PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCLAW UNIVERSITY OF ECONOMICS AND BUSINESS

2020, vol. 64, nr 9

ISSN 1899-3192
e-ISSN 2392-0041

Joanna Perzyńska
West Pomeranian University of Technology in Szczecin
Faculty of Economics
e-mail: joanna.perzynska@zut.edu.pl
ORCID: 0000-0002-7182-2381

APPLICATION OF KOHONEN NETWORKS FOR CLUSTERING OF THE ZACHODNIOPOMORSKIE VOIVODESHIP DISTRICTS IN TERMS OF THE LEVEL OF SOCIO-ECONOMIC DEVELOPMENT

DOI: 10.15611/pn.2020.9.08
JEL Classification: C38, C45, R11, R15.

Quote as: Perzyńska, J. (2020). Application of Kohonen networks for clustering of the Zachodniopomorskie Voivodeship districts in terms of the level of socio-economic development. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, *64*(9).

Abstract: The author presents the possibilities of using artificial neural networks in a multidimensional analysis – cluster analysis. The empirical example using districts of the Zachodniopomorskie (West Pomeranian) Voivodeship is the illustration of theoretical considerations. The study used statistical data from many areas related to socio-economic development: demography, labour market, natural environment, recreation, culture, social and technical infrastructure, and the economy. The aim of the study was to divide the voivodeship into disjointed typological groups of districts using Kohonen networks (Self-Organizing Maps). Several networks differing in structure of the output layer were constructed and trained. Selected diagnostic features of socio-economic development of districts were their input values. Using verified Kohonen networks, various sets of groups of the researched objects were created, and confirmed them are a useful tool for identifying clusters of districts similar to each other in terms of the level of socio-economic development.

Keywords: cluster, district, Kohonen network, socio-economic development, West Pomeranian Voivodeship.

1. Introduction

Artificial neural networks are used in many different areas where problems arise related to data processing and analysis. Neural models allow for the mapping of very complex functions when there are no grounds for the linear approximation of phenomena. They enable solving problems related to the modelling of multidimensional vector functions, i.e. nonlinear functions with a large number of independent variables. Neural networks give the opportunity to search for models mapping the complex relationships between data for phenomena whose structure, operating principles or causal relationships have not been sufficiently known to build effective mathematical models (Lula and Tadeusiewicz, 2001). They were able to discover in the data set undetectable compounds even by applying traditional statistical methods (Masters, 1996). Artificial neural networks do not require programming – they acquire the necessary knowledge in the learning process and are able to generalise it later.

Kohonen networks, also known as Self-Organizing Maps (SOM), are used for exploratory analysis of data. They are used to classify data regardless of similarity criteria, and can detect new products (different classes, not similar to learned). The essence of their operation is analogous to the methods of cluster analysis derived from the patternless classification– unsupervised learning. The goal of cluster analysis is to partition a set of objects into distinct relatively homogeneous groups in such a manner that that all the objects within a group are similar in terms of variable values, while observations in different groups are not similar. Cluster analysis is a tool for classifying objects into groups in a multidimensional space (Timm, 2002).

Kohonen networks cope well with the multidimensionality of the analysed phenomena, and socio-economic development is just such a phenomenon. The level of socio-economic development is affected by many different factors (in addition to obvious social and economic changes, e.g. political and cultural conditions). Socio--economic development is a process concerning important changes taking place in the territorial system. It is a complex phenomenon derived from activities affecting its shape and the strategic goals it serves (Brol, 1998; Kogut-Jaworska, Szewczuk, and Zioło, 2011; Gorynia and Łaźniewska, 2012). Considering the socio-economic processes in their regional distribution allows to identify the factors that shape them, resulting from the regional differentiation of space. The regional distribution of global processes may indicate areas of low efficiency, low level of meeting needs and poor social activity, which is a condition for formulating realistic improvement programs (Gorzelak and Smętkowski, 2019).

The aim of the study was to divide the Zachodniopomorskie Voivodeship into disjointed typological groups of districts using Kohonen neural networks. A research hypothesis that the separated clusters include districts with a similar level of socio-economic development was put forward.

