Associate Professor Dumitru-Iulian NASTAC, PhD
E-mail: nastac@ieee.org
Politehnica University of Bucharest
Professor Alexandru ISAIC-MANIU, PhD
E-mail: alex.maniu@gmail.com
Centre for Industrial and Service Economics, Romanian Academy
Associate Professor Irina-Maria DRAGAN, PhD
E-mail: irina.dragan@csie.ase.ro
The Bucharest University of Economic Studies

ANALYZING THE PROFITABILITY PERFORMANCE OF SMEs USING A NEURAL MODEL

Abstract. *Special models of artificial neural networks (ANNs) have proven their worth in various and sometime unexpected domains. In this paper, our focus was to develop an ANN application in order to analyze the financial performance of the SMEs in Romania. For historical reasons, this sector seems to be still weakly developed in that country, both quantitative (being situated on one of the last places in the EU's entrepreneurial intensity) and qualitative, having a weak economic performance with a modest contribution to GDP. Literature shows the importance of this sector for the economies of different countries, and diverse scientific methods used for its description and analysis. One of our research purposes was the identification of those factors that condition the profitability of companies, thus providing useful directions and possible strategies for developing the SME sector. The selected information source was represented by the annual balance sheets, from about 8000 of medium-sized companies in Romania. As a means of verifying the obtained results, econometric methods were used, such as regression analysis, which could identify and validate the models that emphasize the dynamics with different influence factors. The conclusions obtained could prove their utility in both the investigation of the combining quantitative methods (ANN and regression), and in the SME sector management plan.*
Keywords: SMEs, neural networks, classification, econometric models.

JEL Classification: C10, C38, C52, M11, P12

1. Introduction – The Evolution of SMEs in Romania and among European Countries

A prediction concerning the profitability performances can be made in several ways. One way is to use a long and consistent data series. Here we have a different approach, by trying to establish if a series of economic parameters could

properly indicate a good classification of SMEs, according to their performance. In other words, we are looking to see how strongly correlate could be these parameters with the profitability performance. A neural model is used in order to build such a system to extract information and provide a good classifier. Due to the extreme complex situation here, we will ensure that if the system will work this way, then it will be easy to readapt it for any other EU country or in the rest of the world (any country which may have a complicate economic situation).

It took nearly a decade, after 1990, to transition from a centralized economic system to a market-driven price system, from a state-owned property system to another based on private property, from a controlled pricing system to one based on demand and supply, from a system built on social classes, established on the dominance of the proletarian class, to a democratic, multi-party system.In this context, the private sector and more specifically, the SME sector, had to be reset, and free initiative had to be reinvented (Annual Report of European SMEs 2013-2014; Dragan and Isaic-Maniu, 2012). Starting with 1990, the private sector of the SMBs continuously developed, as we can see, from 81 new company registration in 1990, there were 65479 in 1995, 57197 in 2000, 119048 in 2010, and 123541 in 2015 (Post-Privatization Foundation, 2013).The Romanian SMB sector was still insufficiently developed when it was strongly affected by the economic crisis in 2009-2010 and it managed to slowly overcome the hit, but the comeback lost momentum in 2012. As such, the total number of active non-financial enterprises (504581) existing in 2008, dropped to 489646 in 2009, to 442241 in 2010, to 404338 in 2011, slightly increasing in 2012 to 410210, and to 426295 in 2014. The decrease of almost 20% reflected during the crisis is still to be recuperated (Dragan and Isaic-Maniu, 2013; Annual Report on European SMEs, 2014/2015). The rate of creation of new businesses, an indicator computed in EU countries since 1995 and in Romania since the year 2000 (based on the number of business in 1995, when it was 100%), grew continuously until 2008 when it reached 43.3%, while for the past two years it stopped at 35%.

Entrepreneurial intensity, measured by the number of SMBs to 1000 inhabitants, is a good comparison indicator given.This indicator rated Romania at 21.3‰ compared to the EU average of 42.7‰.The uneven ratio between the share of the added value and the demographic potential in Romania reflects a high gap in the development, productivity and competitiveness among Romanian SMEs (Annual Report on European SMEs 2014/2015). Gradually, SMEs have diversified their activity profile, currently active in 88 NACE activities. Shares of over 5% of the total number are the following sectors: retail sales (21.46%), wholesale commerce (10.62%), land and pipeline transportation (6.05%) and building constructions (5.30%).Most companies are specialized in low technology manufacturing without elements of knowledge economy. Compared to the EU average, Romania has less medium-tech industries (such as producing chemical substances, electric equipment, cars and transportation equipment).

For 2014-2015 we envisage a consolidation of the positive results obtained in the last two years. We estimate that the SMB performance indicators will maintain the upward trend for 2016, at a higher place than the European average. The goal of our research was to find a useful model able to capture the complex relationships among various specific parameters that can describe or influence companies' profit in order to finally classify them. In order to identify some supporting elements for the establishment of some strategies to strengthen the SMEs sector and increase their profitability, the analysis was focused on the strongest segment of SMEs, the medium-sized companies, represented by a number of 7,902 sets of data, which were firstly used to build and evaluate the neural tool and then for the regression models, as alternative validation process of the results obtained in the first phase.

