Econometric and Neural Network Analysis of the Labor Productivity and Average Gross Earnings Indices in the Romanian Industry

MADALINA ECATERINA ANDREICA, NICOLAE CATANICIU
National Scientific Research Institute for Labor and Social Protection
6-8 Povernei Str., District 1, Bucharest
ROMANIA
madalina.andreica@gmail.com, ncataniciu@incsmps.ro

MUGUREL IONUT ANDREICA
Politehnica University of Bucharest
Splaiul Independentei 313, sector 6, Bucharest
ROMANIA
mugurel.andreica@cs.pub.ro

Abstract: - This paper focuses on the identification and short term forecast of the correlation between the Labor Productivity Index (LPI) and the Average Gross Earnings Index (AGEI) in the Romanian Industry. The tools and models that were used consist of several lag econometric models, ARIMA processes, as well as feed forward neural networks. The results proved that the models were suitable and showed a connection between AGEI and LPI.

Key-Words: - labor productivity, econometric model, ARIMA, VAR, neural network, forecast

1 Introduction

In this paper we present the results of a study whose purpose was to identify the correlation between the monthly Labor Productivity and the Average Gross Earnings in the Romanian Industry during the period 2002-2008. Several models were therefore tested, by using both some econometric and neural network methodologies. Regarding the econometric tools used in this study, we estimated the impact of Labor Productivity upon the Average Gross Earnings through different lag models and applied the Box Jenkins methodology to short term predict and to test the efficiency of those models. Further on, we assumed that our variables of interest might also be interrelated and estimated a VAR model with the purpose of analyzing the dynamic impacts between Labor Productivity and Average Gross Earnings in the Romanian Industry. From this analysis some important conclusions were noticed. We then performed a similar analysis using a feed-forward neural network, with which we were able to accurately forecast the values of the two considered indicators on a longer term.

There are several recent papers concerning the dynamic movements of the main branches of the Romanian Industry [3] on the labor market. Some have estimated the impact of the foreign direct investments upon the Romanian Labor Productivity both with econometric and neural network tools [4]. Several other papers studied other macroeconomic indicators using models and tools similar to ours [5,6,7] or could benefit from them [2]. However, no work that we know of, studied the connection between the Labor Productivity and the Average Gross Earnings on the Romanian Industry, despite the quite obvious relation between them, which can be clearly explained from an economic perspective. A growth in earnings is normally expected when labor productivity increases. Moreover, a higher increase in labor productivity than in earnings describes a healthy image in one country's economy. That is why, being able to estimate and to predict the dynamics of these two macroeconomic indicators can become a real advantage in the governmental economic policy and in any risk investing decision problem [1], since it may give a better view of the macroeconomic movements that might occur on the market and also help the investors better predict their cash flows. And just like one of the many economic feedbacks, the higher investments they get, the more it leads to a higher labor productivity, which makes the economy prosper even more.

The paper is organized as follows: Section 2 describes the data used for this study, Section 3 presents the econometric framework and the main neural network concepts used for this study, while Section 4 presents and compares the estimated results of the several methods tested. Section 5 concludes.

2 Data description

We separately estimated the correlation between the Labor Productivity Index (LPI) and the Average Gross

Earnings Index (AGEI) for total Industry, as well as for each of the three Romanian Industrial sections: *Mining and quarrying (mining), Electric and thermal energy, gas and water (energy)* and *Manufacturing (manufact)*.

For this study, monthly (seasonally adjusted) data were used, starting with June 2002 and ending with July 2008. The reason for choosing this period was based on data availability, since there were no available data for the LPI before May 2002. For this particular reason, both indices were calculated with the base in May 2002. The AGEI was also deflated by using the monthly CPI. The main data sources were the *Romanian Statistical Yearbook* and the *Monthly Statistical Bulletins*, both published by the Romanian National Institute of Statistics.

By LPI in Industry we understand the indicator that characterizes the efficiency of work form a certain period of time in the industrial activity. It is therefore calculated as a ratio between the gross, industrial production index and the industry average number of employees' index. On the other hand AGEI is an index that comprises salaries, respectively money rights for the work which was effectively performed. It also includes different benefits and indemnities granted, as well as other legal rises of salary, amounts paid for the non-worked time and bonuses.

3 Models and Methodologies

Seeking for the proper model to estimate the relation between the two targeted variables - LPI and AGEI, we made use of several econometric and neural network tools, which will be further on presented in this section.

