COMBINING PCA ANALYSIS AND ARTIFICIAL NEURAL NETWORKS IN MODELLING ENTREPRENEURIAL INTENTIONS OF STUDENTS

Marijana Zekić-Sušac
University of J.J. Strossmayer in Osijek, Faculty of Economics
Gajev trg 7, 31000 Osijek, Croatia
E-mail: marijana@efos.hr

Nataša Šarlija
University of J.J. Strossmayer in Osijek, Faculty of Economics
Gajev trg 7, 31000 Osijek, Croatia
E-mail: natasa@efos.hr

Sanja Pfeifer
University of J.J. Strossmayer in Osijek, Faculty of Economics
Gajev trg 7, 31000 Osijek, Croatia
E-mail: pfeifer@efos.hr

Abstract

Despite increased interest in the entrepreneurial intentions and career choices of young adults, reliable prediction models are yet to be developed. Two nonparametric methods were used in this paper to model entrepreneurial intentions: principal component analysis (PCA) and artificial neural networks (ANNs). PCA was used to perform feature extraction in the first stage of modelling, while artificial neural networks were used to classify students according to their entrepreneurial intentions in the second stage. Four modelling strategies were tested in order to find the most efficient model. Dataset was collected in an international survey on entrepreneurship self-efficacy and identity. Variables describe students' demographics, education, attitudes, social and cultural norms, self-efficacy and other characteristics. The research reveals benefits from the combination of the PCA and ANNs in modeling entrepreneurial intentions, and provides some ideas for further research.

Key words: *Classification, Entrepreneurial intentions, Modelling, Artificial neural networks, Principal component analysis*

1. INTRODUCTION

Due to the fact that several other authors reported the ability of principal component analysis (PCA) to reduce the dimension of input space in the artificial neural network (ANN) models while lowering the training time and preserving or even improving the ANN model accuracy, the aim of this paper was to investigate if the PCA will be effective in reducing the number of predictors of entrepreneurial intentions of students. The purpose is to model entrepreneurial intentions of students by focusing not only in prediction accuracy but also in finding the most efficient model regarding the cost of data collection and training. The survey was conducted at a Croatian university including a sample of students at the first year of the undergraduate, and the first year of the graduate study.

Policy makers, together with the educational program designers recognize entrepreneurship as a behavioral pattern that can be influenced by formal education (Henry et al. 2005). If more potential entrepreneurs were identified and cultivated throughout educational process national economy could expect more successful start-ups, new jobs openings and more economic growth (Gerry et al. 2008). Therefore, the entrepreneurial intentions of students should be investigated in depth.

2. OVERVIEW OF PREVIOUS RESEARCH

A number of research has been done on using the PCA as a preprocessor to the usual multilayer perceptron ANN in different areas. An ANN for classification and a PCA in the preprocessing stage was used by O'Farrella et al. (2005) to classify the quality of food products by a feedforward ANN with one hidden layer and the backpropagation algorithm. The PCA as a feature extractor was applied to spectral data before training the ANN to reduce the amount of redundant information. Bucinski et al. (2005) combined PCA and ANN in medicine by using a backpropagation ANN to classify patients into two categories, and PCA to extract some important features predictive for patients' survival. Sousa et al. (2007) use PCA analysis with varimax rotation to extract factors using ozone concentrations data. The results showed that principal components as inputs improved MLR and ANN models' prediction by eliminating data collinearity and the number of predictor variables. Ravi and Pramodh (2008) were the first who placed principal components instead of the hidden layer of a multilayer perceptron (MLP) network. They proposed a new principal component neural network (PCNN) architecture to solve bankruptcy prediction problem in commercial banks. They replaced the hidden layer of an ANN by a 'principal component layer' which consists of a few selected principal components that perform the function of hidden nodes. The advantage of such architecture is that it reduces the number of weights by eliminating connections between the input layer and the principal component layer. Liu et al. (2007) also used a hybrid approach by combining a Kung and

Diamantaras's artificial neural network with the adaptive principal components extraction (APEX) algorithm.

