Croatian Review of Economic, Business and Social Statistics (CREBSS)
UDK: 33;519,2; DOI: 10.1515/crebss; ISSN 1849-8531 (Print); ISSN 2459-5616 (Online)
Vol. 4, No. 2, 2018, pp. 57-66

sciendo

Cluster analysis and artificial neural networks in predicting energy efficiency of public buildings as a cost-saving approach

Marijana Zekić-Sušac
Faculty of Economics in Osijek, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia
marijana@efos.hr

Rudolf Scitovski
Department of Mathematics, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia
scitowsk@mathos.hr

Adela Has
Faculty of Economics in Osijek, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia
adela.has@efos.hr

Abstract

Although energy efficiency is a hot topic in the context of global climate change, in the European Union directives and in national energy policies, methodology for estimating energy efficiency still relies on standard techniques defined by experts in the field. Recent research shows a potential of machine learning methods that can produce models to assess energy efficiency based on available previous data. In this paper, we analyse a real dataset of public buildings in Croatia, extract their most important features based on the correlation analysis and chi-square tests, cluster the buildings based on three selected features, and create a prediction model of energy efficiency for each cluster of buildings using the artificial neural network (ANN) methodology. The main objective of this research was to investigate whether a clustering procedure improves the accuracy of a neural network prediction model or not. For that purpose, the symmetric mean average percentage error (SMAPE) was used to compare the accuracy of the initial prediction model obtained on the whole dataset and the separate models obtained on each cluster. The results show that the clustering procedure has not increased the prediction accuracy of the models. Those preliminary findings can be used to set goals for future research, which can be focused on estimating clusters using more features, conducted more extensive variable reduction, and testing more machine learning algorithms to obtain more accurate models which will enable reducing costs in the public sector.

Keywords: artificial neural networks, clustering, energy efficiency, machine learning, prediction model.

JEL classification: C52, C53, C55, F64.
DOI: 10.2478/crebss-2018-0013

Received: June 01, 2018
Accepted: October 30, 2018

Acknowledgment: This work was supported by Croatian Science Foundation through research grant IP-2016-06-8350 "Methodological framework for efficient energy management by intelligent data analytics" MERIDA.

Introduction

Scientific efforts in predicting energy efficiency and consumption align with the European Commission directives about reducing greenhouse gas emissions, increasing energy efficiency and using 20% of energy from renewable resources until 2020. The largest individual energy consumer is the building sector, which contains 40% of total primary energy consumption (Tommerup et al., 2007). That stresses out the importance of creating efficient models that will be able to extract features and predict the energy efficiency level of a building according to its characteristics and planned reconstruction measures. This work deals with creating such prediction model by using data from Croatian Agency for Legal Trade and Real Estate Brokerage (APN), which maintains the centralized information system of energy efficiency in public buildings (ISGE). In order to create prediction models, the machine learning methods were used such as clustering and artificial neural networks (ANN). The real dataset of Croatian public buildings with a large number of attributes, which were reduced in pre-modelling phase. The aim of this research was to investigate the effect of a clustering procedure on ANN model accuracy. For that purpose, the symmetric mean average percentage error (MAPE) of an initial prediction model obtained on the whole dataset is compared to the MAPE of separate models obtained on each cluster, and the results were discussed. Due to a lack of approaches, which integrate clustering and neural networks in modelling energy efficiency, the scientific contribution of this paper is in providing such an approach and in analysing its effects to modelling accuracy. Guidelines for using the model as an approach for saving costs in public sector are also given.

Literature Review

One of the earliest research in methods used to predict energy efficiency is given by Patterson (1996) who defines energy efficiency, and describes its concepts, indicators and methodological issues. Since then, previous research shows a number of efforts to predict energy consumption and cost by mathematical, statistical methods, machine learning and simulations.

