JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

MODELLING UNEMPLOYMENT RATE USING BOX-JENKINS PROCEDURE

Ion DOBRE
PhD, University Professor, Department of Economic Cybernetics
University of Economics, Bucharest, Romania

E-mail: dobrerio@ase.ro

Adriana AnaMaria ALEXANDRU
PhD Candidate, University Assistant, Department of Statistics and Econometrics
University of Economics, Bucharest, Romania

E-mail: adrianaalexandru@yahoo.com

Abstract: *This paper aims to modelling the evolution of unemployment rate using the Box-Jenkins methodology during the period 1998-2007 monthly data. The empirical study relieves that the most adequate model for the unemployment rate is ARIMA (2,1,2). Using the model, we forecasts the values of unemployment rate for January and February 2008.Therefore, the unemployment rate for January 2008 is 4.06%.*

Key words: *Unemployment rate; Box-Jenkins methodology; ARIMA models; Romania*

1. Theoretical Background

The pioneers in this area was Box and Jenkins who popularized an approach that combines the moving average and the autoregressive models in the book[1].Although both autoregressive and moving average approaches were already known (and were originally investigated by Yule), the contribution of Box and Jenkins was in developing a systematic methodology for identifying and estimating models that could incorporate both approaches. This makes Box-Jenkins models a powerful class of models.

The Box-Jenkins ARMA model is a combination of the AR and MA models as follows:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + ...a_p y_{t-p} - b_1 u_{t-1} - b_2 u_{t-2} - ..... - b_q u_{t-q} + u_t$$

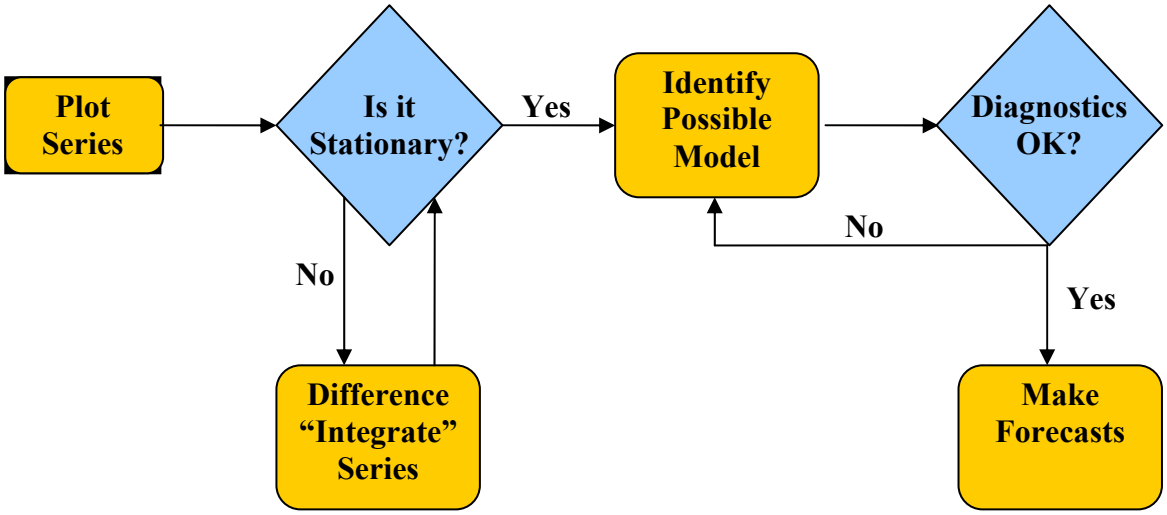


Figure1. Box-Jenkins procedure

There are three primary stages in building a Box-Jenkins time series model:

1. **Model Identification**
2. **Model Estimation**
3. **Model Validation**

1.1. Box-Jenkins Model Identification

The identification stage is the most important and also the most difficult: it consists to determine the adequate model from ARIMA family models. The most general Box-Jenkins model includes difference operators, autoregressive terms, moving average terms, seasonal difference operators, seasonal autoregressive terms, and seasonal moving average terms[2].This phase is founded on the study of autocorrelation and partial autocorrelation.

The first step in developing a Box-Jenkins model is to determine if the series is stationary and if there is any significant seasonality that needs to be modelled.