2. Methods

Artificial neural networks constructed as computer programs reflect only the basic processes occurring in the human brain. They are a system of interconnected artificial neurons that are the smallest elements of the data processing network. Each artificial neuron receives signals called input values – they can be primary values (data fed into the network from the outside) or intermediate values (from the outputs of other neurons in the network). Each intermediate input value is introduced into an artificial neuron through a connection of a certain weight. The neuron processes the input signals, and this takes place in two stages: aggregation of input values (processing them into a single value called total neuron stimulation) and determination of the output value of the neuron (after transformation of the value of total neuron stimulation by activation function). The neurons in the network are arranged in layers, thus depending on the location, the output values of the neurons of a given layer are sent to the neurons of the next layer or are the output values of the network (Bishop, 1995; Gately 1999; Lula and Tadeusiewicz, 2001).

The artificial neural network has an input layer (its neurons are used to input data from outside and pre-processing) and an output layer (its neurons determine the final result of the calculations). There may also be hidden layers in the network, its neurons mediate information processing within the network and an external observer has no direct access to them. The outputs of one layer are connected to the inputs of subsequent layers, the signals can be sent only in one direction or they can circulate the entire network in a cyclical way. Connection weights are modified in the process of learning the network. Depending on the way neurons are connected there can be distinguished a feedforward, radial basis function, recurrent and cellular networks (Lula, Tadeusiewicz, and Wójtowicz, 1999).

Cellular networks are used for problems related to the patternless data classification. They have a regular structure in which interconnections occur between neurons in the immediate neighbourhood. The most commonly used cellular network is the self-organizing Kohonen network (Self-Organizing Map). The architecture of the network is simple, it consists of only two layers: input and output, without hidden layers. The input layer has as many neurons as the number of features, while the size of the output layer is chosen arbitrarily and has as many neurons as the number of classes. The input values introduced to the first layer of the network are subject to pre-processing consisting of normalisation. Normalised values through direct connections reach all neurons of the last layer in which their basic processing takes place. Output neurons are radial neurons, defined by the coordinates of their centre (weight vector) and radius. The total stimulation of each radial neuron is determined as the value of the selected measure of the distance of its weights vector from the input value vector. The smaller this distance is, the bigger the neuron's output value determined using the activation function – the neighbourhood function. The neuron whose output is the largest becomes the winning neuron (Lula and Tadeusiewicz, 2001; Witkowska, 2002).

The Kohonen network is trained without a teacher. It is a form of unsupervised learning – competitive learning. Learning data are the known values of input variables, the network itself learns to recognize their common characteristics and groups them into appropriate classes represented by the network outputs. When learning is based on the iterative Kohonen algorithm, the principles *the winner takes all* or *the winner takes the most* are used. According to the first rule in a given iteration, only the winning neuron is trained and the values of its weights are modified in such a way as to bring them closer to the input values of the network. Using *the winner takes the most* principle, the weights of neurons adjacent to the winning neuron are also modified – their change is in the same direction, however it is smaller than for the winner and is inversely proportional to the distance of individual neurons from the winner. The new values of neuron weights in the neighbourhood of the winning neuron are based on the formula:

$$w_{lj}^{(r+1)} = w_{lj}^{(r)} + \eta s(l,k)\left(z_j^{(r)} - w_{lj}^{(r)}\right), \tag{1}$$

where: j – input neuron number, l – output neuron number, k – winning output neuron number, r – learning cycle iteration number, $w_{lj}^{(r)}$ – the weight of the connection of the j-th input neuron with the l-th output neuron in r-th iteration, $w_{lj}^{(r+1)}$ – the weight of the connection of the j-th input neuron with the l-th output neuron in $r+1$-th iteration, $z_j^{(r)}$ – normalised j-th input value in the r-th iteration, $s(l, k)$ – value of the neighbourhood function of the l-th output neuron relative to the winning neuron, η – monotonically decreasing learning coefficient.

During learning the neighbourhood radius decreases, which means that less and less neurons belong to it, and finally only contains the winning neuron. As a result of the learning process, a topological map is created, which is usually a two-dimensional grid with neurons of the output layer arranged regularly in its nodes. They only recognize a class of values similar to those that previously made them a winner (Kohonen, 1982; Witkowska, 2002).

