Qualitative variables and their reduction possibility. Application to time series models

Ciuiu, Daniel

Technical University of Civil Engineering Bucharest, Romanian Institute for Economic Forecasting

June 2013

XI Balkan Conference on Operational Research
BALCOR 2013
Belgrade, Zlatibor, 07-10 September 2013
UNIVERSITY OF BELGRADE
FACULTY OF ORGANISATIONAL SCIENCES

QUALITATIVE VARIABLES AND THEIR REDUCTION POSSIBILITY. APPLICATION TO TIME SERIES MODELS

DANIEL CIUIU[1]

[1] Department of Mathematics and Computer Science, Technical University of Civil Engineering, Bucharest, Bd. Lacul Tei Nr. 124, Bucharest, Romania & Romanian Institute for Economic Forecasting, Str. 13 Septembrie 13, Bucharest, Romania, dciuiu@yahoo.com

Abstract: *In this paper we will study the influence of qualitative variables on the unit root tests for stationarity. For the linear regressions involved the implied assumption is that they are not influenced by such qualitative variables. For this reason, after we have introduced such variables, we check first if we can remove some of them from the model.*

The considered qualitative variables are according the corresponding coefficient (the intercept, the coefficient of X_{t-1} and the coefficient of t), and on the different groups built tacking into account the characteristics of the time moments.

Keywords: *Qualitative variables, Dickey-Fuller, ARIMA, GDP, homogeneity.*

1. INTRODUCTION

In the general case a time series can be decomposed in three parts [1, 3, 5]: the trend, the sesonal component and the stationary component. If there is no sesonnal component, a method to estimate and remove the trend is the moving average. The moving average of order q is

$$\widehat{m}_t = \frac{\sum_{j=-q}^{q} X_{t+j}}{2\cdot q+1}. \tag{1}$$

In [1] there are considered $X_t = X_1$ for $t < 1$, and $X_t = X_n$ for $t > n$, and in [3] and [5] there are computed only the values for which $q < t \leq n-q$, hence all the terms in the above relation exist in the time series.

A criterion to choose q used in [3] is the minimum variance of $X_t - \widehat{m}_t$.

If the above time series X_t contains also a seasonal component, having the period s, then we remove first this component as follows.

Consider two cases: $s = 2\cdot q+1$ and $s = 2\cdot q$. In the first case we estimate $\widehat{m}_t$ according (1), and in the second case we estimate

$$\widehat{m}_t = \frac{\frac{X_{t-q}+X_{t+q}}{2} + \sum_{j=-q+1}^{q-1} X_{t+j}}{2\cdot q}. \tag{2}$$

Next we compute the average y_k of $X_{k+js} - \widehat{m}_{k+js}$ for $q < k+js \leq n-q$, and from here the seasonal component

$$\begin{cases} \widehat{c}_k = y_k \text{ for } k = \overline{1,s} \\ \widehat{c}_k = \widehat{c}_{k-s} \text{ for } k > s \end{cases}. \tag{3}$$

The time series $\widetilde{X}_t = X_t - \widehat{c}_t$ has no more seasonal component, and we apply (1) (with another q) for removing the trend. Obviously, the criterion to choose s and q_1 is the minimum variance of the obtained stationary time series [3].

Another method to separate the three components is the differenciating method [1, 3, 5]. We denote first

$$\begin{cases} \Delta X_t = X_t - X_{t-1} \\ \Delta_s X_t = X_t - X_{t-s} \end{cases}, \tag{4}$$

where s is the number of seasons.

The above operator Δ is the difference operator, and the operator Δ_s is the seasonal difference operator, with the period s. If the time series X_t has a seasonal component with the period s, then there exists $n_s > 0$ such that

$$Y_t = \Delta_s^{n_s} X_t \tag{5}$$

has no seasonal component. Otherwise, consider $Y_t = X_t$. If the time series Y_t has trend, then there exists $d > 0$ such that

$$Z_t = \Delta^d Y_t \tag{6}$$

is stationary. Analogous to the case of lack of seasonal component, if Y_t has no trend we have $Z_t = Y_t$.

► **Definition 1.** The time series X_t without seasonal component is $ARIMA(p,d,q)$ if the time series $Y_t = \Delta^d X_t$ is $ARMA(p,q)$.

