Model	Significance of coefficients (5% prob.)	AIC	R-squared	Durbin Watson statistic	RMSE
Model 1	All coefficients are significant (2 coeff.)	20.519	0.0964	0.065	6673.765
Model 2	All coefficients are significant (3 coeff.)	19.37	0.722	0.184	3697.722
Model 3	All coefficients are significant (3 coeff.)	18.91	0.831	0.286	3273.628

Source: data from Annex 2 to Annex 6.

Analysing Table 1 we can conclude that Model 3 is the best among the proposed models. We see a continuous improvement of all indicators:

- AIC statistic has decreased to the lowest level (18.91) for Model 3; this indicates a quality improvement of the model;
- R-squared has increased to a maximum value of 0.831 for Model 3 from a low level (0.0964) for model 1;
- DWS has steadily increased but it is still showing significant autocorrelation of errors;
- RMSE has the lowest value for Model 3 indicating the fact that, in the case of using this model for forecasting, it will include all information contained in Model 1 and Model 2.

Taking into account all this information we select Model 3 as the best model to describe the series of natural gas production in Romania, in the period from 1960 to 2016.

The following calculations will be made with Model 3 only.

Final model

As it was said, the main shortcoming of the Model 3 is the fact that a significant autocorrelation of errors is present. This issue will be eliminated by altering Model 3 by adding the autoregressive term AR (1). In consequence we will develop Model 4 according to the next specification:

Eq04: Nat.gas.prod = C(1) + C(2)*TIMP + C(3)*LEVEL3 + C(4)*TREND3 + AR(1)

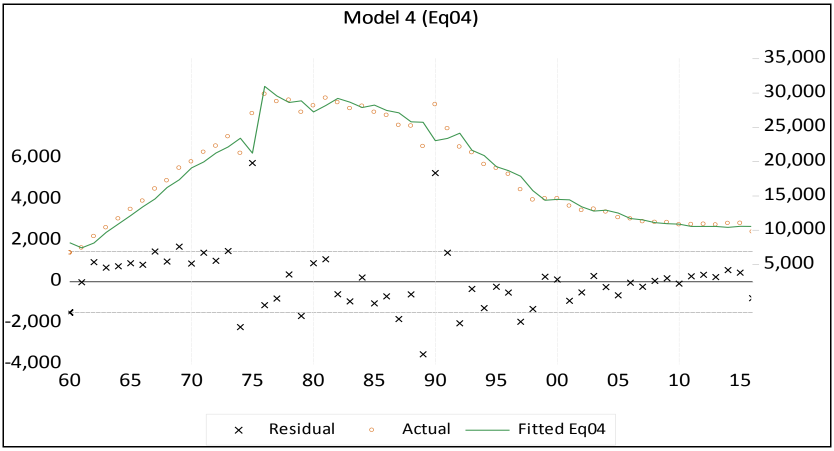
The method used for estimating parameters was ARMA Maximum Likelihood. The estimated output is presented in Annex 7.

As we may see from the next figure, the new model follows the raw data with very few fluctuations or deviations. The residuals have a more random distribution.

If we analyse the output for Eq04 (Annex 7) we may notice the next improvements:

- All coefficients are statistically significant with 5% probability;
- R-squared has a high value: 0.960; this value shows that Model 4 is well fitted;
- AIC has the lowest value among all models: 17.56 showing a quality improvement of the model;
- DWS has a value of 1.968 which is very close to the goal of having random residuals; white noise residuals would have DWS equal to 2.

Figure 6. Graphic representation of Model 4 (AR)



Source: Own processing according to Eurostat data.

As a consequence of all that we consider Model 4 as the best model that describes the natural gas production in Romania, for the period 1960-2016.

Conclusions

As it was shown, the goal of the article to define econometric models for non-stationary time series with structural breaks was fulfilled. The time series considered was that of natural gas production in Romania, in the period from 1960 to 2016. There are a series of difficulties in modelling this series due to the fact that the profile is very irregular: a first period of rapid increase, a second period that has had a more constant tendency and a plateau shape and a third period that displayed a swift decrease. This type of series cannot be modelled well by straight regression line or parabolic functions using OLS method. The solution to accommodate the shape of the series was to add two dummy variables in order to account for the plateau segment and another one to account for the rapid decrease segment.

