Using Data Mining Techniques in Macroeconomic Analysis on Romania's Case

STELIAN STANCU
Department of Economic Informatics and Cybernetics
Academy of Economic Studies, Bucharest
Dumitru Secăreanu, nr.5, ROMANIA
stelian_stancu@yahoo.com http://www.ase.ro

BOŞCOIANU MIRCEA
Department of Military Sciences and Management
"Henry Coandă" Air Force Academy, Braşov
Mihai Viteazul, nr.160, ROMANIA
mircea_boscoianu@yahoo.co.uk http://www.afahc.ro/

ALEXANDRA MARIA CONSTANTIN
Academy of Economic Studies, Bucharest
Boulevard Unirii, nr. 75, ROMANIA
constantin_alexandra_maria@yahoo.com http://www.ase.ro

Abstract: In the present data mining techniques are the most utilized methods of calculus in discovering existent relations between different components, objects, phenomenon, etc. Economic analysis must be achieved by the most advanced methods of calculus, so that eventually, thanks to their evolution, they can generate much more viable data by minimizing information loss in order to better reflect reality.

Romania, as an EU country, is in direct competition with other European countries, classifying in the category of countries heavily dependent on external commerce, and so, by economic relations with other countries. At the same time, its EU membership grants Romania the possibility to develop and implement strategies with regards to economic performance development, in order to better utilize external production factors. In this context, this paper will try to complete a macroeconomic analysis of relations between European GDP, Romanian GDP, EU population, Romania's population, Romania's imports from the EU and its exports to the EU using data mining techniques.

Key-words: data mining, main component analysis, cluster analysis, discriminant analysis, GDP, interest rate, entropy, Bayesian classification, SAS, EU.

1. Introduction

Considering export and GDP are classified as main macroeconomic indicators by which rational economy functions by, a study of their interactions is wholly necessary, but also between them and other economic parameters, such as interest rate and exchange rates.

There are numerous opinions considering causality, both ways, between export and GDP. The first hypothesis claims that export leads to growth, while other hypotheses, just as interesting, consider that increased production determines increased exports.

As for the first hypothesis, Makki and Somwaru[1] claim that increasing exports is a productivity growth factor, thanks to earnings gained by raising scale income and the existence of a larger external market.

On the other hand, increasing exports relax constraints created by the exchange rate, which results in an important capital/intermediate entries from intensive technologies.

Thanks to export growth, efficiency is increased, as exporters are able to compete on foreign markets,

[1] Makki S.S., Somwaru A., *Impact of foreign direct investment and trade on economic growth: Evidence from developing countries*, American Journal of Agriculture Economics, 86, No. 3, 2004, pag. 795-801.

leading to technological progress and psychological manipulation of local entrepreneurs.

Grossman and Helpman[2] consider that regiments that open up commerce contribute to technological imports and, also, create a better climate for investors.

The causality relation between export and GDP isn't necessarily positive, but can also be negative if production growth leads to lowering exports. This may occur when there is a growth of consumer demand on the national market, in the nonnegotiable export sector, which can eventually lead to reduced export growth, due to consumer demand in the national economy.

Similarly, it is difficult to identify the impact commerce has on economic growth and there are arguments that show that countries with higher incomes, for reasons other than commercial ones, have the tendency to increase commercial exchanges.

Another direction whereas the connection between growing commerce and growth is that it is necessary to include institutional factors in estimating commerce coefficients and other variables.

Taking into account the availability of data from official statistics, the multidimensional model is as follows:

$$\ln Y = \alpha + \beta_1 \ln E + \beta_2 \ln I + \beta_3 \ln Rd + \beta_4 \ln C + \varepsilon ,$$

where:

Y -real GDP for Romania or the EU

E - real export

I - Romania's imports from the EU

Rd – real exchange rate

Variables are expressed as logarithms, and the residual, ε, is distributed normally and follows white noise behavior.

As the direct causal approach between the variables taking into account is troublesome, almost impossible, and theoretical inferences are missing, it is recommended that we characterize these variables to be able to classify them in classes, with the help of an ensemble of methods and techniques. This mechanism is recognized in artificial intelligence as pattern recognition technique.

2. Data mining techniques used in macroeconomic analysis

2.1. Unsupervised recognition methods

Unsupervised recognition methods imply having a set of characteristics, models, correlations between entry data, in order to supply homogenized data, so that it ensures informational redundancy and a very defined character at the exit data set.