Stationarity in Box-Jenkins Models

The Box-Jenkins model assumes that the time series is *stationary*. A stationary series has:

1. Constant mean
2. Constant variance
3. Constant autocorrelation structure

Regression with nonstationary variables is a spurious correlation. The random walk $y_t = y_{t-1} + u_t \;\; u_t \sim \mathrm{N}(0, \sigma^2)$ is not stationary, since its variance increases linearly with time t. Stationarity can be assessed from a run sequence plot. The run sequence plot should show constant location and scale. It can also be detected from an autocorrelation plot. Specifically, non-stationarity is often indicated by an autocorrelation plot with very slow decay.

Box and Jenkins recommend differencing non-stationary series one or more times to achieve stationarity. Doing so produces an ARIMA model, with the "I" standing for

"Integrated". But its first difference $\Delta y_t = y_t - y_{t-1} = u_t$ is stationary, so y is „integrated of order 1", or y ~ I (1).

Testing for non-stationarity

1. Autocorrelation function (Box-Jenkins approach)-if autocorrelations start high and decline slowly, then series is nonstationary, and should be differenced.

2. Dickey-Fuller test

$y_t = a + by_{t-1} + u_t$ would be a nonstationary random walk if b = 1. So to find out if y has a "unit root" we regress: $\Delta y_t = a + cy_{t-1} + u_t$ where c = b-1 and test hypothesis that c = 0 against c < 0 (like a "t-test").

Seasonality in Box-Jenkins Models

Box-Jenkins models can be extended to include *seasonal* autoregressive and seasonal moving average terms.

Model identification: seasonality of order s is revealed by "spikes" at s, 2s, 3s, lags of the autocorrelation function.

Model estimation: to make series stationary, may need to take s-th differences of the raw data before estimation. These seasonal effects may themselves follow AR and MA processes.

At the model identification stage, our goal is to detect seasonality, if it exists, and to identify the order for the seasonal autoregressive and seasonal moving average terms. For Box-Jenkins models, it isn't necessary remove seasonality before fitting the model. Instead, it can include the order of the seasonal terms in the model specification to the ARIMA estimation software.

Once stationarity and seasonality have been addressed, the next step is to identify the order (the ***p*** and ***q***) of the autoregressive and moving average terms. The primary tools for doing this are the autocorrelation plot and the partial autocorrelation plot. The sample autocorrelation plot and the sample partial autocorrelation plot are compared to the theoretical behaviour of these plots when the order is known.

Order of Autoregressive Process (p)

Specifically, for an AR (1) process, the sample autocorrelation function should have an exponentially decreasing appearance. However, higher-order AR processes are often a mixture of exponentially decreasing and damped sinusoidal components. For higher-order autoregressive processes, the sample autocorrelation needs to be supplemented with a partial autocorrelation plot. The partial autocorrelation of an AR (***p***) process becomes zero at lag ***p*+1** and greater, so we examine the sample partial autocorrelation function to see if there is evidence of a departure from zero. This is usually determined by placing a 95% confidence interval on the sample partial autocorrelation plot (most software programs that generate sample autocorrelation plots will also plot this confidence interval). If the software program does not generate the confidence band, it is approximately $\pm 2/\sqrt{N}$, with *N* denoting the sample size.

The data is AR (p) if: ACF will decline steadily, or follow a damped cycle and PACF will cut off suddenly after p lags.

Order of Moving Average Process (q)

The autocorrelation function of a MA (***q***) process becomes zero at lag ***q*+1** and greater, so we examine the sample autocorrelation function to see where it essentially becomes zero.

The following table summarizes how we use the sample autocorrelation function for model identification.

Table 1. The type of the model

Shape	***Indicated Model***
Exponential, decaying to zero	Autoregressive model. Use the partial autocorrelation plot to identify the order of the autoregressive model.
Alternating positive and negative, decaying to zero	Autoregressive model. Use the partial autocorrelation plot to help identify the order.
One or more spikes, rest are essentially zero	Moving average model, order identified by where plot becomes zero.
Decay, starting after a few lags	Mixed autoregressive and moving average model.
All zero or close to zero	Data is essentially random.
High values at fixed intervals	Include seasonal autoregressive term.
No decay to zero	Series is not stationary.

The data is MA (q) if: ACF will cut off suddenly after q lags and PACF will decline steadily, or follow a damped cycle.

It's not indicated to build models with:

– Large numbers of MA terms

– Large numbers of AR and MA terms together

You may well see very (suspiciously) high t-statistics. This happens because of high correlation ("colinearity") among regressors, *not* because the model is good.

1.2. Box-Jenkins Model Estimation

The main approaches to fitting Box-Jenkins models are non-linear least squares and maximum likelihood estimation. Maximum likelihood estimation is generally the preferred technique[3].