The structure of the paper is as follows. Section 2 presents the problems that concerned the literature review in connection with our aim. A description of the data, together with the methodology involved in this work and the main features of our experimental results are given in the next section, where we also discuss a comparison with a classic approach. The conclusions are formulated in the last part of our paper.

2. Literature Review

The importance of SMEs to the economy is given by various points of view, including the contribution to national production, the use of internal resources, the increase of labor force occupancy, the ease of workforce migration at local and European levels, the increase of national economy competitiveness, and the better use of human capital. This context quite explains the increasing interest of not only decision makers in the economy but also of the political and academic circles.

Thus, a special attention is given to key-factors which determine the increase of the number of SMEs, using the relationship between economic and financial growth, the way in which SMEs overcame the crisis, econometric modeling of firms' growth and the influencing factors. Exploring the main decisive factors for the growth of SMEs in CEE countries was done by Mateev and Anastasov (2010) for which they used a data panel of 560 rapidly growing companies from six emerging economies. Important factors, with leverage effect, are liquidity rates, future growth opportunities, labor productivity.

Lejárraga et al. (2014) explore issues regarding the business internationalization of manufacturing SMEs and their various related services. Based on the experimental results, the link between the firm size and the business performance is highlighted, but this is less obvious in the export outcome. There is a similar situation in manufacturing firms and those in services.

An assessment of the impact of SME sector growth is given by Subhan et al. (2013), mainly focused on the role of innovation and the effects on the Pakistani economic development. For measuring innovation, the authors propose the level of C&D spending, number of patents, number of publications, technological intensity,

and high-technology exports. Other variables included in this study are the weight of exports in GDP, the increase of SMEs, GDP growth, industry weight in GDP, the level of workforce occupancy, the consumption price index, the exports and imports volume and the exchange rate. The uncertainty raised by predictions on macro-economy evolution was identified by Henzel and Rengel (2013) as being caused by the fluctuations of business cycles, oil and raw material prices.

Using the information from a sample of 144 small and middle-sized companies in China and the involvement of the corporate social responsibility(CSR), in (Tang and Tang, 2012) it was considered the environmental performance in conjunction with the size of your company, as well as the differentiation of the involvement in the programs of social responsibility in the light of the economic power of them. Terdpaopong (2011) proposes to determine whether a statistic model can in fact identify the crisis of a firm's debts. A sample of 159 SMEs was used, including some firms with financial difficulties, as well as others which don't have this problem. Using a logistic regression model, validated by specific testing, is used to determine the probable chance of survival or failure using predictive models. Another way for solving the classification problems it was the use of the induction techniques, such as recursive partitioning algorithm (Frydman et al., 1985). Many other authors developed applications of artificial intelligence models, like neural networks (Ripley, 1994; Nastac et al., 2009). Several hybrid solutions were also proposed (Chou et al., 2006).

3. Methodology and Main Results

In the following an extension of the preliminary results published in (Nastac et al., 2014; idem 2016) is presented. The primary raw data are from the Romanian National Statistics Institute (for synthetic indicators and for SMEs demography indicators), and also from the Romanian Trade Registry (the data used for balance sheets).The initial volume of data includes over 500000 micro enterprises (each having up to 10 employees), about 40000 small enterprises (up to 50 employees), near 8000 medium-sized enterprises (up to 250 employees), and 1500 enterprises with over 250 employees. The data consist of a large matrix, where the columns represent a wide variety of parameters, mostly economics. One by one, each line of this matrix is a set of information for a specific enterprise out of all over collection of firms. An important column of this matrix refers to the commercial profitability, and we select this parameter in order to classify the enterprises. We assume that there is a complex relationship between this parameter and the rest of matrix columns. Our goal is to build a system that has as input all possible relevant parameters without the commercial profitability (which is to be used only at the output for building the classes). This way the resulting model is intended to extract all relevant information from the input in order to indicate the performance of the firms.

There is quite a challenge to classify these enterprises since there must be established some borders between the classes that are used in defining success. A

clear image of this problem is given by the histogram of the rate of commercial profitability (which is a column of the previously mentioned matrix). From such a histogram we have to find the mentioned borders which will split the data in a way that allows the classes to have, among each other, somehow similar (or comparable) amount of data.

In the simplest case, suppose that we have to split the data in two groups, and in this case everything could be viewed in black and white. In literature, this issue is usually denoted as classical binary classifications, or as the two-group discriminant problem. A model which implements such a binary classification would just show a good or bad performance for an enterprise, without giving detailed information about its real problems. Better information can be extracted from a model with more than two performance classes. For example, it would helpful to analyze the enterprises according to their possible position in one ofthreeclasses, or categories (poor/ medium/ high) as a result of the splitting the histogram of the rate of commercial profitability in intervals that denote the economic performance.