Using these methods is recommended when:

- the size of the entry data is large
- homogenous data is required
- a regression model must be perfected so that it precisely reflects as many of the existent information links contained in a group

These methods consider the following aspects:

- entry data set, which will be further processed
- static data properties
- maximizing the quality of the model obtained through the homogenous data.

Starting from the purpose of unsupervised learning methods, we must take into account the problem of grouping data.

Data analysis techniques that are based on unsupervised learning techniques are: Principal Component Analysis (PCA), Factorial Analysis (FA), Cluster Analysis (CA).

These techniques consist of simultaneously modeling data through multiple variables, in order to render the relations that can exist between them.

2.1.1. Principal Component Analysis

In pattern recognition, selecting and extracting hidden data through observations subject to modeling is decisive to choosing any classifier. Thus, trait selection must be approached as a data compression process, which implies a linear transformation from the initial observation space to a smaller space. This transformation implies the conservation of most of the information, and implies the application, in small spaces, of efficient algorithms (in real time), the most utilized being linear algorithms.

Specialist define PCA as[3]: “a multidimensional data mining technique's purpose is to decompose total variability from the initial space under the form of a number composed from less elements, without the decomposition containing informational redundancy”.

[2] Grossman G., Helpman E., *Innovation and Growth in the Global Economy*, MI Press Cambridge, MA, 1991, U.K.

[3] Ruxanda, Gh., *Analiza datelor*, Editura ASE, 2001, Bucureşti, pag. 406.

Principal Component Analysis goes through several steps:

- calculating the covariance matrix of entry data
- maximizing variance
- calculating proprietary values and arranging them in a descending manner (principal components)
- determining proprietary vectors associated to proprietary values
- determining the linear combination in the new situation

2.1.2. Factorial analysis

Factorial analysis has the great advantage of offering the possibility to quantify unobservable data characteristics. It also estimates coefficients that retain patterns, if we know the existent correlation between common factors and the other factors can be described as a linear combination.

The common (causal) factor is the factor that explains the variation of at least two variables. As opposed to PCA, FA tries to model correlations that already exist between the variables subject to analysis.

Steps of FA are:.

- determining the minimal number of causal factors;
- factor rotation in order to discover the factor solution;
- interpreting common factors;
- estimating the factor score matrix.

2.1.3. Cluster Analysis

Based on hierarchic classification algorithms, cluster analysis is useful in summarized data description. This way, the cluster method's purpose is to descriptively classify data by identifying similar groups they could be included in.[4]

The advantages of this technique are:

- studying significant links between data
- retaining the most important characteristics from the data set
- summarizing information extracted from the data

2.2. Controlled pattern recognition. Discriminant analysis and Bayesian classification.

A unique field in which discriminant analysis is applied is supervised pattern recognition. As opposed to unsupervised pattern recognition, controlled forms of pattern recognition (or supervised pattern recognition) require previous knowledge of a set number of classes and a set of patterns[5]. Controlled pattern recognition refers to using methods and techniques to partition a number of classes based on partitioning criteria. Supervised pattern recognition takes into account the characteristics and optimal appurtenance of the object in a class, based on a representative sample.

Discriminant analysis unites explicative, descriptive and prediction models in order to study a data set split into *p* classes. Every object is characterized by an ensemble m of quantitative variables and a qualitative variable, thus identifying appurtenance class.

The recommended selection criteria for new variables is to maximize class cohesion with the help of intra-class variance and inter-class variance. The new variables must, according to the general principle from factorial analysis, be uncorrelated with each other, two at a time.

Discriminant analysis is used to determine the characteristics that differentiate group members. This type of analysis can be used to:

- determine which groups are similar;
- identify specific variables within a group;
- compare results of different sets of variables;
- identify members with special characteristics.

The discriminant analysis model implies finding a linear combination of ratios and independent variables: $D = p_0 + p_1V_1 + p_2V_2 + ... + p_nV_n$ (1)

where:

D - is the value of the discriminant

p - is the ratio or discriminant coefficients

V - is the predictor

2.1. The role of informational entropy in measuring information quantity

In order to measure information quantity at a data set level, we must take into account:

- the covariance matrix;
- the net quantity of information;
- informational entropy.

[4] Babucea, A. G., Dănăcică, E. G., „*Analiza cluster în studii de priximitate a şomajului la nivelul judeţelor României la începutul crizei economice*”, Analele Univesrtităţii „Constatin Brâncuşi” din Târgu Jiu, Seria Economică, Nr. 1/2009.