1.3. Box-Jenkins Model Diagnostics

Model diagnostics for Box-Jenkins models is similar to model validation for non-linear least squares fitting. Model diagnostics for Box-Jenkins models is similar to model validation for non-linear least squares fitting.

That is, the error term u_t is assumed to follow the assumptions for a stationary unvaried process. The residuals should be white noise (or independent when their distributions are normal) drawings from a fixed distribution with a constant mean and variance.

If the Box-Jenkins model is a good model for the data, the residuals should satisfy these assumptions. If these assumptions are not satisfied, we need to fit a more appropriate model. That is, we go back to the model identification step and try to develop a better

model. Hopefully the analysis of the residuals can provide some clues as to a more appropriate model. The residual analysis is based on:

1. Random residuals: the Box-Pierce Q-statistic: $Q(s) = n\sum r(k)^2 \approx \chi^2(s)$ where r (k) is the k-th residual autocorrelation and summation is over first s autocorrelations.

2. Fit versus parsimony: the Schwartz Bayesian Criterion (SBC):

SBC = ln {RSS/n} + (p+d+q) ln (n)/n, where RSS = residual sum of squares, n is sample size, and (p+d+q) the number of parameters.

2. The data

The variable used in the analysis is the unemployment rate that ran from 1998 to the end of 2007 and its available monthly. The source of data is the Monthly Bulletins of National Bank of Romania.

Stage 1: The time series analysis

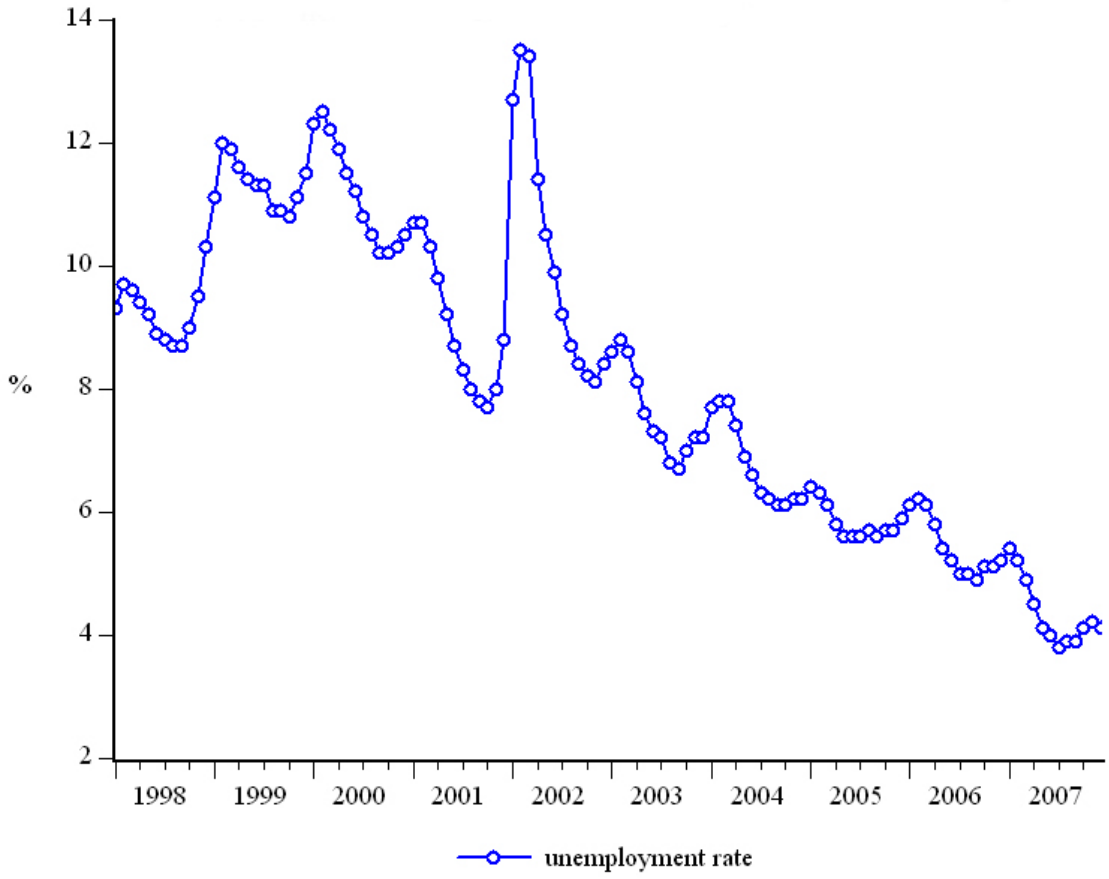


Figure 2. The unemployment rpate evolution during the period 1998-2007
Source: Montly Bulletins of National Bank of Romania

The data presents some seasonal fluctuations and that is the reason for with data has been seasonally adjusted, using the moving average method implemented in Eviews program.