Out of three models analysed the third one was selected based on a pool of criteria/statistics: significance of indicators, R-squared, AIC, RMSE, DWS.

The remaining issue, autocorrelation of errors, was solved by adding to Model 3 the autoregressive term AR (1). In such circumstances the method called ARMA Maximum Likelihood was used. Thus, Model 4 resulted, having residuals close to white noise request (DWS is very near to the target value 2). All other parameters of Model 4 are superior to the other models taken into account. As a consequence, we have selected Model 4 as the model that is best for describing the evolution of the time series of natural gas production in Romania.

Annexes

Annex 1. Natural gas production, in Romania, 1960-2016 (mil. cm)

Year	Mil. m^3
1960	6707
1961	7424
1962	9091
1963	10388
1964	11672
1965	13038
1966	14252
1967	16036
1968	17220
1969	19066
1970	19971
1971	21365
1972	22287
1973	23639
1974	21217
1975	27001
1976	29834
1977	28755
1978	28973
1979	27189
1980	28156
1981	29263
1982	28620
1983	27719
1984	28083
1985	27196
1986	26763
1987	25301
1988	25195
1989	22222
1990	28336
1991	24807
1992	22138
1993	21318
1994	19598
1995	19016
1996	18162
1997	15916
1998	14441
1999	14617
2000	14607
2001	13524,138
2002	12896,928
2003	13077,018
2004	12663,432
2005	11843,712
2006	11668,59
2007	11271,15
2008	11155,644
2009	11133,288
2010	10811,61
2011	10835,208
2012	10892,34
2013	10789,254
2014	10996,668
2015	11021,508
2016	9764,603

Source: data from Romanian Yearbook of Statistics; 1990-2018.

Annex 2. Estimated output for Model 1

Dependent Variable: Nat.gas.prod				
Method: Least Squares				
Date: 27/01/20 Time: 11:18				
Sample (adjusted): 1960 2016				
Included observations: 57 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	281670.3	108743.5	2.590226	0.0123
TIMP	-132.4991	54.69810	-2.422371	0.0187
R-squared	0.096404	Mean dependent var		18262.18
Adjusted R-squared	0.079975	S.D. dependent var		7083.161
S.E. of regression	6794.023	Akaike info criterion		20.51993
Sum squared resid	2.54E+09	Schwarz criterion		20.59162
Log likelihood	-582.8180	Hannan-Quinn criter.		20.54779
F-statistic	5.867881	Durbin-Watson stat		0.065270
Prob(F-statistic)	0.018742			