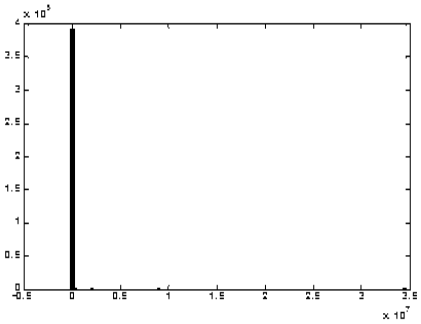

Figure 1. The histogram of the rate of commercial profitability for micro enterprises

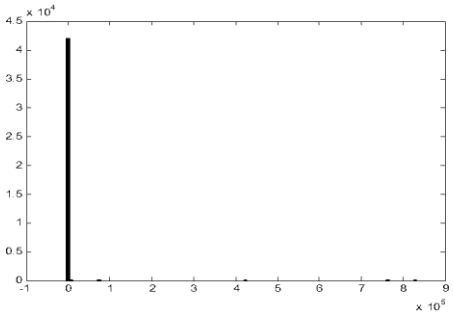

Figure 2. The histogram of the rate of commercial profitability for small enterprises

Depending on the enterprises' type, this parameter could have a range of values starting from zero to more than one million. For example, the left limit could be almost 35 million as we can see in the figure 1, where is presented a histogram of more than 400,000 micro enterprises (after eliminating the outliers).

The whole interval was divided into one hundred parts in order to obtain this histogram, where all other elements (bars), except the one on zero, are small and therefore almost invisible (figure 1). As we can easily remark, it is very difficult to split the observed interval into, e.g., three regions with a similar number of components. We probably need to divide the whole interval into more than 10000 parts in order to obtain a histogram, where we can put such significant borders. Almost the same situation we found for small enterprises (figure 2) where the whole range was divided again into 100 parts. In the case of medium-size enterprises, the

situation is better since, on the histogram (figure 3) we could mark two points that split the whole range in three classes. We denote these points (borders or delimiters) with B_{12} and B_{23}, where each numerical index (the pair of digits as subscript of B) consecutively indicates two neighbor classes for a specific delimiter.

For the last case (large-sized enterprises), there it is also possible to use the delimiters between classes (figure 4).For such a system, the processing of the data was made using a neural model and then, the interpretation of the classifier output has confirmed a better result than a classic regression econometric model, which was employed for comparison at the end of this section.

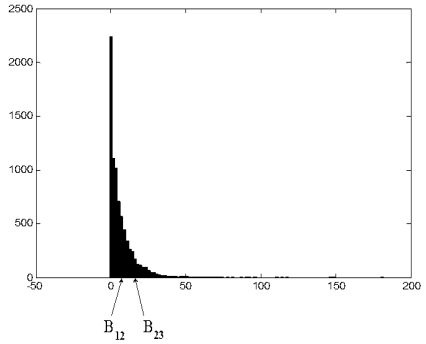

Figure 3. The histogram of the rate of commercial profitability for medium-size enterprises

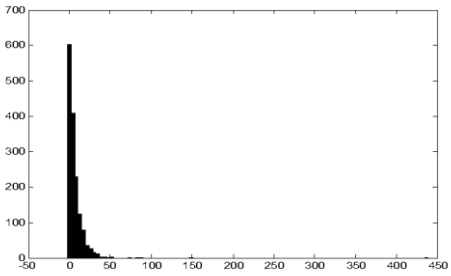

Figure 4. The histogram of the rate of commercial profitability for large enterprises

Firstly, we built the neuro-classifier system, which was employed in order to capture distinctive aspects of the dataset. The respective system includes a feedforward artificial neural network (ANN) of which a descriptive diagram is presented in figure 5. There are several features especially designed in order to obtain maximum performance by capturing some properties of the dataset. The inputs are preprocessed by normalization, and then the model includes a principal component analysis block (PCA) (Jolliffe, 2002) in order to reduce the dimension of input space.

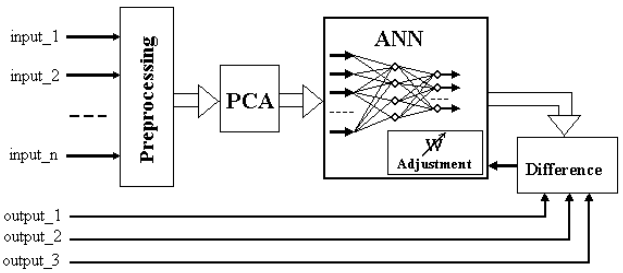

Figure 5. The ANN classification model during the learning process

We selected as indicators, related to the involved SMEs, various parameters such as the company location – the county, ownership status, regional territorial area, type of economic activity according to international classification (NACE), and a set of balance sheet indicators. If we are going into details, the input variables under consideration were: counties (jj); property type (pp); NACE total fixed assets (AIT); total current assets (ACT); debts: amounts due in a period up to a year (DSA); debts: amounts due in a period larger than one year (DPA); total owner's equity (CPT); total equity (CT); personnel expenses (CHP); total operating expenses (CHE); total financial expenses CHF); total expenses (CHT); corporate tax (IMP); average number of employees (NMP); total assets minus current debts (TAMDC). As previously stated the output is related to the commercial profitability rate (RPC).