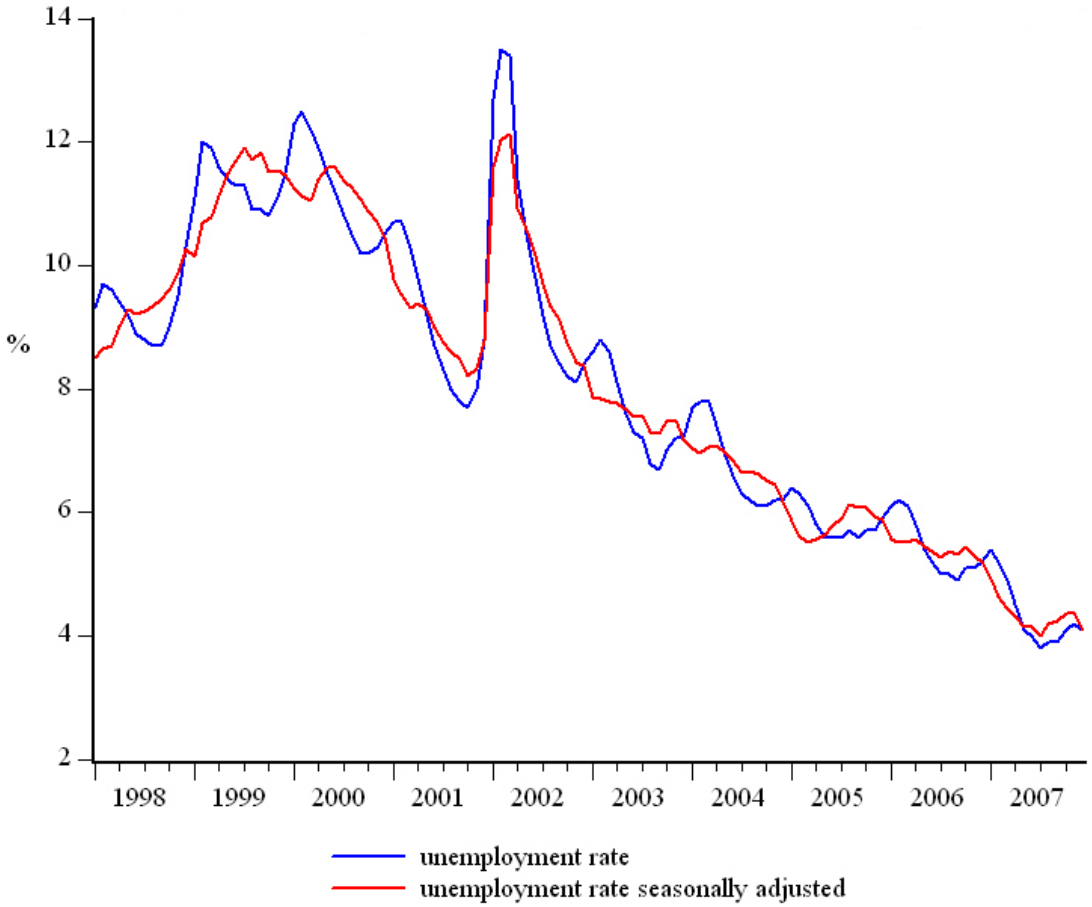
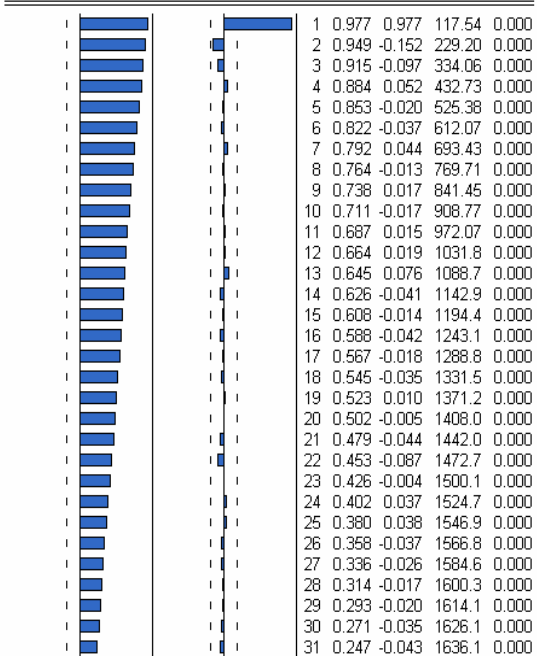


Figure 3. The unemployment rate and the unemployment rate seasonally adjusted

The first step in developing a Box-Jenkins model is to determine if the series is stationary. For this, we use the autocorrelation function (ACF) and Augmented Dickey-Fuller test (ADF).

