Economic Performance Competitor Benchmarking using Data-Mining Techniques

Assist. Adrian COSTEA, PhD
Academy of Economic Studies Bucharest

In this paper we analyze comparatively the macroeconomic performance of different Central and Eastern European countries by the means of Data Mining (DM) techniques. We analyze the economic situations of three EU countries (Poland, Slovenia and Latvia), a newly-accepted one (Romania), and other two non-EU countries (Russia and Ukraine). We have depicted economic performance of countries using a number of macroeconomic variables for the time period from 1993 till 2000. The economic variables that we used were: Currency Value (CV) (the inverse of the Exchange Rate - ER), Domestic Prime Rate (Refinancing Rate – RR), Industrial Output (IO) compared to previous periods in percentages, Unemployment Rate (UR), Foreign Trade (FT) in millions of USA dollars. The dataset consists of monthly/annual data during the period 1993-2000, in total 225 cases with five variables each. As DM techniques we firstly applied Self-Organizing Map (SOM) to group the countries according to their economic performance. We applied a so-called "two-step" SOM clustering: firstly, we built larger maps that contained "raw" clusters and then, we re-grouped the "raw" clusters to form a smaller number of "real" clusters. We characterize each "real" cluster, using for each variable the following linguistic terms: VL – very low, L – low, A – average, H – high, VH – very high. Secondly, we go one step further and use the clustering results (of SOM) to build hybrid classification models to help position new countries' performance within the existent economic performance clusters. This type of analysis can benefit the countries involved, EU in its monitoring process, business players such as international companies that want to expand their business and individual investors.
Keywords: *Self-Organizing Map, Multinomial Logistic Regression, Performance Benchmarking, Data Mining Techniques.*

Introduction

The aim of this paper is to analyze comparatively the macroeconomic performance of a number of Central and Eastern European countries (Russia, Ukraine, Romania, Poland, Slovenia and Latvia) by the means of Data Mining techniques. The process of analyzing comparatively the economic performance of different entities is a type of benchmarking also known as *economic performance competitor benchmarking*. Table 1 shows different types of benchmarking depending on the goal and scope. As Table 1 shows some scope-goal benchmarking pairs have more relevance than others. For example, it is irrelevant to compare the entity's strategy with itself, whereas comparing the performance and strategies of different competitors in the same area should be relevant. The type of benchmarking process that constitutes the business application in this paper is at the intersection of performance and competitor benchmarking (the bold word in Table 1).

Table 1. Suitability of goal and scope-oriented types of benchmarking.

	Internal benchmarking	Competitor benchmarking	Functional benchmarking	Generic benchmarking
Performance benchmarking	Medium	**High**	Medium	Low
Process benchmarking	Medium	Low	High	High
Strategic benchmarking	Low	High	Low	Low

(Source: Bhutta & Huq, 1999, originally adapted from McNair & Liebfried, 1992)

We narrow even more our benchmarking applications by only looking at countries' *economic* performance measures. Therefore, benchmarking in our case is understood as the process of continuously comparing and measuring a country' performance against its competitors or against the performance leaders anywhere in the world to gain information that will help the country to take action improving its performance.

We employ DM techniques to perform the economic benchmarking. We associate the business problem of analyzing comparatively the economic performance of countries with the DM clustering and classification tasks. The algorithms used to perform data-mining tasks described above are numerous and they come from different research fields (statistics, machine learning, artificial intelligence, fuzzy logic, etc.). In this paper, we have used one approach for performing the DM clustering task: a heuristic method (neural networks with unsupervised learning algorithm known as Self-Organizing Map algorithm), and a statistical approach for performing the DM classification task: multinomial logistic regression.

2. Methodology and the dataset

2.1. The SOM

The SOM (Self-Organising Map) algorithm is a well-known unsupervised-learning algorithm developed by Kohonen in the early 80's and is based on a two-layer neural network (Kohonen, 1997). The algorithm creates a two-dimensional map from *n*-dimensional input data. After training, each neuron (unit) of the map contains input vectors with similar characteristics, e.g. countries with similar economic performance.