An overview of prediction models in the area of energy efficiency of public buildings is conducted by Zekić-Sušac (2017). It reveals that regarding input variables the most of the authors use data on previous consumption, weather data, and building characteristics available in energy certificates or data collected by their own surveys. However, several authors emphasize the importance of occupancy and usage data (Wang, Ding, 2015, Mangold et al., 2015).

Among methodology used in modelling energy efficiency, operations research and statistical methods prevail, while machine learning methods have been only recently tested in this area (Kalogirou, 2006).

Previously, the authors emphasize the potential of integrating statistical, mathematical and machine learning methods, as well. Sabo et al. (2011) have suggested several mathematical models of natural gas consumption. Hsu (2015) used integrated clustering methods for predicting energy consumption of buildings,

using cluster wise regression, k-means and model-based clustering. Naji et al. (2016) used several machine learning methods, such as support vector regression (SVR), adaptive neuro-fuzzy inference system (ANFIS), and applied them in Energy Plus simulation program to predict energy consumption of residential buildings. However, the efficiency of machine learning methods integration in this area is still not investigated enough.

Data and methodology

Data and sampling procedure

The total dataset used in this research consisted of 3659 buildings that had their geospatial, construction, heating, cooling, meteorological, occupational, and energetic characteristics available in the database of the Information System for Energy Management managed by the Agency for Legal Trade and Real Estate Brokerage (APN) in Croatia. Those were the spatial data collected in 2017. In order to create prediction models for estimating energy efficiency level of buildings, only 655 buildings, for which the real output variable was available could be used. After removing outliers, the final data sample consisted of 621 cases. Each sample was divided into the train data (70%) and test data (30%) for modelling purposes. The structure of the samples is presented in Table 1.

Table 1 Sample structure

Sample	No. of cases in the sample	No. of cases in train subsample	No. of cases in test subsample
Total sample before removing outliers	655 (100%)		
Total sample after removing outliers used in clustering procedure	621 (100%)	-	-
Sample used for NN modelling – all buildings that had available output variable	621 (100%)	434 (70%)	187 (30%)
Sample used for NN modelling on cluster 1	322 (100%)	225 (70%)	97 (30%)
Sample used for NN modelling on cluster 2	109 (100%)	76 (70%)	33 (30%)
Sample used for NN modelling on cluster 2	190 (100%)	133 (70%)	57 (30%)

Source: Authors.

The total number of 141 input variables were initially available, while for clustering procedure the 3 most important continuous attributes of buildings were selected by using Pearson correlation. The variables used to cluster buildings and their descriptive statistics are shown in Table 2. The output variable was the relative value of yearly energy needed for heat (QHNDREL) in percentages, which can be used for calculating energy efficiency levels (A, B, C, D, E, F, and G). Due to a large number of variables, the descriptive statistics is shown only for the variables that had the highest correlation with the output and were therefore used for clustering.

Due to a large number of variables, a reduction procedure was conducted in the pre-modelling phase by using correlation coefficients for continuous variables and the chi-square test for categorical input variables. The variable reduction extracted 81 continuous and 18 categorical variables (99 in total) out of 141 initial variables. The input space of 99 input variables was used in ANN modelling.

Table 2 Variables used for clustering procedure and their descriptive statistics

Variable code	Variable description	Correlation with the output variable	Descriptive statistics
V84	H1TRND (H1 max. allowed coefficient of transmission heat loss per surface)	0.3941	Min.: 0.0, Mean: 0.5595, Max.: 1.6500
V86	annual thermal energy needed for heat	0.9425	Min.: 0, Mean: 41784, Max.: 3945170
V93	object construction surface d1	0.5057	Min.: 0, Mean: 110.5, Max.:18850.3
Output variable	relative value of yearly energy needed for heat (QHNDREL) in %	-	Min.: 0, Mean: 24.33, Max.: 955.51

Source: Authors.