3.1 Regression models

The simplest econometric method for describing certain relationships based on economic theory among our variables of interest consists of estimating a univariate regression model. However, in order to be able to make use of the econometric results, the residuals should pass the following important conditions:

- The residuals should not be correlated
- The residuals should have a normal distribution
- The residuals should be homoskedastic

In case the residuals are serially correlated or do not pass all of the above tests, the estimated coefficients will be biased and inconsistent and the equation should be re-specified before using any of the estimated results.

Another fact that should be taken into consideration in our econometric analysis consists of admitting that the variables may influence each other with some delay in time. That is why the need of including lags in the econometric model should be as well attentively tested.

3.2 ARIMA model

A more complex technique that is used for describing the behavior of one's series and to short term forecast its dynamics, is the Autoregressive Integrated Moving Average model, notated as ARIMA(p,d,q). The general model was introduced by Box and Jenkins and is a method which allows both autoregressive and moving average parameters to be included. Moreover, it explicitly includes differencing in the formulation of the model. *p* is the autoregressive parameter, *d* is the number of differencing needed so that the series become stationary, while *q* represents the moving average parameter. The general form of an autoregressive AR(p) model is:

$$Y_t = \alpha + \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + \ldots + \rho_p Y_{t-p} + \varepsilon_t \quad (1)$$

while a moving average MA(q) is described by:

$$Y_t = \varepsilon_t + \vartheta_1 \varepsilon_{t-1} + \vartheta_2 \varepsilon_{t-2} + \ldots + \vartheta_q \varepsilon_{t-q} \quad (2)$$

where Y_t is a stationary series, Y_{t-i} represents lag *i* of Y_t, ε_t is a random walk of 0 mean and σ^2 variance and $\rho_{1,..}$ $\rho_p, \theta_{1,..}\theta_p$ are the parameters to be estimated. When considering an ARMA(p,q) process, there is the following representation:

$$Y_t = \alpha + \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + \ldots + \rho_p Y_{t-p} + \varepsilon_t + \vartheta_1 \varepsilon_{t-1} + \vartheta_2 \varepsilon_{t-2} + \ldots + \vartheta_q \varepsilon_{t-q} \quad (3)$$

The Box Jenkins methodology consists of the next steps:

- Checking if the series is stationary, meaning constant mean, variance and autocorrelation in time, by using ADF test and by analyzing the correlogram. In case of non-stationarity, *d* differences will be applied until the series becomes stationary.
- Identifying the possible ARMA(p,q) processes, based on the correlogram and of the ACF and PACF functions.
- Testing the validity of the selected processes and deciding upon the process (if any) that best describes the behavior of the series, based on R^2 values or on information criterions such as Akaike or Schwarz.

3.3 VAR model

One further step consists of accepting the possibility that the two variables might be interrelated. That means, that both variables might influence each other through several *p* lag periods, and therefore each variable should be described by a separate equation. The mathematical representation for a VAR (2) is:

$LPI_t = a_{11}LPI_{t-1} + a_{12}AGEI_{t-1} + b_{11}LPI_{t-2} + b_{12}AGEI_{t-2} + c_1 + \varepsilon_{1t}$

$AGEI_t = a_{21}LPI_{t-1} + a_{22}AGEI_{t-1} + b_{21}LPI_{t-2} + b_{22}AGEI_{t-2} + c_2 + \varepsilon_{2t}$

where ε_{1t} and ε_{2t} are the innovations that may be contemporaneously correlated but are uncorrelated with their own lagged values and uncorrelated with all of the right-hand side variables.

3.4 Neural network concepts

Neural networks are commonly used in order to find correlations between various time series and in order to produce future estimates. For our analysis we considered only feed forward neural networks. These networks were used in two distinct ways. First, we used them in order to

obtain a model which is similar to the VAR model, except that it is not necessarily linear. Let's assume that *LAG* is the number of lag periods. We have:

$LPI_t=f_{LPI}(LPI_{t-1}, \ldots, LPI_{t-LAG}, AGEI_{t-1}, \ldots, AGEI_{t-LAG})$ (4)