In fact, each output of this model corresponds to one specific class (the number of outputs is equal to the number of classes). To be clearer, for a specific configuration of the inputs, when a class is selected, the value of the corresponding output has the value of one, while the rest of outputs are zeros.

In order to find the proper architecture of the two-layers feedforward neural model there is an iterative process, which implies a series of imbricate loops to get the proper number of neurons on each layer. Going into details, we varied the number of neurons for each hidden layer (e.g. between 4 and 10 for the first hidden layer N_{h1}, and from 2 to 8 for the second hidden layer N_{h2}). Moreover, each of the neural architectures was tested several times (about 5 times, depending on the situation) with random initial settings of the weights and different training-validation sets. Finally, we chose the adequate model with respect to the smallest error between the simulated and the desired outputs. This error (E_{tot}) was computed for both training and validation data sets. Consequently, we were able to select the best neural network (separately for each set of borders) from a total of 7×9×5=245 different instances. A supplementary condition was used:

$$E_{val} \leq \frac{6}{5} \cdot E_{tr} \tag{1}$$

because we want to eliminate those combinations which provide a value of the validation set (E_{val}) that is at least 20% greater than the error of the training set (E_{tr}). This way, we can improve the result of the model for the test set, since in our approach, the validation data also acts as a kind of test set (providing pretty similar errors).The volume of medium-sized firms used in the analysis was 7902, for which we selected relevant indicators related to these companies, such as: the company location – the county, ownership status, regional territorial area, type of economic activity according to international classification (NACE), and a set of balance sheet indicators. Remember that we use two borders (B_{12} and B_{23}), where the numerical indices (each pair of digits as subscript of B) consecutively indicate two neighbor classes for a specific delimiter. As an example we started considering B_{12}=0.5 and

B_{23}=5, and then we varied these values or, even expanded the number of classes. To be more explicit, as we can see in figure 3, those firms whose parameter profitability rate varies from zero to B_{12}=0.5 (inclusive) we can consider that fall in the poor class. Then, the class of medium-performance includes all companies which indicate values strictly greater than B_{12}=0.5 till maximum B_{23}=5. And finally, the last group (with best performances) includes firms that have the mentioned parameter strictly greater than B_{23}=5.But there is a practically endless range of possibilities to choose these borders. We may start with a first attempt (as in the previous example) and then there we can study what happens, when one changes the borders between the classes. Different changes of the borders will be taken into consideration. Additionally, we also extend the number of classes to see if it is possible to capture detailed aspects of firm performances. In our attempt, an important issue is to properly classify those companies which are very close to the borders. The challenge is how to capture a suitable correlation between input parameters and the output of the classification model in order to minimize wrong classification, especially in the vicinity of the borders.

Having in mind these circumstances, for the ANN model, it is essential to have enough data for each important process (training, validation and test). It is worth to mention that in our approach, we have randomly split the initial set of data into about 70% for the effective training process, 10% for validation, and 20% for the test set. Otherwise a class with a greater amount of data will influence the result (by capturing the false matching from the less representative classes) and during the test process this will have a negative impact over the results. Therefore, in order to keep this splitting under observation, we established successive trials with specific predefined borders, each of these trials being repeated three times to see if the randomization is effective. These distributions for 6 trials (totalizing 18 sub-trials) are presented in Table 1.

Table 1. Random splitting of the data for each separate trial

			N_{tr}			N_{val}			N_{test}		
	B_{12}	B_{23}	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Trial 1	0.5	5	1434	1850	2338	182	247	331	413	461	646
Trial 1 (second iteration)	0.5	5	1461	1837	2357	174	261	314	394	460	644
Trial 1 (third iteration)	0.5	5	1477	1804	2359	204	253	297	348	501	659
Trial 2	1.5	7	1852	1948	1852	245	264	241	495	514	491
Trial 2 (second iteration)	1.5	7	1843	1928	1872	250	265	238	499	533	474
Trial 2 (third iteration)	1.5	7	1838	1916	1835	268	270	233	486	540	516

	B_{12}	B_{23}	N_{tr}			N_{val}			N_{test}		
			Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Trial 3	1.5	8	1827	2172	1632	242	289	226	523	555	436
Trial 3 (second iteration)	1.5	8	1828	2166	1634	264	274	220	500	576	440
Trial 3 (third iteration)	1.5	8	1886	2151	1630	227	293	225	479	572	439
Trial 4	1.55	6.85	1881	1866	1896	259	254	240	493	514	499
Trial4 (second iteration)	1.55	6.85	1872	1860	1905	252	262	241	509	512	489
Trial 4 (third iteration)	1.55	6.85	1887	1908	1848	239	240	274	507	486	513
Trial 5	2	8	2085	1896	1626	301	246	218	536	544	450
Trial 5 (second iteration)	2	8	2077	1954	1606	277	255	223	568	477	465
Trial 5 (third iteration)	2	8	2120	1904	1631	270	246	233	532	536	430
Trial 6	1	10	1653	2729	1267	203	365	183	455	688	359
Trial 6 (second iteration)	1	10	1633	2666	1311	227	374	163	451	742	335
Trial 6 (third iteration)	1	10	1659	2735	1300	204	371	161	448	676	348

There is a reasonable distribution of the data in all trials. According to Table 1, it seems that both second and four trials show almost equal data partitions for each class. It is worth to mention that the selection of the borders between successive classes was intuitively suggested by specialists in economics and we wanted to see the result of this approach. Then we have to construct an ANN structure for each of these trials.