Data pre-processing and clustering

In the pre-processing stage, it was evident that the data contained missing values and outliers. Due to existence of different methods for replacing missing values and removing outliers in the literature, it was necessary to select the most appropriate ones that will enable the minimal information loss in the clustering procedure. Previously to optimal partitioning, the data were normalized. The following steps of the data pre-processing conducted in conjunction with the clustering procedure were suggested and performed using Mathematica software:

(1) Data normalization
(2) Replacement of missing data using the Least Squares (LS) distance-like function and l_1-metric function
(3) Elimination of outliers using DBSCAN algorithm
(4) Clustering procedure using the combination of Incremental and k-means algorithm with application of the Davies-Bouldin and Calinski-Harabasz index to optimize the number of clusters.

In step 2, an algorithm proposed by Scitovski et al., (2018) detect missing values of features and put the arithmetic mean of known values in their places. By using minimal distance principle with the Least Squares (LS) distance-like function $d_{LS}(a,b) = \|a-b\|_2^2$, and the l_1-metric function $d_1(a,b) = \|a-b\|_1$ it is shown that this is the best possibility. Following that suggestion, the dataset of public buildings was reconstructed to be used for clustering.

In order to eliminate outliers in step 3, the DBSCAN algorithm was applied according to Viswanath and Babu (2009) using Mathematica software tool. First, for each data point a radius of the circle with *MinPts* elements of data set is defined. After that, the generated set of radii is grouped into two clusters: a cluster with relatively small and a cluster with relatively large radii. The smallest interval containing 95% elements of the cluster with relatively small radii was determined, and the elements outside of that interval were identified as outliers and removed from the dataset. Figure 1 shows the normalized data after removing outliers. The optimal partition of the dataset with the most appropriate number of clusters was conducted in step 4 by the combination of the Incremental Algorithm (Bagirov et al., 2011, Scitovski, Scitovski, 2013)) and the classical k-means algorithm (see e.g. Kogan, 2007).

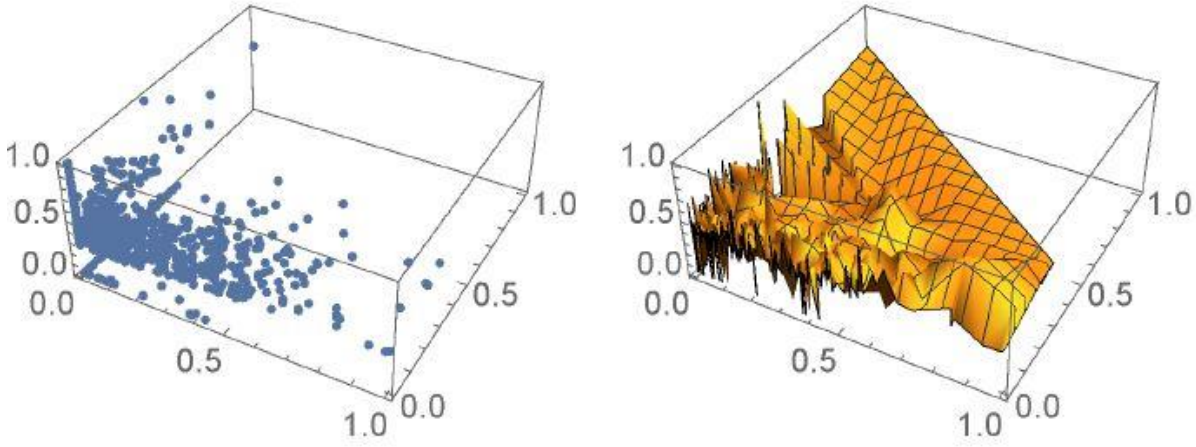

Figure 1 Normalized data after removing outliers
Source: Authors calculations in Mathematica software.

The most appropriate number of clusters is determined based on the Davies-Bouldin index and the Calinski-Harabasz index (Scitovski, Scitovski, 2013). The clusters were further used in ANN modelling.