The verification of the constructed neural model is carried out by determining the values of universal quality measures and their interpretation. For Kohonen networks these measures are determined as network errors (mean square errors or mean absolute errors) based on deviations of the input values from the values of the best patterns, i.e. the weights of individual output neurons. These errors are determined separately for the training, validation and testing set, which are subsets of the set of input values. A learning error allows for the evaluation of the network's ability to approximate. An increase in the approximation capacity may be accompanied by a decrease in the ability to generalise, i.e. the ability to properly operate the network for data outside the training set. The identification of this phenomenon called network overfitting, is made possible by a validation error, its value may be slightly higher than the value of the learning error. After completing the learning process, as

a result of which the learning and validation errors have reached an acceptable level, the final verification of the model is carried out on the test set. The testing error allows the assessment of the network's ability to generalise, its value may exceed the value of the learning error, however if the difference is significant, then it means that there was an excessive overfitting to the learning data (Lula and Tadeusiewicz, 2001; Masters, 1996).

In the empirical research, Kohonen networks were designated in the STATISTICA program. The research procedure comprised the five following stages:

Stage 1. Collecting of potential diagnostic features.

Stage 2. Selecting of diagnostic features (statistical verification and reduction of the initial set of features).

Stage 3. Normalising diagnostic features.

Stage 4. Constructing, learning and verifying Kohonen networks differing in the architecture (with input values being normalised diagnostic characteristics).

Stage 5. Partitioning a set of districts into clusters based on the output values of Kohonen networks.

Cartograms showing the spatial distribution of districts in clusters were made in the STATISTICA MAPS program.

3. Materials

In the empirical research, 21 districts of the voivodeship (including three cities with district status: Koszalin, Szczecin, Świnoujście) were the objects of study. The 2018 data from the Local Data Bank of the Central Statistical Office (Statistical Office in Szczecin, 2019) were used. The initial set included several dozen variables from various areas related to socio-economic development, their list was established following substantive and formal criteria, and based on the work of Gibas and Heffner (2007) and Oesterreich, Perzyńska, Barej-Kaczmarek (2019).

The initial set of collected variables (potential diagnostic features) was verified using statistical procedures. The verification procedures consisted in assessing the degree of fulfilment by the variables of two main criteria: information capacity and the ability to discriminate units of analysis (Nowak, 1990; Panek and Zwierzchowski, 2013). This evaluation was based on coefficients of variation and linear correlation, the critical values of which were arbitrarily assumed as 10% and 0.7, respectively. From the initial set of variables the study eliminated those for which the value of variation coefficient was less than 10%, and the absolute value of the Pearson's linear correlation coefficient was greater than 0.7. As a consequence, the set of initial features was reduced in such a way that selected diagnostic features were characterised by their high variability in relation to the examined objects and low correlation with other characteristics.

Finally, four variables from each of five areas (demography and labour market (DLM), natural environment (NE), recreation and culture (RC), social and technical

infrastructure (STI), and the economy (E)) were selected as diagnostic features. The set of potential diagnostic features is listed below (the selected features are marked with symbols denoting the area and the variable number):

- infant deaths per 1000 live births (DLM_X1),
- natural increase per 1000 population (DLM_X2),
- non-working population per 100 working age people,
- number of people per 1 km^2,
- post-working population per 100 pre-working age people,
- post-working population per 100 working age population (DLM_X3)
- registered unemployment rate (in %) (DLM_X4),
- share of total employed in the population (in %),
- gaseous emissions per 100 km^2 (in t) (NE_X5),
- forest area per 1 km^2 (in ha) (NE_X6),
- percentage of the population connected to wastewater treatment plants (NE_X7),
- share of protected areas in the total area (in %) (NE_X8),
- wastewater treated per 100 km^2 (in dm^3),
- water consumption for the needs of the national economy and population per 1 km^2 (in dam^3),
- accommodation density index (tourist accommodation establishments per 1 km^2),
- average length of stay,
- Charvat's indicator (overnight stays in tourist establishments per 100 residents),
- Defert's indicator (number of beds per 100 residents) (RC_X9),
- Schneider's indicator (tourists accommodated in tourist accommodation establishments per 100 residents),
- tourist density indicator (tourists accommodated in tourist accommodation establishments per 1km^2) (RC_X10),
- book collection per 1000 population (in volumes) (RC_X11),
- population per 1 library (RC_X12),
- beds in hospitals per 10000 population (STI_X13),
- population per 1 healthcare entity (STI_X14),
- population per 1 pharmacy (STI_X15),
- dwellings per 1000 population,
- water supply network per 100 km^2 (in km),
- wastewater network per 100 km^2 (in km),
- districts roads with a hard surface per 100 km^2 (in km),
- number of boiler rooms per 100 km^2,
- length of transmission and distribution heating network in km per 100 km^2,
- length of connections of the transmission and distribution heating network to buildings per 100 km^2 (in km) (STI_X16),
- housing benefits per capita (in PLN) (E_X17),
- investment in enterprises per capita (in PLN),
- gross value of fixed assets per capita (in PLN),

- entities of the national economy per 10000 population (E_X18),
- average monthly gross salary (in PLN) (E_X19),
- revenue of districts budgets per capita (in PLN) (E_X20),
- expenditure of districts budgets per capita (in PLN).