Table 2. The results of different approaches in which we vary the borders (B_{12} and B_{23}) between classes

	B_{12}	B_{23}	N_{tr}	N_{val}	N_{test}	$P_{Class\ 1}$	$P_{Class\ 2}$	$P_{Class\ 3}$	P_{Test}	N_{h1}	N_{h2}
Trial 1	0.5	5	5622	760	1520	0.2833	0.6356	0.8375	0.6257	8	7
Trial 1 (second iteration)	0.5	5	5655	749	1498	0.0127	0.7217	0.8509	0.5907	6	5
Trial 1 (third iteration)	0.5	5	5640	754	1508	0.2155	0.6766	0.8027	0.6253	8	6

	B_{12}	B_{23}	N_{tr}	N_{val}	N_{test}	$P_{Class\ 1}$	$P_{Class\ 2}$	$P_{Class\ 3}$	P_{Test}	N_{h1}	N_{h2}
Trial 2	1.5	7	5652	750	1500	**0.6020**	**0.5623**	**0.7495**	0.6367	5	8
Trial 2 (second iteration)	1.5	7	5643	753	1506	**0.7014**	**0.5291**	**0.7489**	0.6554	8	8
Trial 2 (third iteration)	1.5	7	5589	771	1542	0.7181	0.1944	0.7519	0.5460	7	5
Trial 3	1.5	8	5631	757	1514	0.3939	0.7135	0.7087	0.6017	5	7
Trial 3 (second iteration)	1.5	8	5628	758	1516	**0.4500**	**0.5556**	**0.7477**	0.5765	5	2
Trial 3 (third iteration)	1.5	8	5667	745	1490	**0.5094**	**0.6625**	**0.7221**	0.6309	4	7
Trial 4	1.55	6.85	5643	753	1506	0.6470	0.4105	0.8036	0.6182	8	7
Trial 4 (second iteration)	1.55	6.85	5637	755	1510	0.6503	0.3047	0.7341	0.5603	9	6
Trial 4 (third iteration)	1.55	6.85	5643	753	1506	0.9309	0.0165	0.8070	0.5936	8	4
Trial 5	2	8	5607	765	1530	0.7462	0.3474	0.6911	0.5882	5	4
Trial 5 (second iteration)	2	8	5637	755	1510	0.6866	0.4361	0.6623	0.6000	10	5
Trial 5 (third iteration)	2	8	5655	749	1498	0.7725	0.3227	0.7255	0.5981	7	7
Trial 6	1	10	5649	751	1502	0	0.8983	0.6825	0.5746	8	2
Trial 6 (second iteration)	1	10	5610	764	1528	0.1375	0.8774	0.5672	0.5910	7	6
Trial 6 (third iteration)	1	10	5694	736	1472	0.1384	0.8920	0.6121	0.5965	6	3

In the Trial 1, after a complete scroll of 245 combinations of ANN architecture and training sets (based on previously described inner loops), we obtained a neural network with N_{h1}=8 and respectively N_{h2}=7 neurons on hidden layers. The initial volume of data was randomly split in three parts: N_{tr}=5622, N_{val}=760 and N_{test}=1520 respectively. Each of these numbers is a sum of the values from each class (according with Table 1). In Table 2, the first line of values shows the main results of this described approach. Here, as previously mentioned, we denote by B_{12} the border point between the first and the second class and with B_{23} the next border point, which splits the second and the third class (the same as in Table 1). Several trials, which are shown in Table 2, use different values of these borders. The probability of the correct results in the each class, during the test phase, is separately represented in the same table, including the global probability of the correct results on the test set. As previously mentioned, each trial, with the same configuration of borders, is repeated two times more, from scratch, with other

random initialization of training-validation-test sets, to see if the probabilities of correct classification are not a result of simply chance. This way we can check the repeatability of the results. It is obvious to see in Table 2, that the volumes of data for each set (N_{tr}, N_{val}, and N_{test}), even if were randomly chosen the data, didn't vary too much in order to affect the percentages of their distributions.

During the second trial we established new borders between classes, having B_{12}=1.5 and B_{23}=8. After first and second iterations, the results were slightly better than those obtained after trial 1. Remember that the main idea was to equally distribute the volume of data between classes during the training process. In order to have a broad view we changed (but not quite dramatically) the borders on each trial. As a consequence, in next trial (Trial 3), there is a change of B_{23} (between second and third classes), but without visible improvement. Similar behavior was identified in the next two trials (4 and 5). Condition has worsened, in trial 6, when we moved even farther the second border.

It became obvious that by moving to the left of the first border (B_{12}) we cannot expect to improve the results since it is obvious that the probability from the first class will decrease dramatically. As a consequence of these results it seems that we have to choose some values of the borders somehow similar with those from trials 1, 2, and 3 or even from the trials 4 and 5. Note that even if the trial 4 shows in Table 1 one of the best distributions of the elements, this is not automatically implying a better result.