The exponential smoothing is another method to obtain a stationary time series [1, 3, 5]. Starting from the initial time series X_t and from the real number $a \in (0,1)$, we define

$$\begin{cases} \widehat{m}_1 = X_1 \\ \widehat{m}_t = a \cdot X_t + (1-a) \cdot \widehat{m}_{t-1} \text{ for } t > 1 \end{cases}. \tag{7}$$

From here we obtain for $t > 1$ by computations

$$\widehat{m}_t = (1-a)^{t-1} X_1 + \sum_{j=0}^{t-2} a(1-a)^j X_{t-j}. \tag{8}$$

We notice that the decrease of the coefficients of X_t, X_{t-1},..., X_2 is exponential, and this justifies the name of exponential smoothing.

The criterion for choosing a is such that

$$\sum_{j=1}^{t} (X_t - \widehat{m}_t)^2 \tag{9}$$

is minimum [3].

To decide between time series models, we use unit root tests. One of them is the Dickey—Fuller test [3]. For this, consider the models

$$X_t = \alpha X_{t-1} + a_t, \text{ with } |\alpha| < 1 \tag{10a}$$
$$X_t = X_{t-1} + a_t \tag{10b}$$
$$X_t = \alpha X_{t-1} + \beta + a_t, \text{ with } |\alpha| < 1, \beta \neq 0 \tag{10c}$$
$$X_t = X_{t-1} + \beta + a_t, \text{ with } \beta \neq 0 \tag{10d}$$
$$X_t = \alpha X_{t-1} + \beta + \gamma t + a_t, \text{ with } |\alpha| < 1, \gamma \neq 0 \tag{10e}$$
$$X_t = X_{t-1} + \beta + \gamma t + a_t, \text{ with } \gamma \neq 0 \tag{10f}$$

The above models $(10b)$, $(10d)$ and $(10f)$ are stationary in differences, and the time series of these types are made stationary by differentiating. The models $(10c)$ and $(10e)$ are trend-stationary, and the time series according these models are made stationary by identification and removing trend (moving average, or exponential smoothing).

For the Dickey—Fuller test we group first into pairs the model $(10a)$ with the model $(10b)$, the model $(10c)$ with the model $(10d)$, and the model $(10e)$ with the model $(10f)$. We obtain

$$\Delta X_t = \Phi X_{t-1} + a_t \tag{11a}$$
$$\Delta X_t = \Phi X_{t-1} + \beta + a_t \tag{11b}$$
$$\Delta X_t = \Phi X_{t-1} + \beta + \gamma t + a_t \tag{11c}$$

In fact the Dickey—Fuller test contains three sub-tests. We test first the signification of the coefficients for the linear regression model $(11c)$, but for Φ the test must be left-sided: $H_0 : \Phi = 0$, and $H_1 : \Phi < 0$.

If after the first signification test it results that Φ is significant, it results that the right model from (10) is $(10e)$ if γ is significant (the autoregressive model with temporal trend), $(10c)$ if γ is not significant, but β is significant (the autoregressive model with drift), and, if the other two parameters are not significant, we accept the model $(10a)$ (the autoregressive model).

If in the first test Φ is not significant, we proceed to test the signification of the coefficients of the regression model $(11b)$. If Φ becomes significant, then we choose between the models $(10c)$ and $(10a)$, depending on the signification of β. Otherwise, we do the last test, namely the test for signification of Φ in the model $(11a)$.

If in the last test Φ is significant, we accept the model $(10a)$. Otherwise (if Φ is not significant in all the three tests), we accept the model $(10f)$ if γ was significant in the first test (random walk with drift and trend), $(10d)$ if γ was not significant in the first test, but β was significant in the second test (random walk with drift), respectively $(10b)$ if γ was not significant in the first test, and β was not significant in the second test (random walk).

We cannot use the Student test for the signification of Φ, β or γ. This, because if $\Phi = 0$ or $\gamma \neq 0$ X_t is not stationary, for any values of β, hence the common rules of statistical inference (particulary, the Student test) cannot be applied [3]. Dickey and Fuller have estimated by the Monte Carlo method the critical values (instead of the quantiles of Student distribution) with which we have to compare the computed Student statistics for Φ, in the cases of different sizes of time series.

For the qualitative explanatory variables, in [3] there is presented the problem of the dependence of income in terms of number of school years, for m groups. There are obtained the two linear regressions

$$Y = a_0^{(j)} + a_1 X, \tag{12}$$

where $a_0^{(j)}$ is the intercept for the group j.

Considering the dummy variables

$$D_j = \begin{cases} 1 \text{ for the group} j+1 \\ 0 \text{ otherwise} \end{cases}, \ j = \overline{1, m-1} \tag{13}$$

it is obtained the linear regression

$$Y = a_0^{(1)} + \sum_{j=1}^{m-1} \left(a_0^{(j+1)} - a_0^{(j)} \right) D_j + a_1 X. \tag{14}$$

In the same manner there are considered the seasonal data. In this case the number of groups is the number of seasons.