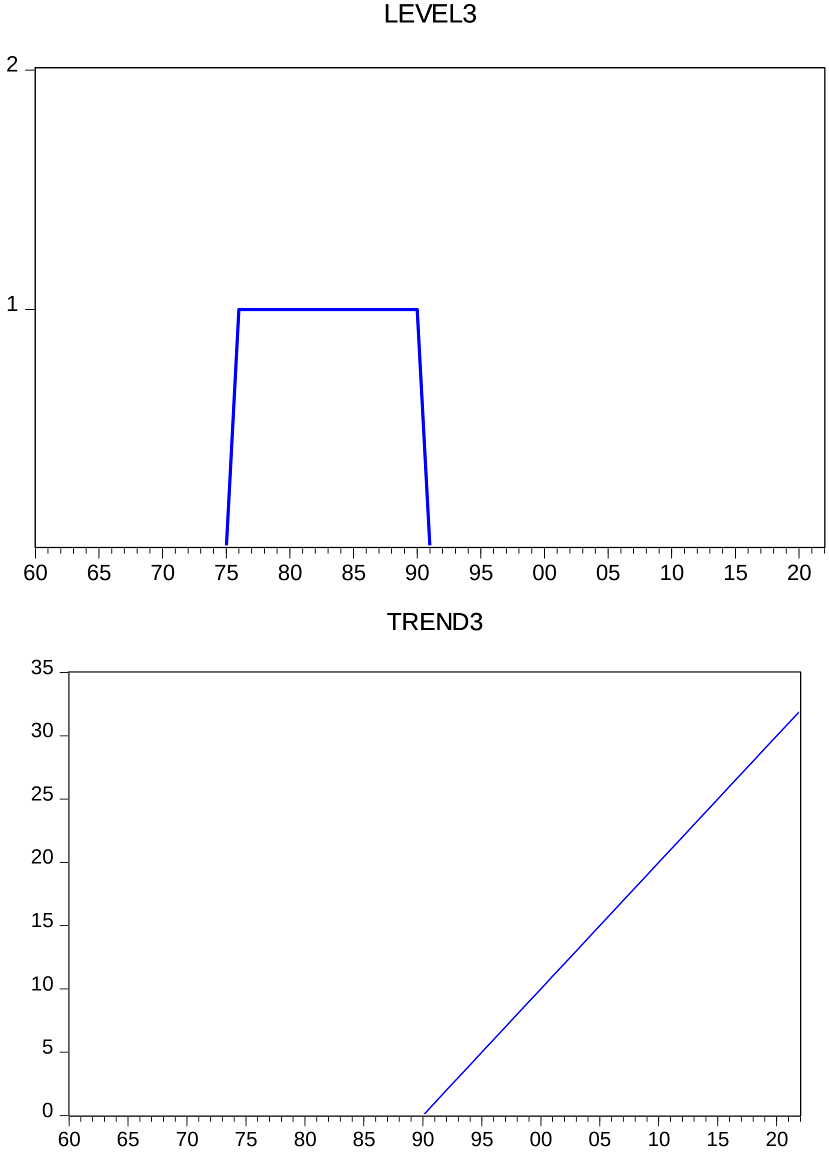
Source: processing data from Annex 1

Annex 3. Estimated output for Model 2

Dependent Variable: Nat.gas.prod				
Method: Least Squares				
Date: 27/01/20 Time: 11:20				
Sample (adjusted): 1960 2016				
Included observations: 57 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-90451084	8218133.	-11.00628	0.0000
TIMP	91154.19	8268.136	11.02476	0.0000
TIMP^2	-22.95943	2.079497	-11.04086	0.0000
R-squared	0.722603	Mean dependent var		18262.18
Adjusted R-squared	0.712329	S.D. dependent var		7083.161
S.E. of regression	3799.048	Akaike info criterion		19.37408
Sum squared resid	7.79E+08	Schwarz criterion		19.48161
Log likelihood	-549.1614	Hannan-Quinn criter.		19.41587
F-statistic	70.33358	Durbin-Watson stat		0.184850
Prob(F-statistic)	0.000000			

Source: processing data from Annex 1

Annex 4. Dummy variable used in Model 3



LEVEL3

TREND3

Source: processing own data.

Annex 5. Estimated output for Model 3

Dependent Variable: Nat.gas.prod				
Method: Least Squares				
Date: 05/02/20 Time: 12:51				
Sample (adjusted): 1960 2016				
Included observations: 57 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-584328.4	117156.9	-4.987572	0.0000
TIMP	304.8447	59.38281	5.133552	0.0000
LEVEL3	7438.715	1145.628	6.493132	0.0000
TREND3	-833.7073	122.5695	-6.801917	0.0000
R-squared	0.831052	Mean dependent var		18262.18
Adjusted R-squared	0.821489	S.D. dependent var		7083.161
S.E. of regression	2992.675	Akaike info criterion		18.91331
Sum squared resid	4.75E+08	Schwarz criterion		19.05669
Log likelihood	-535.0295	Hannan-Quinn criter.		18.96903
F-statistic	86.90210	Durbin-Watson stat		0.286598
Prob(F-statistic)	0.000000			

Source: processing data from Annex 1

Annex 6. Forecast Evaluation of the estimated models

Forecast Evaluation						
Sample: 1960 2022						
Included observations: 63						
Evaluation sample: 1960 2022						
Number of forecasts: 3						
Combination tests						
Null hypothesis: Forecast i includes all information contained in others						
Equation	F-stat	F-prob				
EQ01	62.30814	0.0000				
EQ02	1.356501	0.2662				
EQ03	0.683761	0.5090				
Evaluation statistics						
Forecast	**RMSE**	MAE	MAPE	SMAPE	Theil U1	Theil U2
EQ01	6673.765	5678.386	37.90205	31.62289	0.175824	5.690178
EQ02	3697.722	3207.577	21.14041	21.45574	0.095357	2.876339
EQ03	3273.628	2727.858	18.27135	16.55953	0.083271	2.625469

Source: processing own data.