[5] Ruxanda, Gh., „*Data Mining*”, curs predat la Masterat BDSA, ASE, Bucureşti, 2010, pag. 82.

The net quantity of information is determined as follows: $I(X,Y) = \sigma_X^2 + \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}$ (2).

The covariance matrix is: $\Omega = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{pmatrix}$.

The covariance matrix's determinant is: $\det\Omega = \sigma_X^2\sigma_Y^2 - \sigma_{XY}^2$ from where we can deduce that $I(X,Y) = \sigma_X^2 + \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2} = \sigma_X^2 + \frac{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2} = \sigma_X^2 + \frac{\det\Omega}{\sigma_X^2}$.

Thus, the relation between the net quantity of information and the determinant of the covariance matrix is: $I(X,Y) = \sigma_X^2 + \frac{\det\Omega}{\sigma_X^2}$. We can now determine the net quantity of information through the covariance matrix.

However, this measure is not enough to draw a clear conclusion.

In order to correct this shortcoming, entropy must be determined, a more exact and rigorous measure used to determined informational quantity. Informational entropy (7) measures incertitude associated with a random variable. This measure indicates the quantity of information contained within a message. The higher the entropy, the more likely it is that it does not contain precise information.

3. Empiric results

By using the data mining techniques described and the statistic software SAS 9.1 for Windows, a study was made regarding interactions between considered macroeconomic indicators, as well as the quantity of information contained in the data set.

Observation: Notations used in writing the equations are:

$x_1 = \text{Export\_RO\_UE} \rightarrow \text{Romania's Exports in EU}$

$x_2 = \text{Import\_RO\_UE} \rightarrow \text{Romania's Imports from EU}$

$x_3 = \text{PIB\_UE} \rightarrow \text{EU GDP}$

$x_4 = \text{PIB\_RO} \rightarrow \text{Romania GDP}$

$x_5 = \text{POP\_UE} \rightarrow \text{EU's Population}$

$x_6 = \text{POP\_RO} \rightarrow \text{Romania's population}$

$x_7 = \text{RD\_UE} \rightarrow \text{EU's interest rate}$

$x_8 = \text{RD\_RO} \rightarrow \text{Romania's interest rate}$

$x_9 = \text{Curs\_R\_E} \rightarrow \text{Exchange rate RON/EURO}$

By applying the PCA technique to the official data set[6], with the aid of the PROC procedure, it returned that information can be retained in a principal component at a ratio of 97.39%. Thus, relations between variables are powerful and can be represented in the new causal space through a component. Also, the 9 original variables can be re-expressed through the first component, with an information loss risk of 2.61%. Behavior of the first component is given by relation (3):

$$w_1 = 0.226012x_1 + 0.974122x_2 + 0.000133x_3 + 0.000003x_4 + 0.002316x_5 - 0.000126x_6 - -0.0000003x_7 - 0.0000007x_8 + 0.000003x_9 \quad (3)$$

Starting from the obtained covariance matrix, which shows the tendency of each pair of traits to variate together or co-variate, we can say that:

- Exports and imports grow together ($C_{\text{Export / Import}} > 0$);
- Exports and GDP at an EU level tend to increase in the same direction ($C_{\text{Export / PIB}} > 0$)
- if the interest rate continues to drop, exports tend to grow ($C_{\text{Rd\_RO / Export}} < 0$)

Also, with the help of graphic representation we can infer the impact of the second component, which has increasing impact on each variable after the first component. For example, figure 1 presents the evolution in the new space of the second principal component, in report to the values of the original variable Imports.

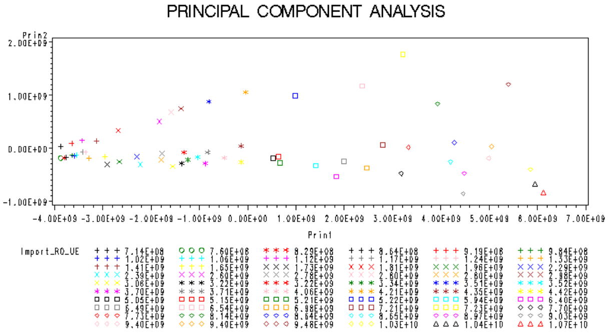

Fig. 1

The FACTOR procedure is based on applying the method of factorial analysis by ensuring the maximization of informational content in a singular factor, in an 80.53%. Thus, factor 1 can be written as a linear combination of macroeconomic indicators taken into consideration with their associated coefficients:

$$w_1 = 85x_1 + 95x_2 + 98x_3 + 90x_4 - 98x_5 - 95x_6 - -95x_7 - 86x_8 + 58x_9 \quad (4)$$

[6] The data set contains trimestrial data from macroeconomic variables like GDP in the EU, GDP in Romania, population in the Eu, Population in Romania, Romanian export to the EU, Imports in Romania from the EU, interest rates in the EU, interest rates in Romania and RON/EURO excahnge rate from 1996 to 2010.