In the above cases the slope is common, and the intercept differs from a group to another. If the slope differs, we denote by $D_{j,0}$ the above dummy variables, and the other qualitative explanatory variables are

$$D_{j,1} = \begin{cases} X \text{ for the group} j+1 \\ 0 \text{ otherwise} \end{cases}, \ j = \overline{1, m-1}. \tag{15}$$

Finally we obtain the linear regression

$$Y = a_0^{(1)} + \sum_{j=1}^{m-1} \left(a_0^{(j+1)} - a_0^{(j)} \right) D_{j,0} + a_1^{(1)} X + \sum_{j=1}^{m-1} \left(a_1^{(j+1)} - a_1^{(j)} \right) D_{j,1}. \tag{16}$$

In [4] there are forecasted 17 economic variables by simulation of three scenarios for the period 2010-2014. The computational assumptions for the first one (base scenario) are the following:

1. A preasure on nominal revenues, either in the private or in the budgetary sector, remains significant. The index of expected disposable income ranges between 1.06 and 1.085.
2. After the elaboration of the 2005 version of the elaboration of the 2005 version of the macromodel, some factors infered and negatively influenced the global return of the Romanian economy. This impact was accentueted during the crisis. Therefore the equation for the total factor productivity, and for the unemployment rate were corrected for all the years of the economic crises. In the case of gross fixed capital formation the correction was for the first two years of the period.
3. The international financial crisis will pass into a moderate global recovery. The parameters concerning the world trade index in real terms and world trade deflator are considered as slowly ascending series, and the short term interest rate is constant.
4. It is expected that the capital flows will increase. This comes from portofolio investments or the net tranfers from abroad, and from a rising degree of absorption of the European structural and cohesion funds.
5. The general consolidated budget is conceived under stability of taxation. Therefore the ratio of direct taxes to GDP, the ratio of other budget revenues to GDP and the ratio of VAT to gross value added are constant.
6. The annual index of broad money (IM2) is projected to exceed slightly the similar index of expected disposable income (IYd), which allows a reduction in interest rates.
7. The rate of tangible fixed assets depreciation is mentained at constant level of 0.075.

For the second scenario (the worsened scenario W1Sc), which generally mentains the assumptions of base scenario, it assumes that the domestic situation (institutional reforms, fiscal systems, etc.) does not allow a significant improvement of the business environment. Consequently, in addition to the base scenario there are considered the following assumptions:

1. The capital inflows are more limited, and this concerns the foreign direct and portofolio investments, current account net transfers and structural European funds.
2. The relationship for total factor productivity is also penalized by slightly increase negative correction coefficients.
3. NBR policy remains able to mentain the exchange rate of RON in a narrow band of fluctuation.

The third scenario (the worsened scenario W2Sc) is derived from the previous one, but it tries to compress the inflation by more restrictive income, monetary and budget expenditure policies. The additional assumptions are as follows:

1. A slower increase in expected disposable income is taken into account.
2. The exogeneous coefficients regarding government transfers and other public expenditures are also reduced in comparison with the other two scenarios.
3. The broad money supply is projected at lower levels.

2. THE TEST FOR IDENTITY OF COEFFICIENTS OF QUALITATIVE VARIABLES

In this section we consider not only one set of coefficients Φ, β and γ in (11): we have gr groups and we consider a set of above mentioned coefficients for each group.

A test for identity of some expectation is the Tukey test [2]. Consider m independent samples having the distributions $N\left(\mu_i, \sigma^2\right)$, having the same size, n.

The Tukey test checks with the first degree error ε the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_m$ against the alternative hypothesis H_1: there exist $i \neq j$ such that $\mu_i \neq \mu_j$.

Consider an unbiased estimator of σ^2 based uppon r degrees of freedom, and we denote it by S^2. We compute the statistics

$$q = \frac{\overline{X}_{\max} - \overline{X}_{\min}}{S \cdot \sqrt{\frac{2}{n}}}, \tag{17}$$

where $\overline{X}_{\max}$ and $\overline{X}_{\min}$ are the maximum, respectively minimum expectation of the above m samples.

It is proved [2] that the q has the Student distribution with r degrees of freedom. Therefore we accept the null hypothesis if and only if $q < t_{r;\frac{\varepsilon}{2}}$, where $t_{r;\frac{\varepsilon}{2}}$ is the quantile of the error $\frac{\varepsilon}{2}$ of the Student distribution with r degrees of freedom.