Because the autocorrelation (ACF) start high and decline slowly, then series is nonstationary, and should be differenced. We have analyzed the data series stationarity by using the Augmented Dickey-Fuller (ADF) test, who reveals the fact that the zero hypotheses is accepted, the series has a root unit and it is non stationary. It becomes stationary by first order differences.

Date: 03/18/08 Time: 00:29
Sample: 1998M01 2007M12
Included observations: 120



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.977	0.977	117.54	0.000
		2	0.949	-0.152	229.20	0.000
		3	0.915	-0.097	334.06	0.000
		4	0.884	0.052	432.73	0.000
		5	0.853	-0.020	525.38	0.000
		6	0.822	-0.037	612.07	0.000
		7	0.792	0.044	693.43	0.000
		8	0.764	-0.013	769.71	0.000
		9	0.738	0.017	841.45	0.000
		10	0.711	-0.017	908.77	0.000
		11	0.687	0.015	972.07	0.000
		12	0.664	0.019	1031.8	0.000
		13	0.645	0.076	1088.7	0.000
		14	0.626	-0.041	1142.9	0.000
		15	0.608	-0.014	1194.4	0.000
		16	0.588	-0.042	1243.1	0.000
		17	0.567	-0.018	1288.8	0.000
		18	0.545	-0.035	1331.5	0.000
		19	0.523	0.010	1371.2	0.000
		20	0.502	-0.005	1408.0	0.000
		21	0.479	-0.044	1442.0	0.000
		22	0.453	-0.087	1472.7	0.000
		23	0.426	-0.004	1500.1	0.000
		24	0.402	0.037	1524.7	0.000
		25	0.380	0.038	1546.9	0.000
		26	0.358	-0.037	1566.8	0.000
		27	0.336	-0.026	1584.6	0.000
		28	0.314	-0.017	1600.3	0.000
		29	0.293	-0.020	1614.1	0.000
		30	0.271	-0.035	1626.1	0.000
		31	0.247	-0.043	1636.1	0.000

Figure 4. The correlogram of unemployment rate seasonally adjusted

Null Hypothesis: RSSA98 has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Fixed)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-3.037670	0.1266
Test critical values:	1% level	-4.036983	
	5% level	-3.448021	
	10% level	-3.149135	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(RSSA98)
Method: Least Squares
Date: 03/26/08 Time: 16:07
Sample (adjusted): 1998M02 2007M12
Included observations: 119 after adjustments

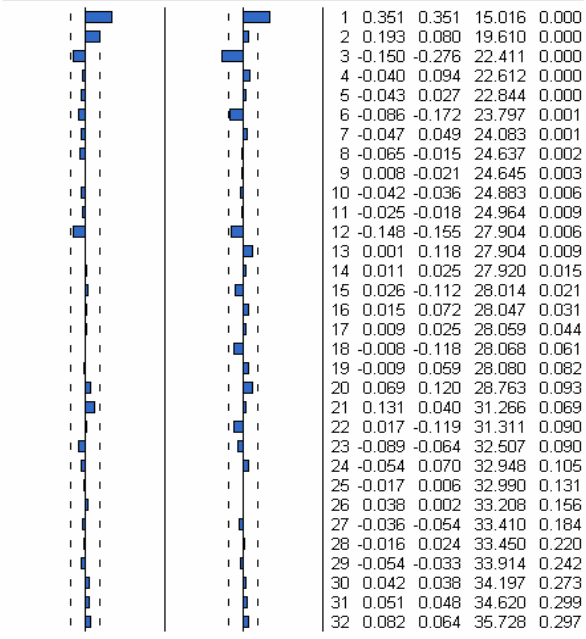
Variable	Coefficient	Std. Error	t-Statistic	Prob.
RSSA98(-1)	-0.085244	0.028062	-3.037670	0.0029
C	1.056913	0.336113	3.144520	0.0021
@TREND(1998M01)	-0.006726	0.001941	-3.464857	0.0007