We notice thus that the variation of Romania's GDP will be most affected by the altering of interest rates in Romania. When interest rates grow up by 1%, GDP drops by 170,049 mil. Euros, if the values of other indicators aren't modified.

The cluster analysis, ran through the CLUSTER procedure, is useful in synthesizing data relations. By applying the centroid method, we obtained, based on the correlation matrix, 9 components that synthesize the original variables.

After applying the cluster method:

- three clusters resulted, with cluster 2 being the largest
- by the method of the centroid, the new principal groups will contain the first component, from which we obtain a large amount of synthesized informational

The 60 observations were divided in three classes of different proportions. For example. The first class contains 21 observations, or 35% of the total observations. Class 2, which incorporates 38.33% of observations, is the class that encapsulates the largest amount of information.

Of the data collected from cluster 1 (homogenous data) we can deduce that we have discovered the existing relations between the 9 original variables based on the observations from this cluster. Thus, between the depend variable Export and the other 8 variables we can say:

$$Export\_RO\_UE = 1.41 + 2.178\,\mathrm{Im}\,port\_UE\_RO + 6.05PIB\_UE + 3.23PIB\_RO + 7.32POP\_UE - -2.272POP\_RO - 39883722541RD\_UE - -1.39835RD\_RO + 3.01Curs\_R\_E$$

The interest rates at EU level have the greatest influence on exports. There are, also, 8 other relations that can be deduced from cluster 1.

By joining clusters we get data similar to the previous (analysis at individual cluster level), in the sense that it highlights, with greater precision, the indisputable influence of EU interest rates over exports.

On the basis of the information obtained through inter-class relations, or the objects which are in different classes, we can deduce, for example, the relation between export and its other variables. Thus, we may say that the interest rate in Romania (coefficient is -27829608137) affects exports from Romania to the EU more than interest rates at in the rest of the EU. When analyzing at cluster level, we have to take into account both the relations between variables from different classes, inter-class, as well as relations between components inside their own class, as they can have different results, in the sense that there can be significant differences between inter-class and intra-class approaches.

On the other hand, canonical approach (method part of the cluster analysis implemented in SAS through the CANDISC procedure), treats relations between components of the entry data set, taking into account relations in all the components from the entry data set. Based on obtained information, we can again deduce that interest rate in the EU strongly affect Romanian exports to the EU.

Another disadvantage brought up by canonical analysis is that it analyzes existing correlations between components in each cluster. In cluster 1, we have $C_{\mathrm{Expor\_RO\_UE/Curs\_R\_E}} = 1.5054 > 0$, and we can conclude that the two variables, Exports and Exchange Rate, in this cluster at least, grow together.

Of course, we can also determine the covariance matrix, by grouping the three clusters. In this case, $C_{\mathrm{Expor\_RO\_UE/Curs\_R\_E}} = -5.87317 < 0$, from which we may say that while the exchange rate keeps dropping, exports are growing.

If every component belongs to different classes, or if the analysis was done inter-classes, different correlation coefficients were obtained. Thus, Exchange Rate and Exports are directly correlated ($COR_{\mathrm{Expor\_RO\_UE/Curs\_R\_E}} = 0.82712 > 0$), this relation being significant to other correlations in 38% proportion.

Also with the help of the CANDISC procedure, we will determine all discriminant functions.

By using the properties of the discriminant functions (and based on information obtained), we may say that objects from the three clusters are not part of common regions, which is to say they are well defined in separate surfaces (see figure 2), as each of them $D_{ij}(x) > 0$, $i = \overline{1,3}$, $j = \overline{1,3}$, $i \neq j$.

Considering the normality hypothesis is respected, a statistics test will be applied to discover if the discriminant is or isn't statistically significant.

As the value of the Lamda Wilks statistic is infinitesimal, it tends towards 0. Thus, $\mathrm{Pr}\,ob < 0.001$. The null hypothesis is rejected and the discriminant exponent is statistically significant.