Our previous work showed that non-linear machine learning methods such as ANNs could be efficient in the area of modeling entrepreneurial intentions of students (Zekic-Susac et al., 2010). The majority of research on career choices of students proposes a huge number of personal inputs that can interact on a variety of levels and directions. It has been presumed that students attitudes, values and career choices can be sufficiently well represented by the following groups of variables: (1) entrepreneurial intentions (Thompson, 2009), (2) altruistic values and empathy (Smith, 2009), (3) subjective norms (Kolvereid and Isaksen, 2006), (4) entrepreneurial self-efficacy (McGee et al., 2009), (4) allocentrism/idiocentrism (Triandis and Gelfand 1998), (5) prior family business exposure (Carr and Sequeira, 2007), (6) entrepreneurial outcome expectations (Krueger, 2000), (7) strength of entrepreneur identity aspiration (Farmer et al. 2009), and (8) social entrepreneurship self-efficacy (Nga, 2010). The career choice theories usually reduces the number of inputs by using construct measures by which the each group of predictors is represented by its construct.

This paper aims to contribute to the methodological aspect of entrepreneurial intentions research by assessing the impact of the different modeling strategies on classifying students' intentions.

3. METHODOLOGY

3.1. Principal component analysis

The aim of PCA is to explain the interdependence of a large number of variables by a smaller number of fundamental or latent variables, i.e. dimensions. Here we bring the formulation of PCA according to Kara and Direngali (2007). For a given *p*-dimensional data set *X*, the *m* principal axes $T_1, T_2, \dots, T_m$, where $1 \leq m \leq p$, are orthonormal axes onto which the retained variance is maximum in the projected space. Generally, matrix *T* can be given by the *m* leading eigenvectors of the sample covariance matrix (Kara and Direngali, 2007):

$$S = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^T (x_i - \mu) \tag{1}$$

where $x_i \in X, \mu$ is the sample mean and *N* is the number of samples, so that

$$ST_i = l_i T_i, i \in 1, \dots, m, \tag{2}$$

where l_i is *i*-th largest eigenvalue of S. The *m* principal components of a given observation vector $x \in X$ are given by:

$$y = [y_1, y_2, \dots, y_m] = [T_1^T x, T_2^T, \dots, T_m^T] = T^T x. \tag{3}$$

The *m* principal components of *x* are decorrelated in the projected space. If taking these principal components as the training set, it will be possible to save computer time and memory, because the size of *y* is much smaller than the size of *x*.

Before entering the PCA in our research data were standardized such that they have zero mean. Exploratory PCA was conducted on the basis of covariance matrix, and principal components were obtained. Kaiser criteria for selecting the number of components was applied in our research according to which the principal components with eigenvalues greater than 1 are retained for further analysis. After extraction the principal components, the varimax rotation was performed in order to get factors linearly more independent. PCA was conducted by using StatSoft Statistica software version 8. In order to test the validity of PCs, a split-sample procedure was conducted, such that the train sample is divided into two subsamples, and PCA was conducted on both subsamples to check the validity of its results. After the validity test, factor score coefficients are obtained on the basis of the train data set, and then the scores were computed separatelly for the validation set as suggested by Ravi and Pramodh (2008).

3.2. Artificial neural networks

Artificial neural networks (ANNs) have been successfully used for classification, prediction, and association in different problem domains (Paliwal and Kumar, 2009). ANNs have the ability to approximate any nonlinear mathematical function, which is useful especially when the relationship between the variables is not known or is complex (Masters, 1995). However, there are some limitations of ANNs such as time-consuming experimentation needed to determine network structure and learning parameters, and a lack of interpretability of the weights obtained during the model building process. The most common type of ANN was tested in this research - the multilayer perceptron (MLP), a feed forward network that can use various algorithms to minimize the objective function, such as backpropagation, conjugate gradient, and other. A simplified architecture of a MLP ANN is presented in Figure 1.

The input layer of an ANN consists of *n* input units with values $x_i \in R$, *i=1,2,..., n*, and randomly determined initial weights w_i usually from the interval [-1,1]. Each unit in the hidden (middle) layer receives the weighted sum of all x_i values as the input. The output of the hidden layer denoted as y_c is computed by summing the inputs multiplied with their weights, according to:

$$y_c = f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{4}$$

where f is the activation function selected by the user (sigmoid, tangent hyperbolic, exponential, linear, step or other) (Masters, 1995).