► Remark ([2]). The denominator from (17), $S\cdot\sqrt{\frac{2}{n}}$, is in fact the estimator of the standard deviation of the numerator, $\overline{X}_{\max}-\overline{X}_{\min}$, with r degrees of freedom. Therefore the Tukey test can be performed also in the case of different variances σ_i^2. It is enough to consider the same degrees of freedom, r, and the statistics becomes $q=\max\limits_{i,j=\overline{1,m}}\frac{\left|\overline{X}_i-\overline{X}_j\right|}{\sqrt{S_i^2+S_j^2}}$, where $\overline{X}_i$ and S_i^2 are the estimators of the expectation and of the variance of the component i, the last one being computed with r degrees of freedom.

If we want to check if some regression coefficients are equal, with given first degree error ε, we consider the formula for the variance-covariance matrix of the vector of coefficients, $\widehat{A}$ [3]:

$$Var\left(\widehat{A}\right)=\sigma_u^2\left(X'X\right)^{-1}, \tag{18}$$

where σ_u^2 is the estimator of the variance of errors. The number of degrees of freedom (for residues and coefficients) is $n-k-1$, where n is the size of data and k is the number of explanatory variables. Therefore the Tukey $q-$statistics becomes

$$q=\max_{i,j=\overline{1,m}}\frac{\left|\overline{X}_i-\overline{X}_j\right|}{\sqrt{S_i^2+S_j^2-2\cdot C_{i,j}}}, \tag{19}$$

where $\overline{X}_i$ and S_i^2 are the estimators of the expectation and of the variance of the coefficient A_i, and $C_{i,j}$ is the covariance of the coefficients A_i and A_j. Of course, the above maximum range only for the pairs (i,j) such that, according to null hypothesis, we have $A_i=A_j$, and the number of degrees of freedom is also $n-k-1$.

Therefore for common regression coefficients we compare the above $q-$statistics with the quantile $t_{n-k-1;\frac{\varepsilon}{2}}$. We accept the null hypothesis of identical coefficients if and only if $q<t_{n-k-1;\frac{\varepsilon}{2}}$. This test can be performed not only to check if one group of coefficients has a single value. We can check for instance if the coefficients of X_1 and X_2 are identical, and in the same time the coefficients of X_3 and X_4 are identical, but the coefficients of X_1 and X_3 are not necessary identical.

The regression coefficients can be considered also for qualitative/ dummy explanatory variables. The conditions that have to be fulfilled are the mutual independence of Y_t and of X_{it}. Therefore in the time series case we cannot use the Student distribution for testing the identity of coefficients, for the same reasons we cannot use it for unit root tests.

More exactly, consider the equation (11). The set of parameters (Φ,β,γ) is replaced by gr sets $(\Phi_i,\beta_i,\gamma_i)_{i=\overline{1,gr}}$ corresponding to gr groups. The qualitative variables are

$$\begin{pmatrix} X_{t-1;i} \\ D_i \\ t_i \end{pmatrix}=\begin{pmatrix} X_{t-1} \\ 1 \\ t \end{pmatrix} \tag{20}$$

for the group i, and the corresponding set of coefficients is $\begin{pmatrix} \Phi_i \\ \beta_i \\ \gamma_i \end{pmatrix}$.

The gr groups are built taking into account the time period (the moment belongs to the economic crisis or not, or, for trimestrial or monthly data, to a given trimester or month).

For each test from the Dikey—Fuller methodology mentioned in introduction, each involved signification test is preceded by homogeneity tests as follows:

1. First we test the total homogeneity: the involved parameter has the same value for all groups.
2. If the total homogeneity fails, we remove a component using the minmax criterion: if we remove a component, the corresponding statistics for identity of the retained coefficients is minimum.
3. If for a partial homogeneity test we accept the null hypothesis, we stop, considering the retained coefficients having the same value. Otherwise, we continue with the above minmax technique, until it remains only one coefficient, or we accept the identity for some coefficients.

Because we cannot use the Student quantile, we generate 1000 sets of parameters such that each of them is uniform in an interval containing zero: if the alternative is that the parameter is less than zero (as for Φ),

the interval is $(-1,0)$. If the alternative is $\neq 0$ (as for γ and β), the interval is $(-1,1)$. We generate also the variance of e_t in the interval $(0,1)$. Of course, for identity between some parameters we do not generate all the coefficients: we generate only one coefficient for each group of equal coefficients. For each set of parameters we generate 10000 such models.

We compute for each model the $q-$statistics, we order the 10000000 $q-$statistics. Because we use also the absolute value, the quantile is the value from the position $10000000\,(1-\varepsilon)$ instead those from the position $10000000\left(1-\frac{\varepsilon}{2}\right)$.