We can test the homogeneity of covariance matrices that make up discriminant functions.[7] As we know (based on previously obtained information) that the afferent probability of Pillai's Trace test is smaller than 0.001, we can say that the matrices are homogenous.

[7] http://www.mpopa.ro/statistica_master11_manovamancova.pdf

As a result of Bayesian classification, a representative model has been estimated, in which the dependent variable is Romania's GDP:

$$PIB\_RO = 160193 + 0.000007 \cdot Export\_RO\_UE +$$
$$0.00008 \cdot \mathrm{Import}\_RO\_UE + 0.0066 \cdot PIB\_UE +$$
$$+ 0.0003 \cdot POP\_UE + 0.037 \cdot POP\_RO +$$
$$98.3474 \cdot RD\_UE + 1064.972 \cdot RD\_RO +$$
$$0.0261 \cdot Curs\_R\_E$$

We have thus constructed a model that shows Romania's GDP's dependency on other economic factors.

On the other hand, the graphic that highlights the repartition of the three clusters, separated by their separation lines:

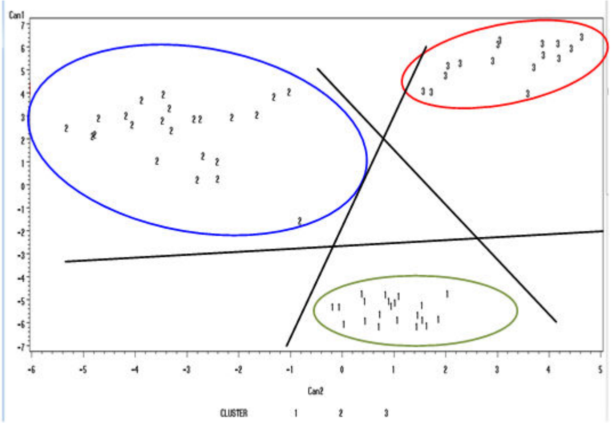

Fig. 2

Not least, in order to measure the quantity of information in two variables, Romania's GDP and the exchange rate, we calculate the net quantity of information:

$$I(X,Y) = \sigma_X^2 + \frac{\det \Omega}{\sigma_X^2} = 92434539 \cdot 10^{11} +$$
$$+ \frac{92434539 \cdot 10^{11} \cdot 104947985.93 - 2868759^2 \cdot 10^{14}}{92434539 \cdot 10^{11}} = 1.5914 \cdot 10^7$$

Thus, according to the result we obtained, we can say that the value of the quantityof information is very large.

In order to draw a clearer conclusion, entropy was determined with the help of the ENTROPY procedure. Results after executing this procedure was that the value of informational entropy was 2.998984. The obtained value highlight that entropy is relatively small in the context of informational quantity contained within the two variables.

4. Conclusions

Connections governed by social, economic, political, etc. elements tend to influence procedures in growth or economic decline in a considerable measure. Thus, any decision taken by the players (banks, investors, government etc.) at a national economic level, must be based on concrete data with complete informational analysis that can be extracted from it.

Due to analysis of the evolution of the economy, of the trajectory of the economic system, the research of causal relations must be taken into account, in order to econometrically measure them.

In order to restore economic balance, it is necessary to diminish incertitude, through knowledge, but also through causality, and also regulations regarding stability of the economic indicators (interest rates, taxes, etc.).

Data mining allows for an abstraction of causal relations between the considered variables and extracting, as precisely as possible, information from the original data in order to apply them in the decision making process.

In conclusion, through its implications, the economy must be studied in such a way that it radiograms as much of the impact the decision has, in the wish of optimizing the game on an economic, national and international, scene.

References:

[1] Babucea, A. G., Dănăcică, E. G., *Analiza cluster în studii de proiximitate a şomajului la nivelul judeţelor României la începutul crizei economice*, Analele Univestităţii „Constantin Brâncuşi” din Târgu Jiu, Seria Economică, Nr. 1, 2009.

[2] Grossman, G., Helpman, E., *Innovation and Growth in the Global Economy*, MI Press Cambridge, MA, U.K., 1991.

[3] Makki, S.S., Somwaru, A., *Impact of foreign direct investment and trade on economic growth: Evidence from developing countries*, American Journal of Agriculture Economics, 86, No. 3, 2004.

[4] Stancu, S., *Econometrie Teorie şi aplicaţii utilizând EViews*, Editura ASE, Bucureşti, 2011.