The result of SOM training is a matrix that contains the codebook vectors (weight vectors). The SOM can be visualised using the *U-matrix* method proposed by Ultsch (1993). The unified distance matrix or U-matrix method computes all distances between neighbouring weights vectors. The borders between neurons are then constructed on the basis of these distances: dark borders correspond to large distances between two neurons involved, while light borders correspond to small distances. In this way we can visually group the neurons ("raw" clusters) that are close to each other to form supra-clusters or "real" clusters (Figure 1 (a)). The "raw" clusters can automatically be grouped using another clustering technique such as Ward's method (Ward, 1963).

In addition to the U-matrix map, a *component plane* or *feature plane* can be constructed for each individual input variable. In the feature planes light/"warm" colors for the neurons correspond to high values, while dark/"cold" colors correspond to low values (Figure 1 (b)). The component plane representation can be considered a "sliced" version of the SOM, where each plane shows the distribution of one weight vector component (Alhoniemi *et al*. 1999, p. 6). Also, *operating points* and *trajectories* (Alhoniemi *et al*. 1999, p. 6 and Figure 1 (a) red line) are used to find how different points (observations) move around on the map (e.g. how the countries evolved over time with respect to their economic performance).

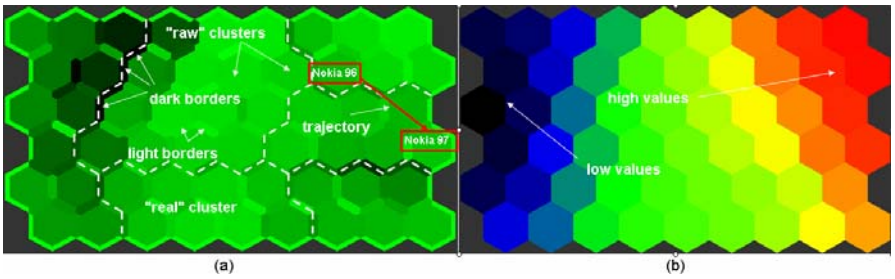

Figure 1. (a) The U-matrix representation with Nenet v1.1a software program and **(b)** some variable component plane

2.2. Multinomial Logistic Regression

MLR classifies cases by calculating the likelihood of each observation belonging to each class. The regression functions have a logistic form and return the likelihood (the odds)

that one observation (x) belongs to a class (C):

$$odds(x \in C) = \frac{1}{1+e^{-logit}} = \frac{1}{1+e^{-(w_0+w_1v_1+...+w_pv_p)}}$$

where v_1, $\ldots v_p$ are the input variables, and $w_0,\ldots,w_p$ are the regression coefficients (weights).

MLR calculates the estimates ($\hat{w}_i$, $i = 0,..., p$) for the coefficients of all regression equations using the maximum likelihood estimation (MLE) procedure. If there are c classes, MLR builds c-1 regression equations. One class, usually the last one, is the reference class.

MLR calculates the standard errors for the regression coefficients, which show the potential numerical problems that we might encounter. Standard errors larger than 2 can be caused by multicolinearity between variables (not directly handled by SPSS or other statistical packages) or dependent variable values that have no cases, etc (Hosmer & Lemeshow, 2000).

Next, MLR calculates the *Wald* statistic, which tests whether the coefficients are statistically significant in each of the c-1 regression equations. In other words it tests the null hypothesis that the logit coefficient is zero. The Wald statistic is the ratio of the unstandardised logit coefficient to its standard error (Garson, 2005).

Next, MLR shows the degree of freedom for the Wald statistic. If "sig." values are less than the 1 – confidence level (e.g. 5%) then the coefficient differs significantly from zero. The signs of the regression coefficients show the direction of the relationship between each independent variable and the class variable. Positive coefficients show that the variable in question influences positively the likelihood of attaching the specific class to the observations.