The parameters for each of the above models are generated uniform on the interval $(-1,1)$ for β and γ coefficients, on the interval $(-1,0)$ for Φ coefficients, respectively on the interval $(0,1)$ for the variance of the errors. The errors are generated as normal variables wis the expectation zero and the variance generated before. The methods to generate the above random variables, and methods to solve optimization problems are presented in [6]. From the methods to generate normal variables presented in the above book, we choose the Box—Muler method, because it is the most rapid.

For signification we use the standard Dickey—Fuller test if after the homogeneity test we conclude that we have only one group for all coefficients. Otherwise we estimate the quantiles by simulation, and we use two-sided tests. Even for Φ, due to the existence of several groups, we can have positive values.

3. APPLICATION

Consider the yearly data of GDP in the period 1990-2011. The data are from [7]. The three periods are 1990-2000, 2001-2007 and 2008-2011 (the economic crisis) inclusive.

In the case of pure data we obtain first, using our $C++$ program, the regression

$$\begin{aligned}\Delta X_t &= 20.79444D_1 - 108.38084D_2 - 269.04604D_3 - 0.84321\widetilde{X}_{t-1;1} - 0.22976\widetilde{X}_{t-1;2} - 0.55302\widetilde{X}_{t-1;3} + \\ &\quad 1.28323\widetilde{t}_1 + 10.08695\widetilde{t}_2 + 17.93309\widetilde{t}_3.\end{aligned}$$

The variance of the residues is $\sigma_u^2 = 24.05647$, and the $q-$statistics using the mentioned minmax technique are 2.51044 (obtained for the first two periods, years 1990—2008) and 1.02822 (obtained for the last two periods, years 2001—2011) for γ, 1.59709 (obtained for the first two periods) and 0.62695 (obtained for the first and the last period, years 1990—2001 and 2008—2011) for Φ, respectively 3.23053 (obtained for the first two periods) and 0.85615 (obtained for the the last two periods) for β.

We order the above q-statistics, and we obtain the following sequence of tests:

1. $\Phi_i = \Phi$, $\gamma_i = \gamma$ and $\beta_i = \beta$.
2. $\Phi_i = \Phi$, $\gamma_i = \gamma$ and $\beta_2 = \beta_3$.
3. $\Phi_i = \Phi$, $\gamma_2 = \gamma_3$ and $\beta_2 = \beta_3$.
4. $\gamma_2 = \gamma_3$, $\beta_2 = \beta_3$ and $\Phi_1 = \Phi_3$.
5. Possible different γ_i, $\beta_2 = \beta_3$ and $\Phi_1 = \Phi_3$.
6. Possible different γ_i and β_i, and $\Phi_1 = \Phi_3$.

Comparing to the quantiles from Table 1, we accept the null hypothesis in the case of the first test, with the threshold of 5%[1]. Therefore we do not proceed to do the other five tests.

Table 1: The quantiles for the homogeneity tests in the case of first degree error being 10%, 5%, 2.5%, respectively 1%.

Model	Test	Quantiles			
		10%	5%	2.5%	1%
III	$\Phi_i = \Phi$, $\beta_i = \beta$ and $\gamma_i = \gamma$	3.25943	3.82568	4.39094	5.17224
II	$\beta_i = \beta$ and $\Phi_i = \Phi$	2.35455	2.71764	3.06432	3.49962
II	$\beta_i = \beta$ and $\Phi_1 = \Phi_2$	2.35467	2.72891	3.08986	3.55251
II	$\beta_1 = \beta_2$ and $\Phi_1 = \Phi_2$	2.02454	2.40515	2.76983	3.22623
I	$\Phi_i = \Phi$	1.93702	2.27258	2.58109	2.97614
I	$\Phi_1 = \Phi_3$	1.49849	1.83887	2.1422	2.53485

[1] The statistics 3.23053 is significant neither for 10%, because the quantile is in this case 3.25943

For the equation $(11b)$ we obtain the regression

$$\Delta X_t = 13.85923D_1 + 3.43493D_2 + 209.915D_3 - 0.40747\widetilde{X}_{t-1;1} + 0.22203\widetilde{X}_{t-1;2} - 1.24332\widetilde{X}_{t-1;3}.$$

The variance of residues is $\sigma_u^2 = 52.55795$, and the lists of q-statistics is 6.15478 (obtained for the last two periods) and 1.43677 (obtained for the first two periods) for Φ, respectively 5.07655 (obtained for the last two periods) and 0.65238 (obtained for the first two periods) for β.