$AGEI_t=f_{AGEI}(LPI_{t-1}, \ldots, LPI_{t-LAG}, AGEI_{t-1}, \ldots, AGEI_{t-LAG})$

f_{LPI} and f_{AGEI} are two unknown functions which have to be learnt by the feed forward neural network. While in VAR f_{LPI} and f_{AGEI} are linear functions, there are many situations where these functions are not linear and we do not know their structure in advance. A feed forward neural network is able to learn any function which has only a finite number of discontinuities. The structure of a feed forward neural network is the following: an input layer, one or more hidden layers and one output layer. Every neuron x on a layer L is connected to all the neurons y on the next layer L+1. A (directed) connection between two neurons x and y has a weight w(x,y) and every neuron has an activation function (e.g. tansig for the hidden layer and linear for the output layer). The network starts with arbitrary weights and modifies them during the training stage, in order to minimize the error function (the difference between the output of the network and the desired output). A non-linear optimization method like the gradient descent technique is used to adjust the weights. In our experiments, we used only one layer of hidden neurons (see Fig.1). The network is trained in order to learn the next value LPI_t (or $AGEI_t$), given the previous *LAG* values of both indices. Thus, its purpose is to also produce future estimates, based on its previous ones. That is, its outputs LPI_t and $AGEI_t$ will be used as inputs in order to estimate $LPI_{t'}$ and $AGEI_{t'}$ $(t+1 \leq t' \leq t+LAG)$.

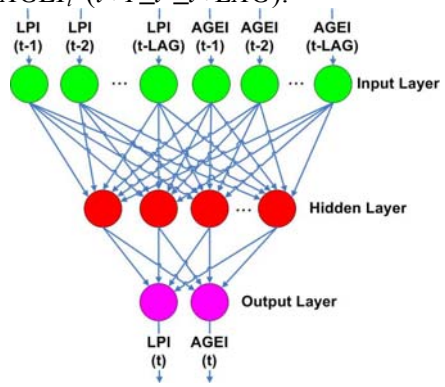

Fig.1 A feed forward neural network.

Secondly, we used the neural network in order to find a correlation between the *LPI* values corresponding to the industry sections and the *AGEI* value corresponding to the total industry. If we denote by *LAG* the number of lag periods again, the model is as follows:

$AGEI_{total,t}=g_{AGEI}(LPI_{mining,t-LAG}, \ldots, LPI_{mining,t}, LPI_{energy,t-LAG}, \ldots, LPI_{energy,t}, LPI_{manufact,t-LAG}, \ldots, LPI_{manufact,t})$ (5)

g_{AGEI} is an unknown composition function which has to be estimated by the feed forward neural network. If the indices of the three industry branches are independently forecasted, then we can use the model in order to produce forecasts of the total industry AGEI.

4 Experimental Results

Since the econometric and the neural network models have different requirements on the characteristics of the series, we studied the following four cases for the three sections and for the total industry: (1) the series in level; (2) the first differenced series; (3) the natural log series; (4) the first differenced natural log series. Section 4.1 shows the econometric results, while section 4.2 tests the correlation from a neural network approach.

4.1 The econometric results

After first testing the series in level as well as in first difference, no valid econometric model was found, mainly because of heteroskedasticity issues. Thus, the following results were obtained for the LPI and AGEI log series (the series of natural logarithms of the values of the series in level). Regarding the stationarity issue, both the correlograms and the ADF Unit Root Test indicated that all the LPI and AGEI log series were first order integrated, meaning that the series become stationary after applying the first difference.

4.1.1 Lag models

When trying to estimate the univariate regression model we encountered the problem of residuals correlation. In order to solve it we had to include some lag variables in the model. The resulted models when using log series are described below. Notice that AGEI was considered the dependent variable, while LPI the independent one:

The case of Electric and thermal energy, gas and water:

$lnAGEI\_energy(t) = \mathbf{-0.01} + \mathbf{0.09}*lnLPI\_energy(t-7) + \mathbf{0.36}*lnAGEI\_energy(t-1) + \mathbf{0.45}*lnAGEI\_energy(t-2) + \mathbf{0.21}*lnAGEI\_energy(t-6)$ (6)

The model was then validated. The residuals passed the normality Jarque Bera test with a high probability of 82% and the White Heteroskedasticity Test with 67%, while the LM Test indicated the absence of any autocorrelation for the first 8 lags with a probability of 59%. From the model specification, we notice that the only significant lags of the two variables upon AGEI log series are its own first, second and the 6-th lag as well as the 7-th LPI log series lag.