Neural network methodology

As one of the machine learning methods, ANN has shown its success in solving prediction, classification and association problems (Prieto et al., 2016, Zekić-Sušac et al., 2016). The most common type of ANN is the multilayer perceptron (MLP) suggested by Werbos in 1974 and improved by Rumelhart et al. (Masters, 1995). An MLP network used in this paper had a three-layer structure, with an input layer, a hidden layer, and an output layer. The computation in the hidden layer includes a summation function, which sums weighted inputs from the input layer units, and an activation function, which computes the output of the hidden layer by using a linear or a nonlinear function. The computation can be summed into:

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

where y_c is the computed output, x_i are the elements of the input vector, w_i are the elements of the weight vector (values of the weights are initially randomly determined from the interval $[-1,1]$ and later adjusted by the error term) and n is the number of units in the layer. In this paper, the nonlinear logistic and tangent hyperbolic activation functions have been used. In the output layer of a neural network, the local error ε is calculated as $\varepsilon = y_c - y_t$, where y_t is the actual (target) output. In the next iteration another input vector is loaded into the input layer, the weight values are adjusted according to a learning rule (Masters, 1995), and the above process of computing the output is repeated. This learning process is conducted in a cross-validation procedure, where the neural network was trained on the training sample for a k number of iterations, than tested on the test sample, and the process is repeated until the error on the test sample stops decreasing. The maximum number of iterations in the experiments was set to 100000. The number of hidden units varied from 1 to 15, the learning rate was set to 0.01. The four ANN models were created: (a) model with all buildings, (b) model with buildings in cluster 1, (c) model with buildings in cluster 2, and (d) model with buildings in cluster 3. The

accuracy of the models is evaluated by using SMAPE objective functions according to (Tofallis, 2015):

$$SMAPE = 100\frac{1}{n}\sum_{i=1}^{n}\frac{|y_t - y_c|}{|y_t| + |y_c|} \quad (2)$$

The ANN models' calculations and graphs were generated by using R software tool, and the results of ANN models before and after clustering were compared by the statistical t-test of equality of means for independent samples, also conducted in R software tool.

Results

In the clustering procedure, Davies-Bouldin index and the Calinski-Harabasz index have extracted 3 clusters as the most acceptable option of partitioning. The number of elements in each cluster, the standard deviation of the clusters, which shows the intensity of dispersion, as well as the centres of clusters are given in Table 3. The number of elements in each cluster equals the number of cases in the ANN samples (Table 1).

Table 3 Properties of optimal partition clusters

Cluster	Number of elements	Cluster center c_j^*
Cluster 1 – center π_1^*	322	c_1^* = { 54570,13.0891, 1.22424 }
Cluster 2 - center π_2^*	190	c_2^*= { 210433, 21.2811, 1.20905 }
Cluster 3 - center π_3^*	109	c_3^* = { 434672, 34.3106, 1.18908 }

Source: Authors' calculations in Mathematica software.

It can be seen in Table 3 that the first cluster is the largest one with 322 elements. The suggested 3 clusters of buildings are graphically presented in Figure 2. The points of each cluster are shown in different colour.

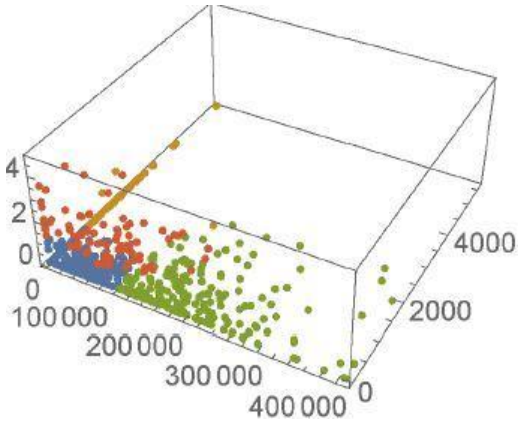

Figure 2 Clusters of public buildings (cluster 1- blue dots, cluster 2 – red dots, cluster 3 – green dots), x axis = *annual thermal energy needed for heat,* y axis= *H1TRND,* z axis = *object construction surface d1*

Source: Authors' calculations in Mathematica software.