We order the above q-statistics, and we obtain the following sequence of tests:

1. $\beta_i = \beta$ and $\Phi_i = \Phi$.
2. $\beta_i = \beta$ and $\Phi_1 = \Phi_2$.
3. $\Phi_1 = \Phi_2$ and $\beta_1 = \beta_2$.
4. Possible different Φ_i, and $\beta_1 = \beta_2$.

In the case of the first test we reject the null hypothesis for 5%, because $6.15478 > 2.71764$. The same thing we can say about the second test, because $5.07655 > 2.72891$. We notice that the above statistics are also significant for 1%.

In the case of the third test, we accept the null hypothesis for 5%, because $1.43677 < 2.40515$. The statistics is significant neither for 10%.

For the equation $(11a)$ we obtain the regression

$$\Delta X_t = -0.004\widetilde{X}_{t-1;1} + 0.25366\widetilde{X}_{t-1;2} - 0.05185\widetilde{X}_{t-1;3}.$$

The variance of residues is $\sigma_u^2 = 126.32111$, and the list of q-statistics is 5.34024 (obtained for the last two periods) and 0.44964 (obtained for the first and the last period).

We test first if all the values of Φ_i are identical, and we reject the null hypothesis for 5%, because $5.34024 > 2.27258$, and the statistics is also significant for 1%.

Next we test first if $\Phi_1 = \Phi_3$, and we accept the null hypothesis for 5%, because $0.44964 < 1.83887$. The statistics is neither significant for 10%.

In the case of logarithmic data, we obtain first the regression

$$\begin{aligned}\Delta X_t &= 2.55954D_1 + 1.07965D_2 + 0.59702D_3 - 0.79454\widetilde{X}_{t-1;1} - 1.34646\widetilde{X}_{t-1;2} - 0.53005\widetilde{X}_{t-1;3} + \\ &\quad 0.03795\widetilde{t_1} + 0.33396\widetilde{t_2} + 0.105\widetilde{t_3}.\end{aligned}$$

The variance of the residues is $\sigma_u^2 = 0.00582$, and the list of $q-$statistics is 1.6956 (obtained for the first two periods) and 0.6333 (obtained for the first and the last period) for γ, 0.74806 (obtained for the first two periods) and 0.30465 (obtained for the first and the last period) for Φ, respectively 1.82297 (obtained for the first two periods) and 0.07644 (obtained for the last two periods) for β.

We order the above q-statistics, and we obtain the following sequence of tests:

1. $\Phi_i = \Phi$, $\gamma_i = \gamma$ and $\beta_i = \beta$.
2. $\Phi_i = \Phi$, $\gamma_i = \gamma$ and $\beta_2 = \beta_3$.
3. $\Phi_i = \Phi$, $\gamma_1 = \gamma_3$ and $\beta_2 = \beta_3$.
4. $\gamma_1 = \gamma_3$, $\Phi_1 = \Phi_3$ and $\beta_2 = \beta_3$.
5. Possible different γ_i, $\Phi_1 = \Phi_3$ and $\beta_2 = \beta_3$.
6. Possible different γ_i and Φ_i, and $\beta_2 = \beta_3$.

Because the statistics 1.82297 is less than the same quantile of 5% from the case of pure data, we accept also the null hypothesis of total homogeneity. We accept also the null hypothesis for the threshold of 10%, as for pure data.

For the equation $(11b)$ we obtain the regression

$$\Delta X_t = 1.30746D_1 + 0.13816D_2 + 6.445D_3 - 0.37205\widetilde{X}_{t-1;1} + 0.02095\widetilde{X}_{t-1;2} - 1.25707\widetilde{X}_{t-1;3}.$$

The variance of residues is $\sigma_u^2 = 0.01384$, and the lists of q-statistics is 1.88925 (obtained for the last two periods) and 1.25023 (obtained for the first and the third period) for Φ, respectively 1.81271 (obtained for the last two periods) and 1.30958 (obtained for the first two periods) for β.

We order the above q-statistics, and we obtain the following sequence of tests:

1. $\beta_i = \beta$ and $\Phi_i = \Phi$.
2. $\beta_i = \beta$ and $\Phi_1 = \Phi_3$.
3. $\beta_1 = \beta_2$ and $\Phi_1 = \Phi_3$.
4. Possible different β_i, and $\Phi_1 = \Phi_3$.

For the first test we accept the null hypothesis for 5%, and the statistics of 1.88925 is neither significant for 10%.

For the equation $(11a)$ we obtain the regression

$$\Delta X_t = 0.00051\widetilde{X}_{t-1;1} + 0.0523\widetilde{X}_{t-1;2} - 0.00739\widetilde{X}_{t-1;3}.$$

The variance of residues is $\sigma_u^2 = 0.01636$, and the list of q-statistics is 3.32351 (obtained for the first two periods) and 0.43747 (obtained for the first and the last period).