In order to see how efficient the model is, we re-estimated the AGEI series, by using eq. (6) and the new predicted LPI values obtained after applying the Box Jenkins methodology. The best ARIMA process considered suitable for the LPI in this Industrial section was an **ARIMA(4,1,4).** Further on, we only forecasted a 5 months long period due to the difficulty of long term prediction when using only 73 observations. Finally, we plotted both the real and the estimated AGEI series in Fig.2 and calculated an average error of 0.019 (the ratio between the sum of the absolute values of the errors and

the number of estimated values; the error of an estimated value is the difference between that value and the real value).

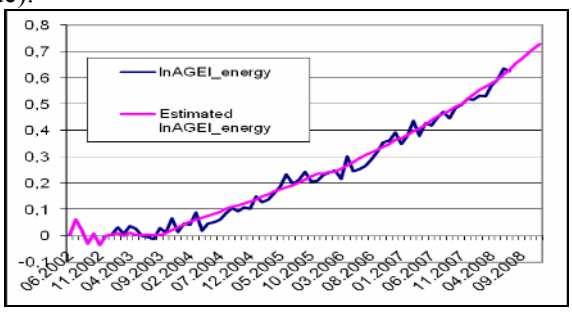

Fig.2 Real and estimated *lnAGEI_energy* series

The case of Mining and quarrying:

$$lnAGEI\_mining(t) = -0.03 + 0.92*lnLPI\_mining(t-2) - 0.31*lnLPI\_mining(t-4) + 0.4*lnAGEI\_mining(t-6) \quad (7)$$

The residuals passed the Jarque Bera test with a high probability of 82% and the White Heteroskedasticity Test with 56%, while the LM Test indicated the absence of any autocorrelation for the first 8 lags with a high probability of 95%. The lags identified as being significant in the behavior of AGEI in the Mining and quarrying Section were lags 2 and 4 of LPI as well as lag 6 of AGEI. Moreover, by the estimated values of the LPI coefficients we conclude that LPI has a quite meaningful impact upon AGEI behavior.

When applying the Box Jenkins procedure for the LPI series, it resulted an **ARIMA(2,1,2)** process. After short term prediction of LPI, the values were introduced in eq. (7), and the estimated AGEI series were calculated. For comparison, we plotted both the real and estimated AGEI values in Fig.3 and calculated the average error of the model of 0.04.

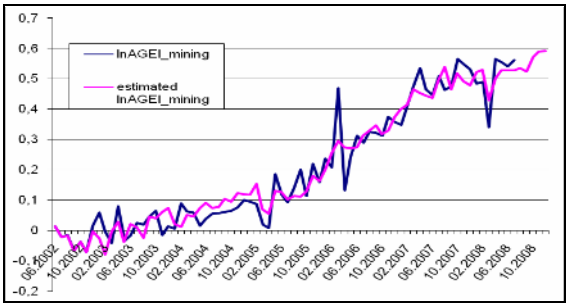

Fig.3 Real and estimated *lnAGEI_mining* series

The case of Manufacturing:

$$lnAGEI\_manufact(t) = -0.007 + 0.32*lnLPI\_manufact(t) - 0.17*lnLPI\_manufact(t-4) + 0.52*lnAGEI\_manufact(t-1) + 0.36*lnAGEI\_manufact(t-2) \quad (8)$$

The residuals passed the Jarque Bera test with a lower probability of 45%, but the White Heteroskedasticity Test with 99%, while the LM Test indicated the absence of any autocorrelation for the first 8 lags with a probability of 63%. We easily notice from eq. (8) that AGEI behavior in the Manufacturing Section is mostly affected by its first and second lags, while the LPI has both a positive influence in the present and a slower negative impact with a 4 months delay.

After applying the Box Jenkins methodologies for the LPI series an **ARIMA(2,1,0)** resulted, which is actually a second order autoregressive model. The estimated AGEI series based on eq. (8) and on the predicted LPI values were plotted in Fig.4 together with the real AGEI values. The average error of the model was of 0.015.