Extending the numbers of borders in order to obtain four (or even more) classes wasn't a good choice for the model. In Table 3, we selected one of the best trials (denoted here with number 7) in which we had split the data in four classes.

Table 3. Using a model with three borders (B_{12}, B_{23} and B_{34}) between classes

	B_{12}	B_{23}	B_{34}	$P_{Class\ 1}$	$P_{Class\ 2}$	$P_{Class\ 3}$	$P_{Class\ 4}$	P_{Test}	N_{h1}	N_{h2}
Trial 7	0.4	4	10	0	0.9396	0.0315	0.8172	0.4657	5	2
Trial 7 (second iteration)	0.4	4	10	0.5163	0.5875	0.2332	0.7216	0.5073	6	6
Trial 7 (third iteration)	0.4	4	10	0.2643	0.6286	0.3683	0.7508	0.4940	6	4

But, as we can see in Table 3, the probabilities, to fall in one out of those four classes, have varied dramatically. Even if the second class and the fourth one have got better probabilities, the global probability of correct classification is around 0.5. It is quite a challenge to choose those numbers that separate successive classes.We had checked several combinations of borders but the results were even worse. This shows that the task of choosing the borders is difficult. From the whole results presented here, it seems that some optimal borders were selected in the second and

third trials. When more data will be available then we will include the adaptive retraining procedure (Nastac 2010) in the model. Having more data we also expect to solve the problem of choosing the borders. This will allow us to refine the model each time when new data becomes available. It will be an adaptation in which the old knowledge is not completely forgotten and the classification system will perform better as long as it progressively acquires more experience.In order to validate the results obtained, statistical regression models were used.The econometric analysis is based on the following steps: setting up a theoretic model; specifying the adequate mathematic model for the theoretic model; specifying the econometric model, estimating model parameters; statistic inference for the models and its parameters; using results to argue economic decisions.

In the case of multiple regressions, we need to solve the following issues: identifying the variables for the regression model; defining the hypotheses of the classic regression model and testing them; estimating parameters and validating the model as well as the parameters; launching predictions for the dependent variable based on the model. The linear regression model is based on a set of hypotheses which describe the form of the model and the relationship between variables, the nature of the residual value, etc. The linearity hypothesis for the regression model implies it has the following form:

$$Y_t = \beta_1 + \beta_2 \cdot x_{2t} + ... + \beta_k \cdot x_{kt} + \varepsilon_t, t = \overline{1,n} \quad (2)$$

in original variables, either when using variables which were conveniently transformed. In the context of a regression analysis, linearity refers to the way in which parameters and noise variable are included in the equation, not necessarily the type of function which reflects the dependence between variables. When the noise variable is not zero, we must also identify the cause and include it in the model as an independent variable, and ε will represent only the unknown part. The estimation for model parameters is made through the generalized Ordinary Least Squares (OLS) method. The OLS estimators for β_0, β_1: $b_0$and $b_1$respectively, are given by:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{Sxy}{S^2x}, \qquad b_0 = \bar{y} - b_1\bar{x} \quad (3)$$

The theoretical value of $Y_i$in function by x_i is: $\hat{y}_i = b_0 + b_1 x_i$, and the residual value for the index *i* is: $\hat{e}_i = y_i - \hat{y}_i$

For classic hypotheses, the noise variable is uncorrelated with a constant dispersion. These restrictive conditions can be relaxed and the ε covariance matrix can have a more general form:

$$E(ee'/X) = \delta^2 V \quad (4)$$

in which: δ^2 is positive, finite parameter; *V* – symmetrical matrix.

The generalized OLS method can also be used to obtain the β estimator:

$$b = (X'V^{-1}X)^{-1}X'V^{-1}Y \quad (5)$$

The validation of the model and computing the determination ratio is based by decomposing the data series variation $y_1,...,y_T$ depending on the influence of the factors included in the regression model and the unregistered random factors. The ANOVA model validation method involves the determination of intermediate components, as follows: the total sum of squares $SST = \sum (y_i - \bar{y})^2$ that quantifies the dispersion of the endogenous variable series under the action of all inference factors; the regression sum of squares $SSR = \sum (\bar{y}_i - \bar{y})^2$; and the error sum of squares $SSE = \sum (\hat{y} - y)^2 = \sum e_i^2$, which measures the influence of random and unregistered factors on the variation of y.

The strong influence of SMEs on macroeconomic results is well described by the correlation between the dynamic of the GDP, as macro result indicator, and the rate of new firm creation, as an indicator of the initiative spirit and the business environment.

We remark a parallel evolution of GDP computed towards 1995 (IPIB) and the rate of new firm creation (RCNF), described by the equation estimated based on data for 17 years:

$$IPIBi = 72.939 + 1.827 \cdot RNCFi \quad (6)$$

The results of estimating the regression equation (6) are valid statistical, t-test probabilities of being lower than the level of significance ($P<\alpha=0.05$). The parameter value estimated explanatory variable and indicated that a 1% increase in the rate of new firm creation leads to an average increase of 1.83% of GDP compared to 1995. Tests confirm the validity of the regression model (F=43.24, t=6.54, P_{value}=0.000009).