R-squared	0.093806	Mean dependent var	-0.037015
Adjusted R-squared	0.078182	S.D. dependent var	0.354721
S.E. of regression	0.340572	Akaike info criterion	0.708507
Sum squared resid	13.45477	Schwarz criterion	0.778569
Log likelihood	-39.15619	F-statistic	6.003934
Durbin-Watson stat	1.312423	Prob(F-statistic)	0.003302

Figure 5. The Augmented Dickey-Fuller test results

Stage 2: The identification - the autocorrelation is computed on the first differences series

Sample: 1998M01 2007M12
Included observations: 119



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.351	0.351	15.016	0.000
		2	0.193	0.080	19.610	0.000
		3	-0.150	-0.276	22.411	0.000
		4	-0.040	0.094	22.612	0.000
		5	-0.043	0.027	22.844	0.000
		6	-0.086	-0.172	23.797	0.001
		7	-0.047	0.049	24.083	0.001
		8	-0.065	-0.015	24.637	0.002
		9	0.008	-0.021	24.645	0.003
		10	-0.042	-0.036	24.883	0.006
		11	-0.025	-0.018	24.964	0.009
		12	-0.148	-0.155	27.904	0.006
		13	0.001	0.118	27.904	0.009
		14	0.011	0.025	27.920	0.015
		15	0.026	-0.112	28.014	0.021
		16	0.015	0.072	28.047	0.031
		17	0.009	0.025	28.059	0.044
		18	-0.008	-0.118	28.068	0.061
		19	-0.009	0.059	28.080	0.082
		20	0.069	0.120	28.763	0.093
		21	0.131	0.040	31.266	0.069
		22	0.017	-0.119	31.311	0.090
		23	-0.089	-0.064	32.507	0.090
		24	-0.054	0.070	32.948	0.105
		25	-0.017	0.006	32.990	0.131
		26	0.038	0.002	33.208	0.156
		27	-0.036	-0.054	33.410	0.184
		28	-0.016	0.024	33.450	0.220
		29	-0.054	-0.033	33.914	0.242
		30	0.042	0.038	34.197	0.273
		31	0.051	0.048	34.620	0.299
		32	0.082	0.064	35.728	0.297

Figure 6. The Correlogram of first differences of unemployment rate

By applying the ADF test for the series of the first order differences one can observe that the series becomes stationary, so the initial series of the monthly unemployment rate is integrated by first order.

As a result, we have applied the Box- Jenkins procedure on the stationary data series and we want to identify the corresponding ARIMA (p, q) process. The series corelogram has allowed us to choose appropriate p and q for the data series. We have estimated more models in order to determine the right specification, by choosing from both the different models estimated on the informational criteria Akaike and by generating predictions on the basis of estimated models. The series corelogram suggests the necessity of introduction in the process estimation of both the analyzed variable lags and the lags of the error. We have started with an AR (1) process and further analyzed the residual corelogram in order to catch the correlations and autocorrelations from lags bigger that 1. From Akaike criteria's point of view, the proper model to best adjust the data is ARIMA (2, 1,2).

Stage 3: The Estimation

Dependent Variable: D(RSSA98)
Method: Least Squares
Date: 03/27/08 Time: 15:16
Sample (adjusted): 1998M04 2007M12
Included observations: 117 after adjustments
Convergence achieved after 21 iterations
Backcast: 1998M02 1998M03

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	-0.443316	0.077808	-5.697588	0.0000
AR(2)	-0.515496	0.072797	-7.081316	0.0000
MA(1)	0.813233	0.009985	81.44546	0.0000
MA(2)	0.977082	0.013197	74.03841	0.0000

R-squared	0.281409	Mean dependent var	-0.039311
Adjusted R-squared	0.262331	S.D. dependent var	0.357265
S.E. of regression	0.306846	Akaike info criterion	0.508651
Sum squared resid	10.63948	Schwarz criterion	0.603084
Log likelihood	-25.75609	Durbin-Watson stat	1.848075