For this model we perform the same test and we have the same conclusions and significance levels as in the case of pure data.

In the following we will test the signification of coefficients considering the resulting homogeneity. In the case of pure data, we test first the signification of the model

$$\Delta X_t = \beta + \Phi X_{t-1} + \gamma t.$$

We obtain

$$\Delta X_t = -6.20162 - 0.20553X_{t-1} + 2.51461t,$$

and the variance of residues is 246.8959. The Dickey—Fuller statistics are -0.085175 for β, -1.73753 for Φ, and 2.34101 for γ.

In this case we perform the standard Dickey—Fuller test, model $(11c)$. It results that Φ is not significant for 5%, because $-1.73753 > -3.6$ for $n = 25$, and the threshold decrease with n. The same thing we can say about the threshold for 10% and $n = 25$, -3.24.

Next we test the signification of parameters for the model

$$\Delta X_t = \beta_{1,2}D_{1,2} + \beta_3 D_3 + \Phi_{1,2}\widetilde{X}_{t-1;1,2} + \Phi_3\widetilde{X}_{t-1;3}.$$

We obtain

$$\Delta X_t = -7.99486D_{1,2} + 209.915D_3 + 0.31148\widetilde{X}_{t-1;1,2} - 1.24332\widetilde{X}_{t-1;3},$$

and the variance of residues is 69.76394. The statistics are -2.31095 for $\beta_{1,2}$, 4.53263 for β_3, 5.96962 for $\Phi_{1,2}$, and -2.86293 for Φ_3.

For the right-sided signification of $\Phi_{1,2}$ we have to compare the statistics 5.96962 with the 5% quantile, which is 0.87408. It results that $\Phi_{1,2}$ is right-significant, hence the model is exploding for the period before crisis. The statistics is significant also for 1%, when the quantile is 1.6717. For Φ_3, we compare the statistics of -2.86293 with the 5% threshold, -1.54511. The statistics is also significant for 1%. Therefore the series is exploding before crisis, and stationary during it.

The above quantiles are listed in Table 2. We do not need now to check the significance of β coefficients, but we can conclude, using the two-sided thresholds from Table 3, that $\beta_{1,2}$ is significant for 10%, but it is not for at most 5% error. β_3 results to be significant, even for 1%.

Table 2: The quantiles for the one-sided signification tests for Φ in the case of first degree error being 10%, 5%, 2.5%, respectively 1%.

Model	Test	Quantiles							
		10%		5%		2.5%		1%	
		Left-sided	Right-sided	Left-sided	Right-sided	Left-sided	Right-sided	Left-sided	Right-sided
II	$\Phi_1 = \Phi_2 = 0$, $\beta_1 = \beta_2$	-2.30614	0.47384	-2.7102	0.87408	-3.07123	1.22858	-3.51806	1.6717
II	$\Phi_1 = \Phi_2$, $\beta_1 = \beta_2$, $\Phi_3 = 0$	-1.18335	0.62469	-1.54511	1.04062	-1.87267	1.46114	-2.29839	2.14064
I	$\Phi_1 = \Phi_3 = 0$	-1.5439	1.13348	-1.92707	1.55075	-2.27791	1.93551	-2.69031	2.42289
I	$\Phi_1 = \Phi_3$ $\Phi_2 = 0$	-1.4381	1.18241	-1.782	1.60629	-2.09037	1.996	-2.46236	2.51302

Table 3: The quantiles for the two-sided signification tests for β in the case of first degree error being 10%, 5%, 2.5%, respectively 1%, Model II.

Test	Quantiles			
	10%	5%	2.5%	1%
$\beta_1 = \beta_2 = 0, \Phi_1 = \Phi_2$	2.02645	2.40952	2.76976	3.21575
$\beta_1 = \beta_2, \Phi_1 = \Phi_2, \beta_3 = 0$	1.83231	2.20667	2.54475	2.9877

Finally, we test the signification of parameters for the model

$$\Delta X_t = \Phi_{1,3}\widetilde{X}_{t-1;1,3} + \Phi_2\widetilde{X}_{t-1;2}.$$

We obtain

$$\Delta X_t = -0.04605\widetilde{X}_{t-1;1,3} + 0.25366\widetilde{X}_{t-1;2},$$

and the variance of residues is 119.69177. The statistics are -1.36267 for $\Phi_{1,3}$, and 5.97619 for Φ_2.

Comparing to the 5% thresholds, we conclude that $\Phi_{1,3}$ is not significant, even for 10%, and Φ_2 is significant even for 1%. Therefore the GDP series is random walk for the periods 1990—2001 and 2008—2011, and exploding during the economic increasing period, 2001—2008.