The variation in the GDP dynamic is explained to a large extent of 73%, by the variation in the creation of new private firms (Adjusted **R** Square = 0.73), while the covariance is 206.34. The equality between the correlation and the linear correlation coefficient confirms the correctness of the linear regression model (Multiple R and Pearson Correlation = 0.86).

The hypothesis of normal distribution of the residual variable is verified by the computations and the analyses of the statistics referring to its distribution, the value of Skewness being close to 0 while the value of Kurtosis is 2.2, thus indicating an almost symmetrical and slightly arched distribution, the Jarque-Bera test also indicating that this distribution is relatively normal (Bera and Jarque, 1981), by the fact that the probability associated to the test is higher than the chosen significance threshold ($0.8>\alpha=0.05$), which leads to the acceptance of the null hypothesis that the distribution is normal. Also, the hypothesis regarding errors' homoscedasticity was verified and confirmed.

The depth of the analysis on influencing factors is given by the attempt to identify the causes behind SME performance. Exogenous variables used in the regression models have been previously defined.All data are taken from the financial reports of medium-sized firms (with 50-249 employees).

Table 4.ANOVA TEST

ANOVA[a,b] for RPC						
Model		***Sum of Squares***	***df***	***Mean Square***	***F***	***Sig.***
1	Regression	2.342	12	0.195	72.555	0.000
	Residual	21.219	7888	0.003		
	Total	23.561	7900			
a. Predictors: (Constant), NMP, DPA, CT, DSA, IMP, CHF, CHP, CHE, AIT, ACT, CPT, TAMDC						
b. Dependent Variable: RPC; Excluded Variables: CHT						

The endogenous variable, the commercial profitability rate (RPC), was determined as percentage between the operating profit and total turnover. In preparing the data, all variables were normalized using two methods (Myatt, 2007). In the first case, we used the minimum and maximum, transposing the values in an interval, e.g. 0 to 1, or -1 to 0. In the second case, we used the Z score, normalizing the values around the average, with the standard deviation as the alternative.

Table 5. **SUMMARY OF RESULTS**

Coefficients[a]						
		Unstandardized Coefficients		***Standardized Coefficients***		
Model		***B***	***Std. Error***	***Beta***	***t***	***Sig.***
1	(Constant)	0.034	0.001		26.048	0.000
	CHP	-0.047	0.011	-0.051	-4.325	0.000
	CHE	-0.865	0.041	-0.325	-21.284	0.000
	CHF	0.074	0.029	0.030	2.576	0.010
	IMP	0.774	0.027	0.423	28.639	0.000
[a)] Dependent Variable: RPC; Excluded Variables: CHT						

We established that between the exogenous and endogenous variables, there are no significant correlations, the Pearson correlation coefficient being between 0.19 and -0.19. The results of the regression model which involve all independent variables (Table 4), although statistically significant ($F=72.555$; $p<0.0005$), show the fact that the exogenous variables taken into consideration explain to a small extent the endogenous variable (R square = 0.099).Next, eliminate variables with little influence and composed a reduced model with the following independent

variables (CHP - personnel expenditure, operating expenses - CHE, financial expenses - CHF and profit tax-IMP), the results are relatively similar.

The model is statistically significant (F=210.013; p<0.0005), but the variables give a weak explanation of the commercial profitability rate (R square = 0.096). The coefficients (Table 5) are statistically significant different from zero (p<0.05), and the general form of the model which estimates the commercial profitability rate is:

$$RPC = 0.034 - 0.407 \cdot CHP - 0.865 \cdot CHE + 0.074 \cdot CHF + 0.774 \cdot IMP \quad (7)$$

These final results confirm a strong dependence between the performance indicators and the balance sheets' information, thus validating the initial results that were obtained when using the ANN model.

4. Conclusions

We have classified companies based on their profitability. We envisage that when more data will be available to include the adaptive retraining procedure. This will allow us to refine the model each time when new data becomes available. It will be an adaptation in which the old knowledge is not completely forgotten and the classification system will perform better as long as it progressively acquires more experience.

Furthermore, when more data will be available, then we also expect to solve the problem of choosing the borders. It is worth to remember that the selection of the borders between successive classes was intuitively suggested by specialists in economics and we saw the result of this approach especially on the second and the third trials (where we obtained best results). But we expect to find even a better selection of these borders. The total number of possibility is practically infinite and the searching for an optimal solution could be improved by using a genetic algorithm starting with a set of population defined by these trials, combined with constructing an ANN structure for each of them.

For the present data, the neural classification system was able to identify the profitability rate as long as there are no more than three classes. This model is very flexible and can be easily adapted for any possible country. As a conclusion of these results, it is difficult to extract a higher classification model by using these data. It seems that there might be necessary to find further information in order to obtain a better separation between classes (more than three). The model may include, through a future extension, many new supplementary parameters. Having an extended number of inputs with economic relevance could lead to an improvement of the results. Our future goal is to include in the neural model the previously mentioned retraining mechanism which implies a huge amount of historical data from previous years. This way it will finally result not only an intelligent system but also an adaptive one, which can be easily retrained on successive predefined intervals of time.