Inverted AR Roots	-.22+.68i	-.22-.68i
Inverted MA Roots	-.41+.90i	-.41-.90i

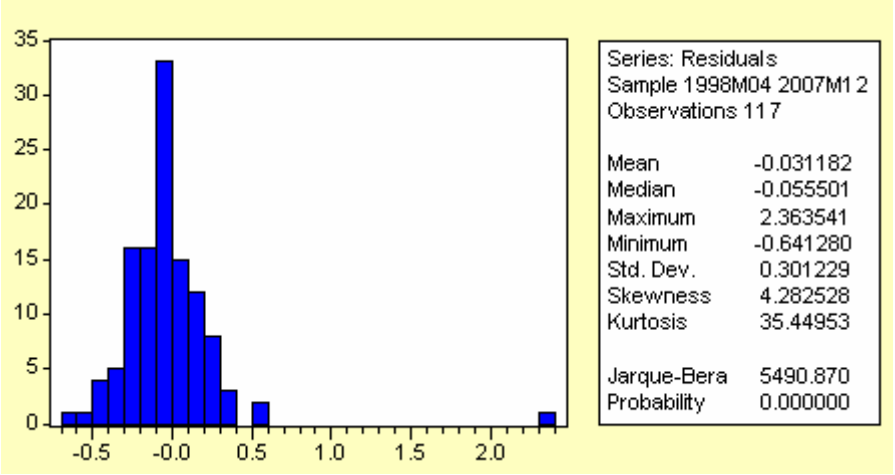
Figure 7. The ARIMA model estimation

Stage 4: The Model's Adaptation

The coefficients of the model are significantly different of 0 (the t-test). The others statistics (DW, F-stat) let portend a good fitting. The determination coefficient R-squared is 28.14%.

The residual analysis is based on two criterions:

- The normality test point out that the average of residuals is approximately 0.



- The residual is a white noise, analysing the autocorrelation. Any term isn't exterior to the confidence intervals and the Q-statistic has a critical probability near to 1. The residue it may be assimilate to a white noise process.

Date: 03/27/08 Time: 15:52
Sample: 1998M04 2007M12
Included observations: 117
Q-statistic probabilities adjusted for 4 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.072	0.072	0.6292	
		2	-0.002	-0.007	0.6297	
		3	0.026	0.026	0.7097	
		4	0.010	0.006	0.7214	
		5	0.017	0.016	0.7555	0.385
		6	-0.002	-0.005	0.7558	0.685
		7	-0.005	-0.005	0.7593	0.859
		8	0.002	0.002	0.7599	0.944
		9	-0.017	-0.018	0.7974	0.977
		10	-0.018	-0.015	0.8384	0.991
		11	-0.014	-0.012	0.8655	0.997
		12	0.083	0.087	1.7869	0.987
		13	0.001	-0.011	1.7870	0.994
		14	-0.024	-0.021	1.8626	0.997
		15	-0.014	-0.015	1.8896	0.999
		16	-0.013	-0.012	1.9128	1.000
		17	-0.023	-0.024	1.9875	1.000
		18	-0.021	-0.017	2.0483	1.000
		19	-0.023	-0.018	2.1214	1.000
		20	-0.022	-0.019	2.1935	1.000
		21	-0.011	-0.004	2.2113	1.000
		22	-0.027	-0.022	2.3163	1.000
		23	-0.020	-0.013	2.3749	1.000
		24	-0.015	-0.020	2.4072	1.000

Figure 8. The Correlogram of Residuals Squared

Therefore, the estimation of ARIMA (2,1,2) model is validated, the time series can be described by an ARIMA(2,1,2) process. The unemployment rate seasonally adjusted times series and in first differences (DRSSA) is described by the process:

$$DRSSA = -0.4433 \cdot RSSA_{t-1} - 0.5154 \cdot RSSA_{t-2} + 0.8132 \cdot u_{t-1} + 0.9777 \cdot u_{t-2}$$

Stage 5: The forecasting

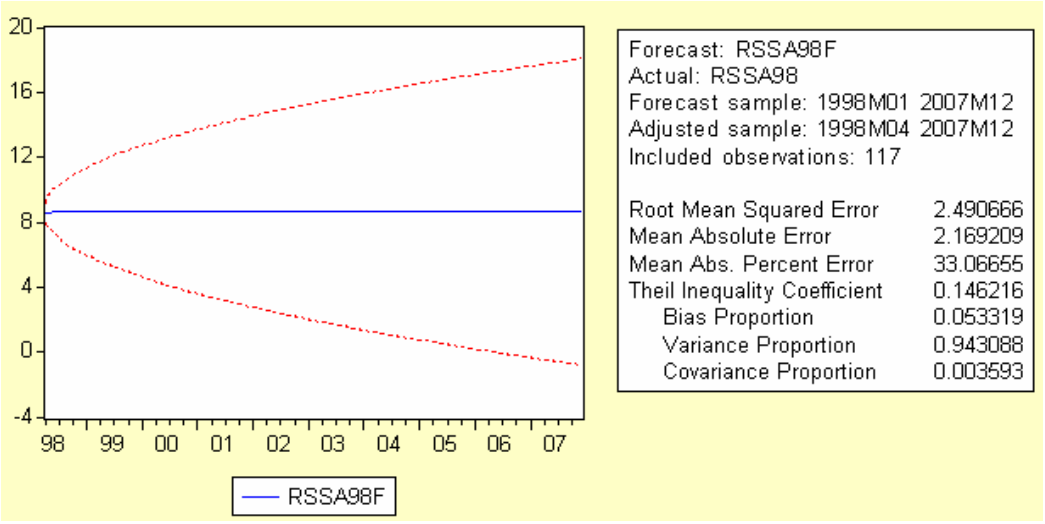
The forecasting is computed by reaggregation of different components. The residual values for the months of December and November are: $u_{2007:12} = -0.21465$, $u_{2007:11} = -0.15538$.