Table1 presents the basic descriptive statistics of the selected diagnostic features, while Table 2 shows the correlation matrix for them.

Table 1. Basic descriptive statistics of selected diagnostic features

Diagnostic feature	Descriptive statistics			Diagnostic feature	Descriptive statistics		
	mean	median	variation coefficient		mean	median	variation coefficient
DLM_X1	4.2	4.2	37.2	RC_X11	3711.4	3533.0	24.1
DLM_X2	–0.9	–1.3	149.5	RC_X12	4100.2	3190.0	64.4
DLM_X3	32.9	32.0	12.5	STI_X13	35.5	28.6	65.2
DLM_X4	11.7	11.7	45.7	STI_X14	1438.5	1443.0	17.5
NE_X5	79.0	79.2	13.0	STI_X15	3228.7	3160.0	23.4
NE_X6	486.2	27.1	190.3	STI_X16	6.0	3.6	150.2
NE_X7	33.7	35.3	34.3	E_X17	23.8	23.6	42.9
NE_X8	19.1	15.0	91.9	E_X18	1180.6	1041.0	23.0
RC_X9	10.1	1.2	140.6	E_X19	3849.5	3761.6	10.6
RC_X10	279.1	22.3	184.4	E_X20	1868.1	1185.4	98.5

Source: own study.

In order to bring the selected variables to comparability and unification of the character, they were subjected to normalising transformation in the process of zero unitarisation according to the formulas:

$$z_{ij} = \frac{x_{ij} - \min\limits_{i} x_{ij}}{\max\limits_{i} x_{ij} - \min\limits_{i} x_{ij}}, \tag{2}$$

$$z_{ij} = \frac{\max\limits_{i} x_{ij} - x_{ij}}{\max\limits_{i} x_{ij} - \min\limits_{i} x_{ij}}, \tag{3}$$

where: i – district number (i = 1, 2, …, 21), j – feature number (j = 1, 2, …, 20), x_{ij} – value of j-th feature (X_j) in i-th district, z_{ij} – normalised value of x_{ij}, $\max\limits_{i} x_{ij}$ – highest value of j-th feature among districts, $\min\limits_{i} x_{ij}$ – lowest value of j-th feature among districts $(\max\limits_{i} x_{ij} \neq \min\limits_{i} x_{ij})$.

Formulas (2) and (3) were used for stimulants (variables positively affect the analysed phenomenon) and destimulants (variables negatively affect the analysed