REFERENCES

[1] **Altman, E.I. (1968),** ***Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy****; The Journal of Finance* Vol. 23, No.4, pp. 589-609;
[2] **Beaver, W.H. (1966),** ***Financial Ratios as Predictors of Failure, Empirical Research in Accounting: Selected Studies****; Journal of Accounting Research*, Vol. 4, pp. 71-111;
[3] **Bera,A.K.Jarque,C.M.(1981),** ***Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals: Monte Carlo Evidence****; Economics Letters,*7(4): 313–318;
[4] **Chou, C.H., Lin, C.C., Liu, Y.H., and Chang, F. (2006),** ***A Prototype Classification Method and Its Use in a Hybrid Solution for Multiclass Pattern Recognition;*** *Pattern Recognition,* Vol. 39, Issue 4, April 2006, pp. 624–634;
[5] **Dragan, I. M., Isaic-Maniu, Al (2012),** ***Performance of Small and Medium Enterprises****; LAP Lambert Academic Publishing*, Germany;
[6] **Dragan I. M., Isaic-Maniu Al. (2013),** ***A Barometer of Entrepreneurial Dynamics above the Crisis; Revista Romana de Statistica***, No.7, pp.42-64;
[7] **Frydman H., Altman E.I. and Kao D.L. (1985),** ***Introducing Recursive Partitioning for Financial Classification :The Case of Financial Distress****; The Journal of Finance,* V.XL, No.1,March 1985, pp. 269-291;
[8] **Henzel, S.R. and Rengel, M. (2013),** ***The Macroeconomic Impact of Economic Uncertainty: A Common Factor Analysis***. Munich: IFO Institute;
[9] **Jolliffe I.T. (2002),** ***Principal Component Analysis***; 2nd edition, *Springer, New York;*
[10] **Lejárraga, I., Lopez, H., Oberhofer, H., Stone, S., Shepherd B. (2014*), Small and Medium-Sized Enterprises in Global Markets: A Differential Approach for Services*?** Doc. No. 165, OECD Trade Policy Papers from OECD Publishing;
[11] **Mateev, M. and Anastasov, Y. (2010),** ***Determinants of Small and Medium Sized Fast Growing Enterprises in Central and Eastern Europe: A Panel Data Analysis****; Financial Theory and Practice*. 34(3), pp. 269-295;
[12] **Myatt, G. J. (2007),** ***Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining****; John Wiley*;
[13] **Nastac, I. (2010),** ***An Adaptive Forecasting Intelligent Model for Nonstationary Time Series****; Journal of Applied Operational Research*, Vol. 2, 2010, No. 2, pp. 117–129;
[14] **Nastac,I.,Bacivarov,A. and Costea, A. (2009),** ***A Neuro-Classification Model for Socio-Technical Systems****;* Romanian Journal of Economic Forecasting, Vol. XI, No. 3/ 2009, pp. 100-109;
[15] **Nastac, D.I.; Dragan, I.M.; Isaic-Maniu, A. (2014),** ***Estimating Profitability Using a Neural Classification Tool***; IEEE Proceedings of NEUREL 2014, Belgrade, Serbia, 25-27 November 2014, pp. 111-114;

[16] **Nastac, D.I.; Dragan, I.M.; Isaic-Maniu, A. (2016),** ***Profitability Analysis of Small and Medium Enterprises in Romania Using Neural and Econometric Tools*****;** Proceedings of the NSAIS'16 Workshop on Adaptive and Intelligent Systems, August 2016, pp. 62-70;
[17] **Ripley, B.D. (1994),** ***Neural Networks and Related Methods for Classifications****; Journal of the Royal Statistical Society*. Series B (Methodological), Vol. 56, No. 3, pp. 409-456;
[18] **Subhan Q.A., Mehmood, R. M.,Sattar, A. (2013***), Innovation in Small and Medium Enterprises(SME's) and Its Impact on Economic Development in Pakistan;* Proceedings of 6th International Business and Social Sciences Research Conference 3 – 4 January, 2013, Dubai, UAE;
[19] **Tang, Z., Tang, J.T. (2012),** ***Stakeholder–firm Power Difference, Stakeholders' CSR Orientation and SMEs' Environmental Performance in China****; Journal of Business Venturing*, vol. 27, no. 4, pp. 436-455;
[20] **Terdpaopong, K. (2011),** ***Identifying an SME's Debt Crisis Potential by Using Logistic Regression Analysis****. RJAS- Rangsit Journal of Arts and Sciences*, Vol. 1 (1), pp.17-26;
[21] **European Commission (2014), Annual Report on European SMEs 2013/2014 – A Partial and Fragile Recovery Final Report** -July 2014 SMEs *Performance Review;*
[22] **European Commission (2016), Annual Report on European SMEs, 2014 / 2015**,**SMEs start hiring again**;
[23] **Post-Privatization Foundation (2013), Post-Privatization Foundation Report on the SME Sector in Romania**.