The fitting values of ARIMA model for unemployment rate are:

$\hat{RSSA}_{2007:12} = -0.06526$, $\hat{RSSA}_{2007:11} = 0.15115$.

Table 2. The unemployment rate forecasts

	u_t	DRSSA	RSSA	Seasonal Coefficients	Unemployment rate(%)
November 2007	-0,21465				
December 2007	-0,15538		4,08929		
January 2008		-0,37544	3,71384	1.0948	4,06
February 2008		-0,0098	4,15817	1.1226	4,15
March 2008		0,197846	4,10594	1.1048	4,35



Using an ARIMA (2,1,2) model of monthly values series of unemployment rate we can predict the value of unemployment rate for January and February 2008. In January 2008 the unemployment rate forecasted by the model was 4, 06% and for February 4,15%. The result troves sustainability into the monthly bulletin of National Institute of Statistics. According to this publication the unemployment rate is 4.3% for January 2008.

Bibliography

1. Bourbonnais, R. **Econométrie,** 6nd. ed., Dunod, Paris, 2005
2. Box, G. E. P. and Jenkins, G. M. **Time Series Analysis, Forecasting and Control,** Holden Day, San Francisco, 1970
3. Box, G. E. P., Jenkins, G. M. and Reinsel, G. C.**Time Series Analysis, Forecasting and Control,** 3rd ed., Prentice Hall, Englewood Clifs, 1994
4. Brockwell, P.J. and Davis, R. A. **Introduction to Time Series and Forecasting,** 2nd. ed., Springer Verlag, 2002
5. Chatfield, C. **The Analysis of Time Series,** 5th ed., Chapman & Hall, New York, NY, 1996
6. Cuthbertson, K. and Hall, S. **Applied Econometrics Techniques,** Whreatons Ltd., 1995
7. Dobre, I. and Alexandru, A. **Scenarios regarding the unemployment rate evolution evolution in Romania during the period 2006-2009,** Economic Computation and Economic Cybernetics Studies and Research, vol. 41, No. 2, 2007, pp. 39-52
8. Enders, W. **Applied Econometric Time Series,** Wiley, New York, 2004
9. Gujarati, D. **Basic Econometrics,** McGraw-Hill Inc., N. Y., 1995
10. Hamilton, J. **Time Series Analysis,** Princeton University Press, Princeton, 1994
11. Johnston, J. **Econometric Methods,** McGraw-Hill, N. Y., 1991
12. Maddala, G. S. **Introduction to Econometrics,** McMillan Public., 1988
13. Pecican, E., S. **Econometrie,** Ed. All, Bucharest, 1994
14. Pecican, E. S. and Tanasoiu, O. **Modele Econometrice,** Ed. ASE, Bucharest, 2001
15. Pecican, E. S. **Econometrie,** Ed. C.H. Beck, Bucharest, 2006
16. Pecican, E. S. **Econometria pentru...economisti: Econometrie-teorie si aplicatii,** Ed. Economica, Bucharest, 2005
17. www.bnr.ro, National Bank of Romania, **Monthly Bulletins**

---

[1] Box, G. E. P., Jenkins, G. M. and Reinsel, G. C.**Time Series Analysis, Forecasting and Control,** 3rd ed., Prentice Hall, Englewood Cliffs, 1994

[2] Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. **Time Series Analysis, Forecasting and Control,** 3rd ed. Prentice Hall, Englewood Cliffs, 1994
Chatfield, C.**The Analysis of Time Series,** 5th ed., Chapman & Hall, New York, NY, 1996
Brockwell, P. J. and Davis, R. A.**Introduction to Time Series and Forecasting,** 2nd. ed., Springer-Verlag, 2002

[3] Bourbonnais, R.**Econométrie,** 6e éd, Dunod, Paris, 2005