It can be seen (Figure 2) that the centre of cluster 1 has smaller values of the first two attributes (*annual thermal energy needed for heat and H1TRND*, and the largest value of the third attribute (*object construction surface d1*). The centre of cluster 3 has the highest values of the first two attributes but the smallest value of the third attribute.

The results of the ANN prediction models obtained on the test samples before and after clustering are presented in Table 4. Since the logistic function has outperformed the tangent hyperbolic function in all four models, only the results obtained by that function are shown. The most accurate prediction model is obtained for all available buildings (SMAPE error of 35.48%).

Table 4 Results of ANN models before and after clustering

Model	ANN structure	Activation function	SMAPE (%)
a – All available buildings	MLP 180-3-1	Logistic	35.48
b – Buildings in cluster 1	MLP 180-2-3-1	Logistic	35.87
c – Buildings in cluster 2	MLP 165-2-1	Logistic	36.59
d – Buildings in cluster 3	MLP 180-2-1	Logistic	37.26

Source: Authors' calculations in R software.

Table 4 shows the optimal structure of each of the three ANN models obtained in a cross-validation procedure, activation functions tested, and the accuracy of each model expressed as the symmetric mean square error (SMAPE). Due to the fact that there were 81 continuous and 18 nominal (categorical) variables in the input space, and that the each category of a nominal variable requires an input unit, there were 180 input units in all ANN models. The output layer consisted of one unit (continuous output variable), while the optimal number of hidden units varied. In the most accurate ANN Model a, the optimal number of hidden units was 3, in Model c and Model d it was 2. The only model where the two hidden layers were required was Model b (for buildings in cluster 2), which consisted of 2 units in the first hidden layer and 3 units in the second hidden layer. The ANN model obtained on cluster 1 has produced the SMAPE of 35.87%, while the SMAPEs of cluster 2 (36.59%) and cluster 3 (37.26%) were a little higher. Therefore, the accuracy of the ANN models was not improved after clustering. In order to test our hypothesis, the t-tests of equalities of variances and means (two-way) for independent samples were performed. The results of the t-tests are shown in Table 5.

Table 5 Results of the one-way test of difference in model accuracy

Comparison	Test of equalities of variances	Test of equalities of means
Model a vs. Model *b*	F = 0.89826, p = 0.5321	p = 0.9138
Model a vs. Model *c*	F = 0.79049, p = 0.3402	p = 0.8413
Model a vs. Model *d*	F = 3.0199, p = 5.239e-06	p = 0.5585

Source: Authors' calculations in R software.

The results of t-tests have shown that the variances of errors between Model *a* and Models *b* and *c* respectively are equal, while the variances of errors between Model *a* and Model *d* are not. The test of equalities of means shows that there is no significant difference between SMAPEs of Model a (with all available buildings) and any of the models obtained on separate clusters (Model b, c, and d).

Thus, it can be concluded that the clustering procedure did not improve the accuracy of energy efficiency prediction for buildings after clustering. An insight into the characteristics of buildings in each of the clusters has shown that cluster 1 contains buildings with a high dispersion of energy consumption (mean *QHNDREL* of 112.3365 and standard deviation of 106.51621), that buildings in cluster 2 have a higher mean value of energy consumption with smaller deviation (mean *QHNDREL* of 128.0911and standard deviation of 67.6094), while buildings in cluster 3 have the highest mean QHNDREL of 168.0902 and standard deviation of 75.3624. The results indicate that a lower accuracy of prediction in cluster 3 could be influenced by

higher values of output in that cluster comparing to other two clusters. It implies that some other attributes should be included in the clustering procedure to obtain partition that will enable more accurate later prediction.