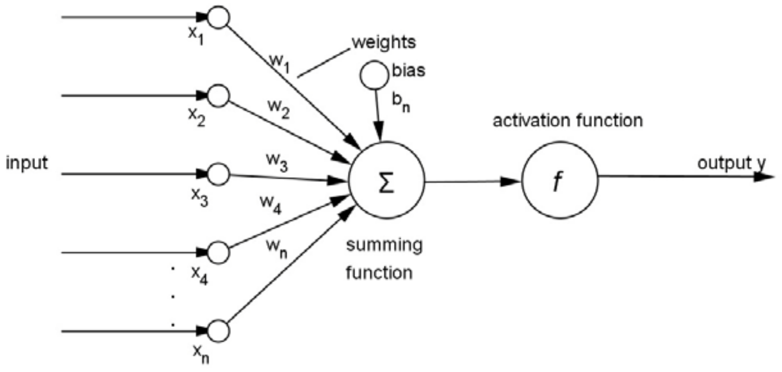


Figure 1. Architecture of the MLP network (Haykin, 1999, modified)

The computed output is compared to the actual output y_a, and the local error ε is computed. The error is then used to adjust the weights of the input vector according to a learning rule, usually the Delta rule according to:

$$\Delta w_i = \eta \cdot y_c \cdot \varepsilon \tag{6}$$

where Δw_i is the weight adjustment, η is the learning parameter that could be experimentally determined. The above process is repeated in a number of iterations (epochs), where the three different algorithm were tested to minimize the error: gradient descent, conjugate gradient descent, and Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. Conjugate gradient descent is faster than gradient descent and performs a series of line searches through error space, therefore avoiding a local minima. BFGS belongs to the second-order algorithms with very fast convergence but memory intensive due to storing the Hessian matrix (Dai, 2002). In order to produce probabilities in the output layer, a softmax activation function is added. The output layer of all ANN models in our experiments consisted of a binary variable (valued as 1 for the existence of entrepreneurship intention, and 0 for the absence of entrepreneurship intention). The number of hidden units varied from 2 to 20, and the training time is determined in an early-stopping procedure which iteratively trains and tests the networks on a separate test sample in a number of cycles, and saves the network which produces the lowest error on the test sample.

3.4. Combining PCA with artificial neural networks – modelling strategy

The modeling strategy was led by the idea to find the most efficient ANN model that will be able to classify students according to their entrepreneurial intentions with a satisfactory accuracy but with a minimal number of predictors needed as the input to the model. The aims were to lower the training time of the ANN, and to lower the time needed for the respondents to fill up the survey, which means to find the minimal number of items needed to predict the output variable. Therefore, the strategy included the creation of 4 models:

(1) Model 1 – ANN with all available original input variables (94 variables corresponding to 94 items in the survey)

(2) Model 2 – ANN model obtained by constructing feature subsets (constructs of variables) on the basis of Cronbach alpha test (The variables that explain the same instrument were grouped together into one construct variable if the Cronbach alpha is greater than 0,7)

(3) Model 3 – ANN with principal components (PCs) extracted by the PCA as input variables (PCA was used only for continuous variables)

(4) Model 4 – ANN with input variables extracted from rotated factors obtained by PCA, where the most representative variable is selected from each factor and used in ANN model to represent that factor.

Model 1 is assumed to be the most costly one, because of the highest dimension of input space, which requires the longest training time of the ANN model, and also the longest time needed for respondents to complete the survey. Model 2 and Model 3 still retain all the items in the survey needed to be filled up, but could reduce the training time of the ANN. The model which could actually save the time of data collection (i.e. the time needed to complete the survey) and the training time is Model 4.

The performance of all models is measured by the hit rate of class 0 (i.e. the "lack of entrepreneurial intentions") denoted as *hit₀*, or “negative hit rate”, hit rate of class 1 (i.e. the "existence of entrepreneurial intentions") denoted as hit_1 or “positive hit rate”, and the total hit rate (*total hit*) according to:

$$hit_0 = \frac{c_0}{t_0},\ hit_1 = \frac{c_1}{t_1},\ total\ hit = \frac{c_0 + c_1}{t_0 + t_1} \tag{11}$$

where c_0 is the number of students accurately recognized to have output 0, t_0 is the number of students with actual (target) 0 output, c_1 is the number of students accurately recognized to have output 1, and t_1 is the number of students with the actual output 1. All tested models were validated on the same out-of-sample dataset, and a 10-fold cross-validation procedure for testing generalization ability of the models was conducted. The cross-validation procedure (or leave *k* cases out, where *k*=1/10 of the total sample) is used in this paper because it produces no statistical bias of the result since each tested sample is not the member of the training set. After the 10-fold cross-validation procedure, the average of the total classification rate (i.e. hit rate) is computed, which is used to estimate the generalization error. This estimate is also used as the model selection criterion.