In the case of logarithmic data, we test first the signification of the model

$$\Delta X_t = \beta + \Phi X_{t-1} + \gamma t.$$

We obtain

$$\Delta X_t = 0.82185 - 0.29444 X_{t-1} + 0.03963 t,$$

and the variance of residues is 0.01863. The Dickey—Fuller statistics are 2.60596 for β, -2.76856 for Φ, and 3.34774 for γ. We have again $-2.76856 > -3.24$, hence Φ is significant neither for 10%.

Next we test the signification of parameters for the model

$$\Delta X_t = \beta + \Phi X_{t-1}.$$

We obtain

$$\Delta X_t = -0.04464 + 0.02941 X_{t-1},$$

and the variance of residues is 0.02864. The statistics are -0.19983 for β, and 0.53683 for Φ.

Because $\Phi > 0$, it results that it is not significant from the Dickey—Fuller test point of view.

Finally, we test the signification of parameters for the model

$$\Delta X_t = \Phi_{1,3}\widetilde{X}_{t-1;1,3} + \Phi_2\widetilde{X}_{t-1;2}.$$

We obtain

$$\Delta X_t = -0.00242\widetilde{X}_{t-1;1,3} + 0.0523\widetilde{X}_{t-1;2},$$

and the variance of residues is 0.01565. The statistics are -0.28404 for $\Phi_{1,3}$, and 4.84361 for Φ_2.

Comparing to the 5% thresholds, we rich to the same conclusions as in the case of the pure data.

4. CONCLUSIONS

In this paper we have study the way we can group data only from the point of view of stationarizing time series. Two groups that have identical coefficients can be stationarized together, using the same scheme. An open problem is to extend the study for stationary data. More exactly, to test if two groups have the same AR and/ or MA coefficients, and/ or the same variance of white noise.

After we will make the groups after the homogeneity tests, considering also the ARMA structure and the variances of the white noises, we can build the scenarios of forecast depending on the group such that the future value X_{n+1} belongs to.

We notice that the logarithmic data are more homogeneous than the pure data. The explanation could be that the differences between values decrease if we apply logarithms. Moreover, for instance an exploding time serries becomes random walk by logarithm.

For only one sequential criterion to group the time moments we have made copies for the common years 2001 and 2008. The same thing we can do for several sequential criteria: we make only one sequential criterion, considering all the separation years from the considered criteria.

An open problem is to study the homogeneity for one or more seasonal criteria. If it is one criterion, we change the signification of groups. For instance, if we consider trimestrial data, T1 has the following new signification: X_t is in $T1$, and X_{t-1} is in $T4$, and so on. If we have several periodic criteria, we make only one, with one period equal to the highest common factor of the periods.

More difficult is the case when we have several criteria sequential and periodical. Of course, as we have mentioned above, we can reduce the problem to the case of two criteria: one sequential, and one periodical. This reduced case is also an open problem.

For the standard significance level of 5% we notice that in the case of the model $(11b)$ we accept identical coefficients for the two periods before the economic crisis. Therfore the economic crisis is separated. In the case of the model $(11a)$ we have another separation: we accept identical Φ coefficients for the first and last periods, and the separated period is those from the middle (2001—2008), of the economic increase.

The identity of coefficients for two periods (first two in the case of model II, first and third in the case of model I) does not mean that we have the same time series. It means that we can use the same stationarising method (differences). The obtained stationary time series can be different.

REFERENCES

[**1**] Brockwell, P.J. & Davis, R.A. (2002). Springer Texts in Statistics; Introduction to Time Series and Forecasting. Springer-Verlag.

[**2**] Ciucu, G. & Craiu, V. (1974). Inferenţă statistică. Bucharest: Ed. Didactică şi Pedagogică (English: Statistical Inference).

[**3**] Jula, D. (2003). Introducere în econometrie. Bucharest: Professional Consulting (English: Introduction to Econometrics).

[**4**] Dobrescu, E. (2010). Macromodel Simulations for the Romanian Economy. Romanian Journal of Economic Forecasting, XIII(2), 7-28.

[**5**] Popescu, Th. (2000). Serii de timp; Aplicaţii la analiza sistemelor. Bucharest: Editura Tehnică (English: Time Series; Aplications to Systems' Analysys).

[**6**] Văduva, I. (2004). Modele de simulare. Bucharest University Printing House.

[**7**] TEMPO-Online database of National Institute of Statistics. Aviable at $<$ *www.insse.ro* $>$ [Accessed February 22 2013].