Model implementation as a cost-saving approach

The implementation of the created models in public sector energy management can have significant economic effects in saving costs in energy consumption. Sajter (2017) suggests a framework for evaluating long-term financial effects of energy efficiency projects. One of the possibilities is using the ANN models based on clustered data in such framework in the process of planning investments in reconstruction measures. Since public buildings are owned or rented by the state, the state bodies could use models to simulate the efficiency of reconstruction process by the following steps:

(1) changing input variables by entering new measures that are planned for reconstruction of buildings,
(2) running ANN models to observe the change of the values in the output variable (energy efficiency) as the effect of reconstruction measures,
(3) identify buildings, for which the change of the output variable is the most significant, and
(4) allocate their resources in the selected buildings to optimize the efficiency of reconstruction measures,
(5) use ANN models to predict the future energy consumption with the allocated resources in reconstruction and estimate the total savings by taking prices into account.

An advantage of the Croatian public sector is its information system for energy management (ISGE) with a centralized database, which is an appropriate platform for integrating ANN models. Therefore, the created machine learning models based on clustering and ANN could improve the efficiency of decision making process during allocating resources in reconstruction measures in public sector and increase the total savings.

Conclusion

The aim of the paper was to investigate if a clustering procedure would improve the accuracy of a neural network prediction in modelling energy efficiency of public buildings or not. A real dataset of Croatian public buildings was used with a large number of building attributes describing geospatial, construction, heating, cooling, occupation, and other characteristics. The total dataset is preprocessed using an algorithm, which replaces missing data and removes outliers in conjuction with clustering procedure based on Davies-Bouldin index and the Calinski-Harabasz index. Three important clusters of public buildings were identified using the combination of the Incremental algorithm and the k-means algorithm. The ANN methodology is then used to create prediction models of energy efficiency for all available buildings and separately for each cluster of buildings. The results show that the clustering procedure has not increased the prediction accuracy of ANN models in any of the three building clusters, showing that a neural network is a robust method that benefits more from the size of the total dataset than from its partition. The limitations of this research are in a small number of attributes used to generate clusters as well as in using only three and four-layered ANN architectures, which is planned to be improved in future research by using more features for clustering and test deep learning ANN.

The main contribution of the paper is in a special approach to integrate data pre-processing, clustering and ANN in modelling energy efficiency, which was not suggested before. Those preliminary findings can be used to create a methodological framework for obtaining more accurate models with an ability of saving costs and planning investments in energy management in public sector.