4. DATA

The total dataset consisted of 443 regular students of business administration at the first year of study at University of J.J. Strossmayer in Osijek, Croatia. The survey was conducted at summer semester 2010 and in 2012. Besides students' demographics, other predictors were used and grouped into eight groups according to previous research described in section 2. There were 48,76% of respondents with intentions to start a business, and 51,24% of them with no intentions to start a business. The total number of 94 input variables was used in Model 1 which consists of all available variables, while the input dimension in other models was reduced according to modeling strategy. For the purposes of model training and testing, the total dataset is divided into three subsamples: train, test and validation subsample in the ANN models. The structure of samples is presented in Table 2.

Table 2: Sample structure

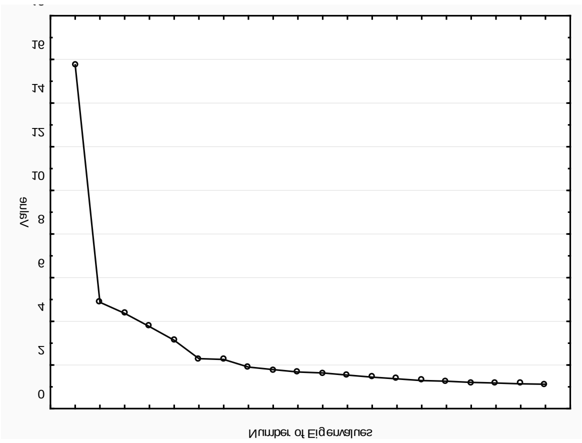
Subsample	1 (existence of intentions to start business)	0 (lack of intentions to start business)	Total	%
Train	171	184	355	81.90
Test	21	23	44	10.16
Validation	24	20	44	10.16
Total	216	227	443	100.00

5. RESULTS

5.1. Results of principal component analysis

The PCA obtained on the covariance matrix revealed that 22 components with eigenvalue greater than 1 can be produced from the training data. The eigenvalues together explain 65,58% of total variance. The first principal component explains the highest percentage of total variance (18,11%), while the second and the third component explain around 5% of total variance each. The validity of principal components is tested by splitting the trainig sample in two subsamples, and the scree plots of eigenvalues obtained on the training sample is presented in Figure 2.

Figure2. *Scree plot of the principal component eigenvalues obtained on the train sample*



The communalities of first 22 factors are greater than 0,5, and the pattern of the factor loadings is the same in both subsamples. Based on factor score coefficients, the factor scores are then computed for the validation sample and 22 principal components are used in the ANN Model 3.

5.2. Result of the artificial neural network models

After testing various ANN architectures by varying the activation function, number of hidden units, and learning algorithm, the MLP with logistic activation function and BFGS algorithm produced the best classification rate obtained on the validation sample in all four tested models. The results of the four ANN models are presented in Table 4.

Table 4: The results of the ANN models obtained on the 10-fold cross-validation procedure

Validation sample		Average hit rates obtained on 10 different validation samples					
		Class 1 - Existence of intentions		Class 0 - Lack of intentions (%)		Total hit rate (%)	
ANN model	ANN type & architecture	Ave.hit rate (%)	St.dev. of ave.hit rates	Ave.hit rate (%)	St.dev. of ave.hit rates	Ave.hit rate (%)	St.dev. of ave.hit rates
ANN Model 1	MLP 115-3-2	84,39	8,25	69,02	8,30	**77,97***	6,96
ANN Model 2	MLP 34-16-2	70,66	13,11	67,32	12,06	**69,32**	7,96
ANN Model 3	MLP 44-20-2	69,62	10,31	73,99	5,99	**72,27**	4,89
ANN Model 4	MLP 44-16-2	76,63	13,05	71,57	13,54	**74,09**	6,36

*The highest average hit rate obtained on 10 validation samples

The bold values in Tables 4 and 5 indicate the average total hit rates of each model that could be used to compare the model accuracy. It can be seen that Model 1 produced the highest average hit rate obtained on 10 different validation samples (77,97%), followed by Model 4 (74,09%). The lowest accuracy is obtained by Model 2 which uses construct variables based on Cronbach alpha (69,32%). The lowest standard deviation i.e. the best stability is observed for Model 3 based on PCs as input variables, while the Model 2 is also the worst regarding the stability of its result.