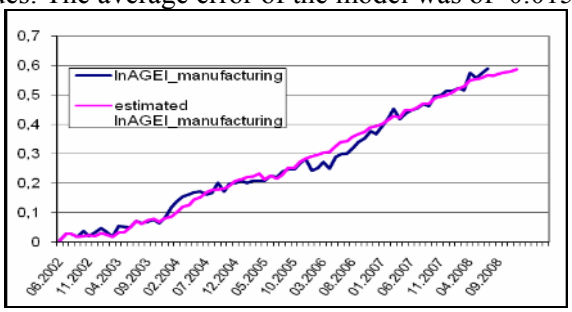

Fig.4 Real and estimated *lnAGEI_manufacturing* series

The case of total Industry:

$$lnAGEI\_total(t) = -0.007 + 0.18*lnLPI\_total(t) + 0.65*lnAGEI\_total(t-1) + 0.22*lnAGEI\_total(t-6) \quad (9)$$

The residuals passed the Jarque Bera test with a high probability of 82.2% but the White Heteroskedasticity Test with only 40%. The LM Test indicated the absence of any autocorrelation for the first 6 lags with a high probability of 74%. From eq. (9) describing the relation between the AGEI and LPI on total Industry, AGEI behavior seems to be influenced by the present value of the LPI and by its own first and 6-th lags.

After applying the Box Jenkins procedure it resulted an **ARIMA(3,1,3)**, which was used for predicting the LPI series. The estimated AGEI series based on eq. (9) and on the predicted LPI were plotted in Fig.5, together with the real values. The average error was of 0.012.

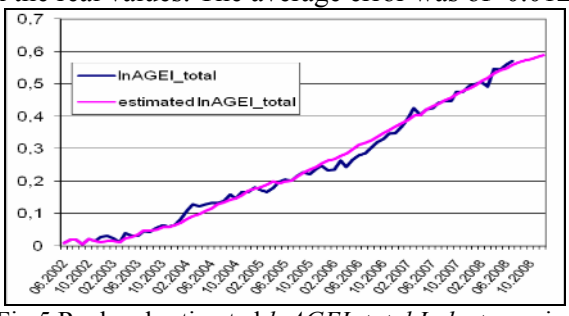

Fig.5 Real and estimated *lnAGEI_total Industry* series

When dealing with first differenced log series, we could only find one valid lagged model, for the case of Manufacturing Industrial section, described below:

$$dlnAGEI\_manufact(t) = 0.008 + 0.28 * dlnLPI\_manufact(t) - 0.26*dlnAGEI\_manufact(t-1) \quad (10)$$

The residuals passed the Jarque Bera test with a high probability of 91.5%, the White Heteroskedasticity Test with 93.3%, while the LM Test indicated the absence of any autocorrelation for the first 8 lags with a probability of 79.3%. We notice from eq. (10) that the first

differenced AGEI behavior in the Manufacturing Section is mostly affected only by its first lag, and by the present value of the first differenced LPI. After using the already predicted values for dlnLPI in eq. (10) we estimated the dlnAGEI values, which were plotted in Fig.6 together with the real dlnAGEI values. The average error of the model was of 0.0062.

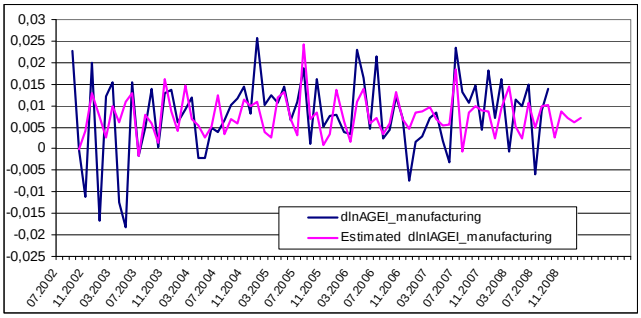

Fig.6 Real and estimated d*lnAGEI_manufacturing* series

From the presented tests (as well as several other tests) it resulted that the above econometric models were able to estimate the AGEI behavior quite well on short terms, being however, inefficient for long term forecasting.

4.1.2 VAR model

When estimating the VAR models we came across the following problems. Although the stability condition was satisfied when working with first differenced series, when estimating a VAR(8) model for example, we noticed that most of the estimated coefficients were statistically insignificant. The possible reasons might be that a VAR model implies the same number of lags for all equations, although some lags are irrelevant. Besides that, we were aware that a VAR(8) with two endogenous variables implies a number of 34 estimations, which is almost 47% of our total number of observations. That is a reason more, why we considered necessary to repeat the tests from a neural network perspective, as well. However, estimating VAR models helped us identify the most significant lags that we considered when building the lagged models.