References

1. Bagirov, A. M., Ugon, J., Webb, D. (2011). An efficient algorithm for the incremental construction of a piecewise linear classifier. *Information Systems*, Vol. 36, pp. 782-790.
2. Hsu, D. (2015). Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. *Applied Energy*, Vol. 160, pp. 153-163.
3. Kalogirou, S. A. (2006). Artificial neural networks in energy applications in buildings. *International Journal of Low-Carbon Technologies*, Vol. 1, No. 3, pp. 201-216.
4. Kogan, J. (2007). *Introduction to Clustering Large and High-dimensional Data*. Cambridge University Press, New York.
5. Mangold, M., Osterbring, M., Wallbaum, H. (2015). Handling data uncertainties when using Swedish energy performance certificate data to describe energy usage in the building stock. *Energy and Buildings*, Vol. 102, pp. 328-336.
6. Masters, T. (1995). *Advanced Algorithms for Neural Networks*, A C++ Sourcebook. John Wiley & Sons, New York.
7. Naji, S., Shamshirband, S., Basser, H., Alengaram, U. J., Jumaat, M. Z., Amirmojahedi, M. (2016). Soft computing methodologies for estimation of energy consumption in buildings with different envelope parameters. *Energy Efficiency*, Vol. 9, No. 2, pp. 435-453.
8. Patterson, M. G. (1996). What is energy efficiency?: Concepts, indicators and methodological issues. *Energy Policy*, Vol. 24, No. 5, pp. 377-390.
9. Prieto, A., Prieto, B., Martinez Ortigosa, E., Ros, E., Pelayo, F., Ortega, J., Rojas, I. (2016). Neural networks: An overview of early research, current frameworks and new challenges. *Neurocomputing*, Vol. 204, pp. 242-268.
10. Sabo, K., Scitovski, R., Vazler, I., Zekić-Sušac, M. (2011). Mathematical models of natural gas consumption. *Energy Conversion and Management*, Vol. 52, pp. 1721-1727.
11. Sajter, D. (2017). Methods of evaluating long-term financial effects of energy efficiency projects. Business and Economic Horizons, Vol. 13, No. 3, pp. 295-311.
12. Scitovski, R., Scitovski, S. (2013). A fast partitioning algorithm and its application to 10 earthquake investigation. *Computers & Geosciences*, Vol. 59, pp. 124-131.
13. Scitovski, R., Zekić-Sušac M., Has A. (2018). Searching for an optimal partition of incomplete data with application in modeling energy efficiency of public buildings, *Croatian Operational Research Review*, Vol. 9, No. 2, in press.
14. Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, Vol. 66, No. 8, pp. 1352-1362.
15. Tommerup, H., Rose, J., Svendsen, S. (2007). Energy-efficient houses built according to the energy performance requirements introduced in Denmark in 2006. *Energy and Buildings*, Vol. 39, No. 10, pp. 1123-1130.
16. Viswanath, P., Babu, V. S. (2009). Rough-DBSCAN: a fast hybrid density based clustering method for large data sets. *Pattern Recognition Letters.* Vol. 30, pp. 1477-1488.
17. Wang, Z. X., Ding, Y. (2015). An occupant-based energy consumption prediction model for office equipment. *Energy and Buildings*, Vol. 109, pp. 12-22.
18. Zekić-Sušac, M. (2017). Overview of prediction models for buildings energy efficiency. *Proceedings of the 6th International Scientific Symposium Economy Of Eastern Croatia – Vision and Growth*, Mašek Tonković A. (Ed.), Faculty of Economics in Osijek, Osijek, May 25-27, 2017, pp. 697-706.
19. Zekić-Sušac, M., Šarlija, A., Has, A., Bilandžić, A. (2016). Predicting company growth using logistic regression and neural networks. *Croatian Operational Research Review*, Vol. 7, No. 2, pp. 229-248.

About the authors

Marijana Zekić-Sušac is a Full Professor at the Faculty of Economics of the University of J. J. Strossmayer in Osijek, Croatia. She has earned her doctoral degree at the Faculty of Organization and Informatics of the University of Zagreb located in Varaždin, Croatia. Her research interests include artificial intelligence, machine learning and data mining in business, education and medicine. She teaches several ICT courses at the undergraduate, the graduate and the doctoral level. She is a member of the International Neural Network Society, the Croatian Operational Research Society, and the Croatian Statistical Association. Author can be contacted at: *marijana@efos.hr.*

Rudolf Scitovski received his PhD in Applied Mathematics from the University of Zagreb in 1984. He works as a Full Professor at the Department of Mathematics of the University of Osijek. He was the Head of the Department of Mathematics for a long period of time. Before that, he had been employed at the Faculty of Electrical Engineering and the Faculty of Economics of the University of Osijek. His research interests include least square and least absolute deviations problems, clustering and global optimization. Author can be contacted at: *scitowsk@mathos.hr.*

Adela Has graduated from the Faculty of Economics of the University of J. J. Strossmayer in Osijek in 2010. Since 2014 she has been a doctoral student at the international inter-university postgraduate interdisciplinary doctoral program Entrepreneurship and Innovativeness. She is employed as an assistant at the Faculty of Economics in Osijek in the scientific field of economics, business information branch. She is a member of the Croatian Operational Research Society. Author can be contacted at: *adela.has@efos.hr.*