Table 2. Correlation matrix for selected diagnostic features

	DLM_ X1	DLM_ X2	DLM_ X3	DLM_ X4	NE_X5	NE_X6	NE_X7	NE_X8	RC_X9	RC_X10	RC_X11	RC_X12	STI_ X13	STI_ X14	STI_X15	STI_X16	E_X17	E_X18	E_X19	E_X20
DLM_X1	1	–0.01	0.04	–0.08	–0.22	0.01	–0.35	–0.23	–0.15	–0.09	0.18	0.09	0.14	–0.21	–0.10	0.19	–0.20	–0.14	–0.03	–0.01
DLM_X2	–0.01	1	–0.56	–0.21	–0.09	0.31	0.16	–0.23	–0.13	–0.40	–0.50	0.14	–0.27	0.21	0.39	–0.10	–0.36	0.08	0.23	–0.37
DLM_X3	0.04	–0.56	1	–0.34	0.55	0.21	–0.30	0.08	0.08	0.62	0.32	0.55	0.48	–0.48	–0.65	0.63	0.35	0.59	0.29	0.70
DLM_X4	–0.08	–0.21	–0.34	1	–0.49	–0.51	0.29	0.08	–0.07	–0.56	0.40	–0.60	–0.32	–0.03	0.39	–0.52	0.24	–0.70	0.61	–0.51
NE_X5	–0.22	–0.09	0.55	–0.49	1	0.32	0.01	0.17	–0.15	0.47	–0.33	0.56	0.61	–0.37	–0.50	0.47	0.42	0.54	0.47	0.52
NE_X6	0.01	0.31	0.21	–0.51	0.32	1	–0.21	–0.14	–0.28	0.32	–0.09	0.64	0.26	0.04	–0.21	0.68	–0.14	0.52	0.61	0.38
NE_X7	–0.35	0.16	–0.30	0.29	0.01	–0.21	1	0.68	–0.31	–0.45	–0.39	–0.25	–0.02	0.15	0.20	–0.27	0.30	–0.37	–0.18	–0.30
NE_X8	–0.23	–0.23	0.08	0.08	0.17	–0.14	0.68	1	–0.29	–0.19	–0.13	–0.20	0.39	0.08	–0.08	–0.01	0.41	–0.25	–0.16	–0.02
RC_X9	–0.15	–0.13	0.08	–0.07	–0.15	–0.28	–0.31	–0.29	1	0.41	0.21	–0.18	–0.12	0.02	–0.13	–0.16	–0.30	0.43	–0.23	0.10
RC_X10	–0.09	–0.40	0.62	–0.56	0.47	0.32	–0.45	–0.19	0.41	1	0.13	0.56	0.30	–0.11	–0.57	0.55	–0.03	0.69	0.58	0.69
RC_X11	0.18	–0.50	0.32	0.40	–0.33	–0.09	–0.39	–0.13	0.21	0.13	1	–0.15	–0.01	–0.27	–0.16	0.10	–0.22	–0.03	–0.24	0.07
RC_X12	0.09	0.14	0.55	–0.60	0.56	0.64	–0.25	–0.20	–0.18	0.56	–0.15	1	0.44	–0.30	–0.42	0.63	0.07	0.60	0.51	0.69
STI_X13	0.14	–0.27	0.48	–0.32	0.61	0.26	–0.02	0.39	–0.12	0.30	–0.01	0.44	1	–0.52	–0.38	0.60	0.42	0.37	0.23	0.33
STI_X14	–0.21	0.21	–0.48	–0.03	–0.37	0.04	0.15	0.08	0.02	–0.11	–0.27	–0.30	–0.52	1	0.53	–0.32	–0.33	–0.26	0.07	–0.04
STI_X15	–0.10	0.39	–0.65	0.39	–0.50	–0.21	0.20	–0.08	–0.13	–0.57	–0.16	–0.42	–0.38	0.53	1	–0.41	–0.24	–0.51	–0.26	–0.53
STI_X16	0.19	–0.10	0.63	–0.52	0.47	0.68	–0.27	–0.01	–0.16	0.55	0.10	0.63	0.60	–0.32	–0.41	1	0.19	0.67	0.57	0.70
E_X17	–0.20	–0.36	0.35	0.24	0.42	–0.14	0.30	0.41	–0.30	–0.03	–0.22	0.07	0.42	–0.33	–0.24	0.19	1	–0.09	–0.21	0.16
E_X18	–0.14	0.08	0.59	–0.70	0.54	0.52	–0.37	–0.25	0.43	0.69	–0.03	0.60	0.37	–0.26	–0.51	0.67	–0.09	1	0.55	0.64
E_X19	–0.03	0.23	0.29	0.61	0.47	0.61	–0.18	–0.16	–0.23	0.58	–0.24	0.51	0.23	0.07	–0.26	0.57	–0.21	0.55	1	0.62
E_X20	–0.01	–0.37	0.70	–0.51	0.52	0.38	–0.30	–0.02	0.10	0.69	0.07	0.69	0.33	–0.04	–0.53	0.70	0.16	0.64	0.62	1

Source: own study.

phenomenon), respectively (Kukuła, 2000). The obtained normalised diagnostic features were designated: Z1, Z2, ... Z20 (in place of names: X1, X2, ..., X20); the headings concerning the areas were left unchanged. The variables normalised in this way have the character of a stimulant, have values in the range [0,1] and retain varied variances proportional to the variance of primary features (Grabiński, 1992). In the empirical studies, the normalised variables were input values of the Kohonen network.