4.2 The neural network results

Our last attempt to describe the relationship between LPI and AGEI is based on neural network models. In order to estimate and forecast the correlation, we used a feed-forward neural network, which was implemented in MATLAB (by using the MATLAB neural network toolbox). As already stated in subsection 3.4, we performed two types of experiments using this network. The first experiments were used in order to forecast the future values of LPI and AGEI, based on their past *LAG=12* estimated values. The real time series had an obvious increasing trend. Since we had to specify a fixed input range to the neural network and since the outputs of the neural network from previous time steps had to be used as inputs in order to estimate the values of the future time steps, it turned out to be rather difficult to use the series in level. Because of this, we used the first differenced series (*dLPI* and *dAGEI*) and the first differenced log series (*dlnLPI* and *dlnAGEI*). These series did not have a pronounced increasing or decreasing trend and, thus, they were suitable for the neural network. Actually, the series were modified further. We first computed their average and subtracted it from the values of the time series and then scaled them from their real interval (approximately *[-0.05, +0.05]*) to the interval *[-F,+F]* (we chose *F* to be approximately *30*). Then, we trained the neural network with this data. A training input consisted of *LAG* consecutive values of *dLPI* and *dAGEI* (respectively, *dlnLPI* and *dlnAGEI*): $dLPI_{t-LAG},\ldots,dLPI_{t-1}$, $dAGEI_{t-LAG},\ldots,dAGEI_{t-1}$ and the corresponding output was the pair $(dLPI_t\ ,\ dAGEI_t)$ (similarly for *dlnLPI* and *dlnAGEI*). For all the series, the results were surprisingly accurate. Fig. 7-12 show both the real values (in blue) and the estimated values (in red) for the first differenced and first differenced log series for the industry sections and the total industry. We used 18 neurons on the hidden layer of the neural network, which was trained for 10000 iterations on all the available data (73 observations). Afterwards, we forecasted 73 values in the future. Thus, in the figures, the real values are plotted only up to the middle of the time range. Note, though, that the estimated values are extremely accurate for the first half of the time range, such that the red graph occasionally completely covers the blue graph. Of course, in practice, we do not expect to use such long term predictions. However, their stability and similarity to the original data (visually observed) showed us that the neural network did not suffer form overfitting, although the unexpectedly accurate estimates for the available data could have indicated that. Thus, we believe that the network estimated the correlation functions quite well and the forecast is believed to be sufficiently reliable.

For the second type of experiments, we trained a feed forward neural network with 6 neurons in the hidden layer for 1000 iterations in order to learn the function g_{AGEI} presented in subsection 3.4. This function estimates the values *dAGEI_total(t)* based on the previous *LAG* values and the current values of *dLPI_manufact*, *dLPI_mining* and *dLPI_energy*. We trained the network on 46 observations and tested its outputs on the remaining ones (see Fig. 13).

5 Conclusions

From our study, we noticed that in all the Romanian Industry sections as well as for the whole Industry, the AGEI is indeed correlated to LPI. Their behavior, as well as short term forecasts, can be properly estimated by both the econometric lagged models and the feed forward neural network. However, the neural network is

able to approximate the long term behavior rather well, too. This was somewhat to be expected, because the neural network can approximate a wide range of functions (e.g. non-linear), while the econometric models we used considered mainly linear functions. Since the results were satisfactory we consider further using these models in future studies as well.

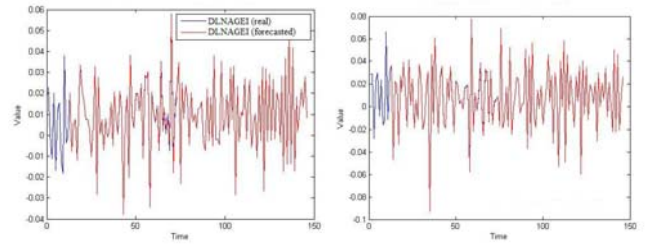

Fig.7 dlnAGEI_manufact (left) and dlnLPI_manufact (right)

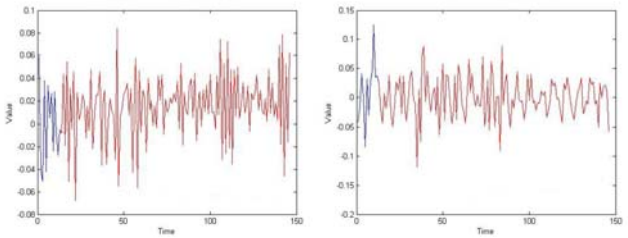

Fig.8 dlnAGEI_energy (left) and dlnLPI_energy (right)