Values greater than 1 for $e^{\hat{w}_i}$ show that the increase in the variable in question would lead to a greater likelihood of attaching the specific class to the observations. For example, if $e^{\hat{w}_1} = 3$ for class c_1 and variable v_1, we can interpret this value as follows: for each unit increase in v_1 the likelihood that the observations will be classified in class c_1 increases by approximately three times.

Finally, MLR shows the lower and upper limits of the confidence intervals for the $e^{\hat{w}_i}$ values at the 95-per cent confidence level.

2.3. The data set

We characterize countries' economic performance with the aid of the following indicators:

- Currency Value (CV) is the inverse of the Exchange Rate (ER), and shows how many US dollars one can buy with 1000 current units of national currency and depicts the purchasing power of each country's currency,
- Domestic Prime Rate (Refinancing Rate – RR), which shows financial performance and level of investment opportunities. This interest rate is established by the central bank of each country and is the interest rate for refinancing the operations of the commercial banks. Hence, it affects all other interest rates.
- Industrial Output (IO[1]) compared to previous periods in percentages, to depict industrial economic development,
- Unemployment Rate (UR), which characterizes labor exploitation and, more generally, the social situation in the country, and
- Foreign Trade (FT) in millions of USA dollars, to reveal the surplus/deficit of the trade budget.

The dataset consists of monthly/annual data for six countries (Russia, Ukraine, Romania, Poland, Slovenia and Latvia) during the period 1993-2000, in total 225 cases with five variables each. In Costea *et al*. (2001) there were two more variables in the dataset: exports (EXP) and imports (IMP) in millions of USD, as intermediary measures to calculate the foreign trade. We discarded them in the later studies as these variables are strongly correlated with the foreign trade variable. Also, we replaced the first variable (Exchange Rate) from Costea *et al*. (2001) with the Currency Value variable to ensure comparability between the different countries'

[1] Industrial Output was preferred to GDP per capita as the latter is an annual indicator and we needed monthly data.

currencies. We encountered in some cases missing values, which we have complemented using the means of existing values.

3. Experiment

In this experiment we assess comparatively the economic performance of different countries from one geo-political area: Central and Eastern Europe. Our results are presented in Costea and Eklund (2003) and Costea (2003). We applied SOM to the countries' economic performance data set. The final 7x5 SOM map with the identified "real" clusters (dotted lines) is shown in Figure 2.

The alphabetical order of the cluster identifiers corresponds to the inverse order of economic performance: A – best performance, B – slightly below best performance, C – slightly above average performance, D – average, E – slightly below average performance, F – slightly above poorest performance, and G – poorest performance.

The characteristics of the identified clusters are summarized in Table 2. We characterize each cluster, using for each variable the following linguistic terms: VL – very low, L – low, A – average, H – high, VH – very high. The last two columns of the Table 2 give the overall characterization of each cluster.

The SOM trajectories can be used to check the economic performance of the different countries over time. For example, Ukraine made steady progress between 1993 and 2000 with respect to its foreign trade balance (Figure 2). In 1993, in spite of its high currency value, Ukraine had the worst economic situation (high negative values for foreign trade), which positioned the country in the worst cluster. Year by year the trade balance improved and in 2000 (April, May, June) became positive, which led to Ukraine being placed in the best economic performance cluster (cluster A).