Annex 7. Estimated output for Model 4

Dependent Variable: nat.gas.prod				
Method: ARMA Maximum Likelihood (OPG - BHHH)				
Sample: 1960 2016				
Included observations: 57				
Estimation settings: tol= 0.00010				
Initial Values: C(1)=13167.3, C(2)=304.845, C(3)=7438.71, C(4)=-833.707,				
C(5)=0.81549, C(6)=2143264				
Convergence achieved after 24 iterations				
Coefficient covariance computed using outer product of gradients				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10426.52	3182.730	3.275968	0.0019
@TREND	422.9703	150.2107	2.815847	0.0069
LEVEL3	4397.378	1029.872	4.269830	0.0001
TREND3	-899.8449	443.9394	-2.026954	0.0479
AR(1)	0.916414	0.055080	16.63793	0.0000
SIGMASQ	1958376.	265884.5	7.365512	0.0000
R-squared	0.960269	Mean dependent var		18262.18
Adjusted R-squared	0.956374	S.D. dependent var		7083.161
S.E. of regression	1479.450	Akaike info criterion		17.56816
Sum squared resid	1.12E+08	Schwarz criterion		17.78322
Log likelihood	-494.6926	Hannan-Quinn criter.		17.65174
F-statistic	246.5270	Durbin-Watson stat		1.968606
Prob(F-statistic)	0.000000			

Source: processing own data.

References

Davidson and MacKinnon (1993), Estimation and Inference in Econometrics, http://qed.econ.queensu.ca/pub/faculty/mackinnon/papers/cea-presadd-2002.pdf .

Greene (2008), Econometric Analysis, 6th Edition, https://link.springer.com/article/10.1007/s00362-010-0315-8 .

Johnston, DiNardo (1997), Econometric Methods, 4th Edition, https://economics.ut.ac.ir/documents/3030266/14100645/econometric%20methods-johnston.pdf .

Pindyck and Rubinfeld (1998), Econometric Models and Economic Forecasts, 4th edition, https://pdfweek.com/downloads/econometric%20models%20and%20economic%20forecasts%204th%20edition%20pdf .

Pontragin L.S, Boltianskii V.G, Amkrelidze R. V, Mischenko E. F (1962) The Mathematical Theory of Optimal Processes (Russian), English translation: Interscience.

Sadoulet, E., De Janvry (1995), Quantitative Development Analysis, John Hopkins University Press.

Taylor, L. (Ed.) (1990), Socially Relevant Policy Analysis, Structuralist Computable General Equilibrium Models for the Developing World, MIT Press., Cambridge, Mass.

Platon V., Turdeanu A., (2006), The Sustainable Development in the EU and Romania: Comparative Analysis, Romanian Journal of Economics 23 (2 (32)), 91-99, https://ideas.repec.org/a/ine/journl/tome23y2006(xvi)i2(32)p91-99.html

Platon V., Constantinescu A., 2019, Econometric Models Of Oil Production In Romania, http://www.strategiimanageriale.ro/images/images_site/articole/article_811a85ac3357b696b57c3a0ebbd0ba25.pdf

Whitney, J.D., (1994), A Course in macroeconomic modelling and forecasting, Harvester Wheatsheaf.

Wooldridge (2013), Introductory Econometrics: A Modern Approach, 5th Edition.

*** Anuarul statistic al României, 1991-2018, Institutul National de Statistica.

*** EUROSTAT, Statistics Explained, https://ec.europa.eu/eurostat/statistics-explained/index.php/Energy_production_and_imports/ro#Produc.C8.9Bia_de_energie_primar.C4.83_a_sc.C4.83zut_.C3.AEntre_anii_2006_.C8.99i_2016