The average training time highly depends on the number of input variables, and is the highest in Model 1 (28,87 seconds), followed by Model 4 (12,01 seconds) and Model 3 (10,39 seconds), while the lowest training time is observed in Model 2 (7,36 seconds) with construct variables (experiments conducted by a computer with Intel Core i5 2410 M processor at 2,3 GHz and 6 GB RAM).

In order to test the significance among the accuracy of the four models, the t-test of difference in proportion was conducted, with the following results (n=44, the size of the validation sample). The results show that there is no statistically significant difference between the total hit rate of any of the four models. However, the accuracy of the models could also be observed based on the hit rate of the individual classes of output variable. Due to the fact that is more important to correctly recognize students with the existence of entrepreneurial intentions, the classification rate of class 1 could also be compared across models. The t-test shows that there is a significant difference on the 10% level between the Model 1 and Model 2 hit rates of class 1 (p=0,0614), and Model 1 and Model 3 (p=0,0498). Therefore Model 1 more accurately recognizes students with entrepreneurial intentions than Model 2 and Model 3, while the difference between Model 1 and Model 4 is not significant.

Regarding the above, the guidelines for selecting the most effective model could be the following:

- since the accuracy of Model 1 in recognizing existence of entrepreneurial intentions is significantly higher than the accuracy of Model 2 and Model 3, it can be concluded that Model 2 (with construct as input variables) and Model 3 (with PCs as input variables) are lower in performance than Model 1 (with all available variables). Model 2 and Model 3 do contribute in lowering the training time, but also significantly lower the model accuracy
- since the accuracy of Model 1 and Model 4 does not significantly differ, the time of data collection and training time should be considered as criteria for further selection. Due to the fact that the number of input items needed for completing the survey is lower in Model 4 (only 28 items are needed for Model 4 comparing to 94 items for Model 1), which also implies the lower training time of the ANN, the Model 4 could be suggested as the most efficient among the tested four models on the observed dataset.

5.3. Extraction of important features for modelling entrepreneurial intentions

The suggested modelling strategy enables two phases of feature extraction: one in the pre-processing stage before running ANN models (based on Cronbach alpha and PCA), and another one in the post-processing by performing the sensitivity analysis of the ANN models. All previously recorded instruments for measuring features of entrepreneurial intentions except allocentrism were also identified in this research. The PCA results were not so consistent to other authors' research in this area, showing a higher dispersion of variables across 21 factors with eigenvalues greater than 1. The most important factor 1 which explains the largest proportion of variance is the factor with high correlations with the variables describing social norms.

The ANN Model 1 that uses all input variables produces the highest sensitivity ratio for variables describing social norms, prior family business exposure to entrepreneurship, gender, and major of study. The sensitivity analysis performed on the ANN Model 4 shows that the variables describing prior family business exposure to entrepreneurship, entrepreneurial identity, gender and major of study have the highest importance in the ANN Model 5, which is similar to predictors extracted by Model 1 with all available variables.

6. CONCLUSION

The paper deals with modelling entrepreneurial intentions of students by PCA and artificial neural network methodology. The input space included nine groups of predictors identified in previous research. PCA was used in the pre-processing stage to extract the important factors, which were then included in the ANN model with some additional categorical variables. This model is compared to other three which uses different strategies of selecting input variables. The results show that the model with variables extracted from principal component factors was the most efficient regarding the criteria of accuracy in recognizing the students that have entrepreneurial intentions, and regarding the time needed for data collection and model training. Since this is a preliminary research, the future plans include testing the methodology on more datasets in order to generalize the conclusions and produce a model that could be implemented at universities for recognizing and bringing more attention to entrepreneurial intentions of students.

Acknowledgements

This work was supported by the Ministry of Science, Education and Sports, Republic of Croatia, through research grant 010-0101195-0872 "Transformation of entrepreneurial potential into entrepreneurial behavior" and research grant 010-0101195-1048 "Models for risk evaluation of enterprises".

REFERENCES

Bucinski, A., Baczek, T., Wasniewski, T., and Stefanowicz, M. (2005), "Clinical data analysis with the use of artificial neural networks (ANN) and principal component analysis (PCA) of patients with endometrial carcinoma", *Reports on Practical Oncology and Radiotherapy*, Vol. 10, pp. 239-248.