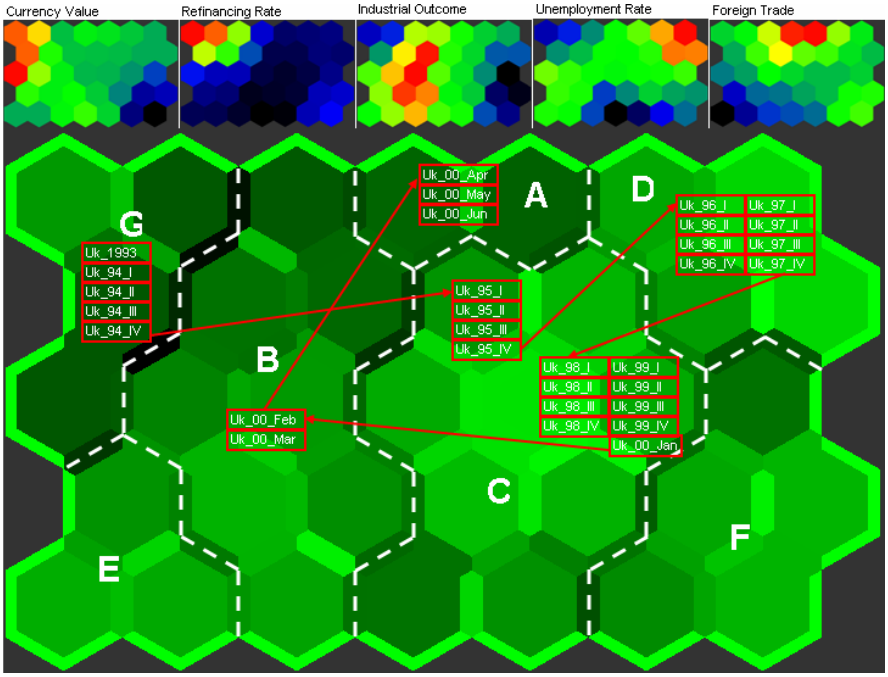

Figure 2. Final 7x5 SOM for the economic data set with identified "real" clusters and feature planes. Trajectories (solid lines) for Ukraine (red).

Once we had constructed the "real" clusters we built the class variable, assigning a class value (1 to 7) to each observation within a cluster. Next, we applied MLR to build the classification models by following some methodological steps (Costea, 2005). We replaced the missing data with the means of existing values. We used SPSS to perform the classification. We standardized input data to zero mean and unit standard deviation (normalization). We validated our models based on the training data by using proportional by-chance and maximum by-chance accuracy rates (Table 3). For example, the training ac-

curacy rate (61.3%) satisfied the proportional by-chance criterion (61.3% > 1.25*29.92% = 37.4%) but slightly failed to satisfy the maximum by-chance criterion (61.3% < 1.25*49.8 = 62.22%). The significance of the Chi-Square statistic ($p<0.0001$) and the overall correlation coefficient (Nagelkerke's R^2 = 74.5%) show a relatively strong relationship between class variable and the economic variables.

Table 2. S*ubjective* characterization of the economic clusters based on the feature planes. Each cluster is characterized subjectively by human interpretation of the feature plane from Figure 2 (e.g. H or VH linguistic term corresponds to "warm" colors). The economic variables are presented in Section 2.3.

	CV	**RR**	**IO**	**UR**	**FT**	**Performance**	**Order**
Cluster A	A&H	L&A	H	A	VH	Best	1
Cluster B	A&H	L&A	VH	L&A	L&A	Slightly below best	2
Cluster C	A&H	VL	A	L&A	L&A	Slightly above average	3
Cluster D	A	L	A	VH	L&A	Average	4
Cluster E	A	L	A	A	VL	Slightly below average	5
Cluster F	VL	L	VL	L&A	A	Slightly above poorest	6
Cluster G	VH	VH	L&VL	L	L	Poorest	7

We interpret the results of MLR by looking at the SPSS output tables. All variables are statistically significant (sig.>0.0001) in explaining the likelihood variations in the dependent variable. Some coefficients in the regression equations are not statistically significant. For example, in differentiating between the average and poorest performance classes (regression equation 4), the "Industrial Output" variable is the only variable that is not statistically significant (sig. of *Wald* statistic = 0.229 > 0.05). Some values in "Std. Error" column are greater than 2, which indicate a multicolinearity problem for our economic dataset. Variable "Unemployment Rate" has a value of 2.086 in column "Exp(B)" for the 5^{th} regression equation, which means that for each unit increase in this variable the likelihood that the observations will be classified in class E (slightly below average) increases by approximately two times. Next, we try to validate our models based on the test data using the general procedure described in Costea (2005).