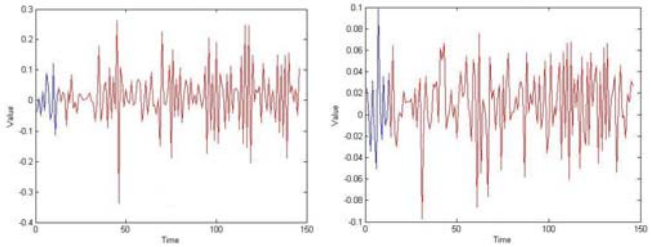

Fig.9 dlnAGEI_mining (left) and dlnLPI_mining (right)

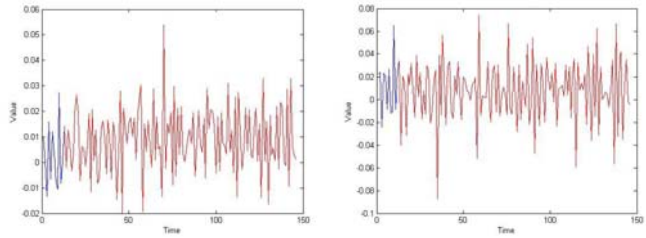

Fig.10 dlnAGEI_total (left) and dlnLPI_total (right)

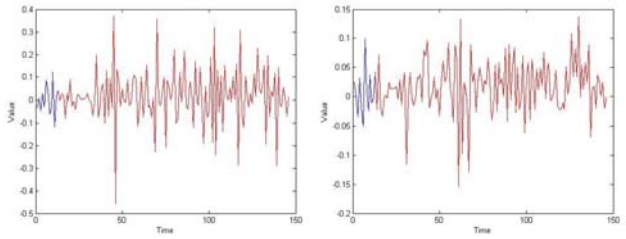

Fig.11 dAGEI_mining (left) and dLPI_mining (right)

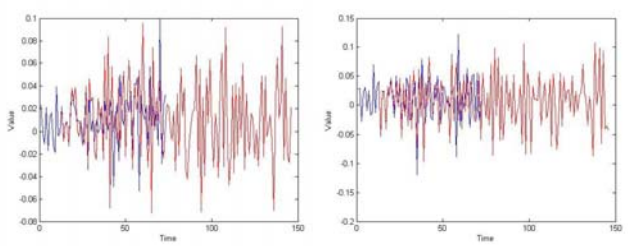

Fig.12 dAGEI_manufact (left) and dLPI_manufact (right)

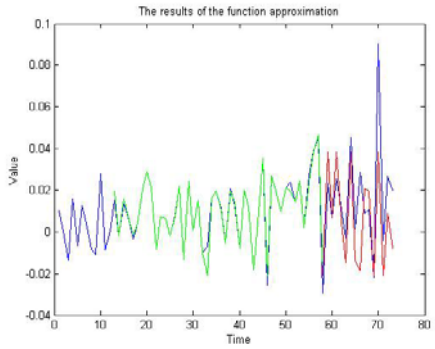

Fig.13 dlnAGEI$_{total}$ (real-blue, trained-green, estimated-red)

References:

[1] M. E. Andreica, I. Dobre, M. Andreica, B. Nițu, R. Andreica, A New Approach of the Risk Project from Managerial Perspective, *Economic Computation and Economic Cybernetics Studies and Research,* vol. 42, no. 1-2, pp. 121-129, 2008

[2] M. I. Andreica, N. Tapus, Maximum Reliability K-Hop Multicast Strategy in Tree Networks, Proc. of the IEEE International Symposium on Consumer Electronics, pp. 169-172, 2008

[3] C. Mereuță, SWOT analysis of the Romanian Manufacturing during 1998–2004 from economic growth perspective, *Working Papers of Macroeconomic Modelling Seminar*, no.071302, 2007

[4] B. Pauna, National and foreign investments impact upon Romanian labor market using VAR and VEC, *Working Papers of Macroeconomic Modelling Seminar*, no.071504, 2007

[5] Saeed Moshiri, Norman Edward Cameron, Neural Network vs Econometric Models in Forecasting Inflation, SSRN *Working Paper Series,* 1998

[6] Lutkepohl H., Econometric Analysis with Vector Autoregressive Models, *EUI Working papers*, 2007

[7] N.R. Swanson, H. White, A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks, *Review of Economics and Statistics* Vol.79, pp. 540-550, 1997