Carr, J.C. and Sequeira, J.M. (2007), "Prior family business exposure as intergenerational entrepreneurial intent: A theory of planned behavior approach", *Journal of Business Research,* 60, pp.1090-1098.

Dai, Y-H., (2002), "Convergence properties of the BFGS algorithm", *SIAM Journal of Optimization*, Vol. 13, No. 3, pp. 693-701.

Farmer, S.M., X. Yao and K. Kung-Mcintyre, (2009), "The behavioral impact of entrepreneur identity aspiration and prior entrepreneurial experience", *Entrepreneurship Theory and Practice,* Volume 35, pp.245-273.

Gerry, C., Marques, C.S., Nogueira F. (2008), "Tracking student entrepreneurial potential: personal attributes and the propensity for business start-ups after graduation in a Portugese University", *Problems and Perspectives in Management*, Vol. 6 (4), pp.45-53.

Henry, C. et al. (2005), "Entrepreneurship Education and Training: Can Entrepreneurship be Taught: Part I-II", *Education and Training,* Vol. 47 (2-3), pp.98-111.

Kara, S. and Direngali, F. (2007), "A system to diagnose atherosclerosis via wavelet transforms, principal component analysis and artificial neural networks", *Expert Systems with Applications*, Vol. 32, pp. 632–640.

Kolvereid, L. and Isaksen E. (2006), "New business start-up and subsequent entry into self-employment", *Journal of Business Venturing,* Vol. 21, pp.866-885.

Krueger, N.F. Jr. (2000), "The Cognitive Infrastructure of Opportunity Emergence", *Entrepreneurship: Theory and Practice,* Vol. 24 (3), pp.5-23.

Liu, G., Yi, Z., Yang, S. (2007) , "A hierarchical intrusion detection model based on the PCA neural networks", *Neurocomputing*, Vol. 70, pp. 1561–1568.

Masters, T. (1995), *Advanced Algorithms for Neural Networks, A C++ Sourcebook,* John Wiley & Sons, Inc., New York, USA.

McGee, J., Peterson, M., Mueller, S. and Sequeira, J.M. (2009), "Entrepreneurial self-efficacy: Refining the measure and examining its relationship to attitudes toward venturing and nascent entrepreneurship", *Entrepreneurship Theory and Practice*, Vol. 33(4), pp. 965-988.

Nga, J., and Shamuganathan, G. (2010), "The influence of personality traits and demographic factors on social entrepreneurship start up intentions", *Journal of Business Ethics,* Vol. 95, pp. 259-282.

O'Farrella, M., Lewisa, E., Flanagana, C., Lyonsa, W.B., Jackman, N. (2005), "Combining principal component analysis with an artificial neural network to perform online quality assessment of food as it cooks in a large-scale industrial oven", *Sensors and Actuators B*, Vol. 107, pp. 104–112.

Paliwal, M. and Kumar U.A. (2009), "Neural networks and statistical techniques: A review of applications", *Expert Systems with Applications*, Vol. 36, pp. 2–17.

Ravi, V., Pramodh, C. (2008), "Threshold accepting trained principal component neural network and feature subset selection: Application to bankruptcy prediction in banks", *Applied Soft Computing*, Vol. 8, pp. 1539–1548.

Smith, T.W. (2009), "Altruism and Empathy in America: Trends and Correlates", *National Opinion Research Center/University of Chicago,* Chicago.

Sousa, S.I.V. , Martins, F.G. , Alvim-Ferraz, M.C.M., Pereira, M.C. (2007), "Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations", *Environmental Modelling & Software,* Vol. 22, pp. 97-103.

Thompson, E.R. (2009), "Individual entrepreneurial intent: Construct clarification and development of an internationally reliable metric", *Entrepreneurship Theory and Practice,* 33, pp. 669-694.

Triandis, H.C., and M.J. Gelfand, (1998), "Converging Measurement of Horizontal and Vertical Individualism and Collectivism", *Journal of Personality and Social Psychology*, 74, pp.118-128.

Zekic-Susac, M. Pfeifer, S., Djurdjevic, I. (2010), "Classification of entrepreneurial intentions by neural networks, decision trees and support vector machines", *Croatian Operational Research Review*, Vol. 1, pp. 62-73.