Table 3. Accuracy rate validations for the economic MLR classification models. The validation is done according to step 5 of the methodology presented at the beginning of Section 5.2 in Costea (2005).

	Main dataset	Part1 (split=0)	Part2 (split=1)
Learning Sample	61.3%	67%	58.4%
Test Sample	no test sample	57.6%	67.1%

The results of MLR classification technique are rather poor for this experiment. There are major discrepancies between the training and test accuracy rates. Moreover, the classifiers did not learn very well the patterns within the data. More robustness in collecting the data is necessary in order for the classification model to be accurate and useful. As it is known, in data analysis, the results are as good as the data that are based upon. However, the statistical nature of each logistic equations confer our models a good interpretability.

4. Conclusion

In this paper we presented how Data Mining techniques, namely Self-Organizing Map (SOM) algorithm and Multinomial Logistic Regression (MLR) can be used in performing

economic performance benchmarking of different countries from one geo-political area: Central and Eastern Europe.

This type of analysis can benefit the countries involved, EU in its monitoring process, business players such as international companies that want to expand their business and individual investors. Using our models, investors would be able to weigh the different investment opportunities by performing the comparisons themselves.

At the time when Romania will join effectively the EU, it will face new challenges with regard to the changes that the EU integration requires. Consequently, research in the field of economic competitor benchmarking is likely to have more and more practical relevance in the near future.

References

1. Alhoniemi E, Hollmen J, Simula O, Vesanto J. 1999. Process Monitoring and Modeling Using the Self-Organizing Map. *Integrated Computer-Aided Engineering* **6**(1): 3-14.

2. Bhutta KS, Huq F. 1999. Benchmarking – best practices: an integrated approach. *Benchmarking: An International Journal* **6**(3): 254-268.

3. Costea A, Kloptchenko A, Back B. 2001. Analyzing Economical Performance of Central-East-European Countries Using Neural Networks and Cluster Analysis. *Proceedings of the Fifth International Symposium on Economic Informatics*, I. Ivan. and I. Rosca (eds), Academy of Economic Studies Press, Bucharest, Romania, pp. 1006-1011.

4. Costea A. 2003. Economic Performance Classification Using Neural Networks. *Proceedings of the Sixth International Symposium on Economic Informatics*, I. Ivan. and I. Rosca (eds), Academi of Economic Studies Press, Bucharest, Romania.

5. Costea A, Eklund T. 2003. A Two-Level Approach to Making Class Predictions. *Proceedings of 36th Annual Hawaii International Conference on System Sciences (HICSS 2003)*, Sprague Jr RH. (ed.), IEEE Computer Society, Hawaii, USA, January 6-9, 2003, Track: Decision Technologies for Management, Minitrack: Intelligent Systems and Soft Computing.

6. Costea A. 2005. Computational Intelligence Methods for Quantitative Data Mining. Turku Centre for Computer Science, Ph. D. thesis, No. 67, Turku, Finland, November 2005.

7. Garson GD. 2005. *PA 765 Statnotes: An Online Textbook.* Work in progress. [Available at: http://www2.chass.ncsu.edu/garson/pa765/statnote.htm] (Accessed on: 10.10.2006).

8. Hosmer DW, Lemeshow S. 2000. *Applied Logistic Regression*. 2nd Edition, John Wiley and Sons, New York, USA.

9. Kohonen T. 1997. *Self-Organizing Maps*. 2nd edition, Springer-Verlag, Heidelberg.

10. McNair CJ, Liebfried KHJ. 1992. *Benchmarking – A Tool for Continuous Improvement*. John Wiley and Sons, Inc., New York, NY.

11. Ultsch A. 1993. Self organized feature maps for monitoring and knowledge aquisition of a chemical process. Gielen S, Kappen B. (eds.), *Proceedings of the International Conference on Artificial Neural Networks (ICANN93)*, London, Springer-Verlag, pp. 864-867.

12. Ward JH. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**: 236-244.