4. Results and discussion

In the empirical research, Kohonen networks were used to divide districts of the Zachodniopomorskie Voivodeship into clusters in terms of the level of socio--economic development. A maximum number of clusters equal to seven was assumed arbitrarily, and therefore networks containing from two to seven output neurons were constructed and trained. The built networks differed in the architecture of the output layer – neurons were regularly arranged (like matrix elements) in a grid with dimensions: 1x2 (SOM12), 1x3 (SOM13), 1x4 (SOM14), 1x5 (SOM15), 1x6 (SOM16), 1x7 (SOM17), 2x2 (SOM2x2), 2x3 (SOM23). Transposing the structure of the output layers (to dimensions: 2x1, 3x1, 4x1, 5x1, 6x1, 7x1, 3x2) only changed the numbering of their neurons, but all the network parameters did not change, hence the results of the research were presented only for the eight mentioned networks.

Features Z1-Z20 were the input values of the constructed SOM12-SOM23 networks – their values for each district were introduced into the input layer neurons as subsequent samples. The networks were trained without a teacher according to *the winner takes the most* principle. The learning process began with the random initialisation of weights, generated as small, different from each other, non-zero numbers. The Euclidean distance was used to determine the total stimulation value of neurons, while the Gaussian neighbourhood function was used as the activation function to determine their output values. Based on repeatedly presented samples (randomly divided into learning, validation and testing sets), the networks themselves modified the initial weights values, as a result of which the output neurons specialised in recognizing only values similar to those that previously made them a winner. The quality assessment of the learned networks was based on their mean absolute errors (MAE). Table 3 shows the MAE values obtained, divided into learning, validation and testing errors.

The information presented in Table 3 shows that the values of learning, validation and testing errors are in adequate relationships – learning errors are small and the remaining errors are only slightly larger. This demonstrates the ability of the constructed networks to approximate and generalise, which confirms their correct verification.

Table 3. Mean absolute errors of Kohonen networks

Set	SOM12	SOM13	SOM14	SOM15	SOM16	SOM17	SOM22	SOM23
Learning	0.1016	0.0878	0.0725	0.0621	0.0584	0.0580	0.0717	0.0537
Validation	0.1193	0.0972	0.0864	0.0734	0.0603	0.0612	0.0793	0.0630
Testing	0.1268	0.1030	0.0898	0.0805	0.0732	0.0749	0.0801	0.0705

Source: own study.

Analysing the information presented in Table 3, it can also be seen that as the number of neurons in the output layer of the network increases, its error determined on the learning, validation and testing sets decreases. For the pairs of networks with the same number of output neurons (SOM14, SOM22 and SOM16, SOM23), the error values are similar – they are slightly smaller for networks with 2x2 and 2x3 output layer dimensions. The values of network errors decreasing with the increase in the number of output neurons testify to the better distribution of districts in clusters. However, it should be noted that these changes are getting smaller, which confirms that there is no need to increase the maximum number of clusters.

Table 4. Distribution of districts in clusters depending on the used Kohonen network

District	SOM12	SOM13	SOM14	SOM15	SOM16	SOM17	SOM22	SOM23
Białogardzki	1	1	1	1	2	2	1	4
Choszczeński	1	1	1	1	1	1	1	1
Drawski	1	1	1	1	1	1	1	1
Goleniowski	1	2	2	2	3	4	3	2
Gryficki	1	2	3	3	4	5	2	5
Gryfiński	1	1	2	2	3	4	3	4
Kamieński	1	2	3	3	4	5	2	5
Kołobrzeski	2	3	3	4	5	6	2	5
Koszaliński	1	2	2	3	4	5	3	4
Łobeski	1	2	2	2	3	3	2	4
Myśliborski	1	1	1	1	1	1	1	1
Policki	1	2	3	4	5	5	3	3
Pyrzycki	1	2	2	2	3	3	2	4
Sławieński	1	2	2	3	4	4	2	4
Stargardzki	1	2	2	2	2	3	3	2
Szczecinecki	1	1	1	1	1	1	1	1
Świdwiński	1	1	2	2	2	2	1	4
Wałecki	1	1	1	1	1	1	1	1
Koszalin	2	3	4	5	6	7	4	6
Szczecin	2	3	4	5	6	7	4	6
Świnoujście	2	3	4	5	6	7	4	6

Source: own study.

With the use of trained and verified SOM12-SOM23 networks, the districts were divided into disjointed groups, thus obtaining eight different sets of clusters. Table 4 presents the cluster numbers to which each district belongs depending on the adopted partition, i.e. the Kohonen network used. The spatial distribution of districts in individual clusters is presented in Figures 1 and 2.

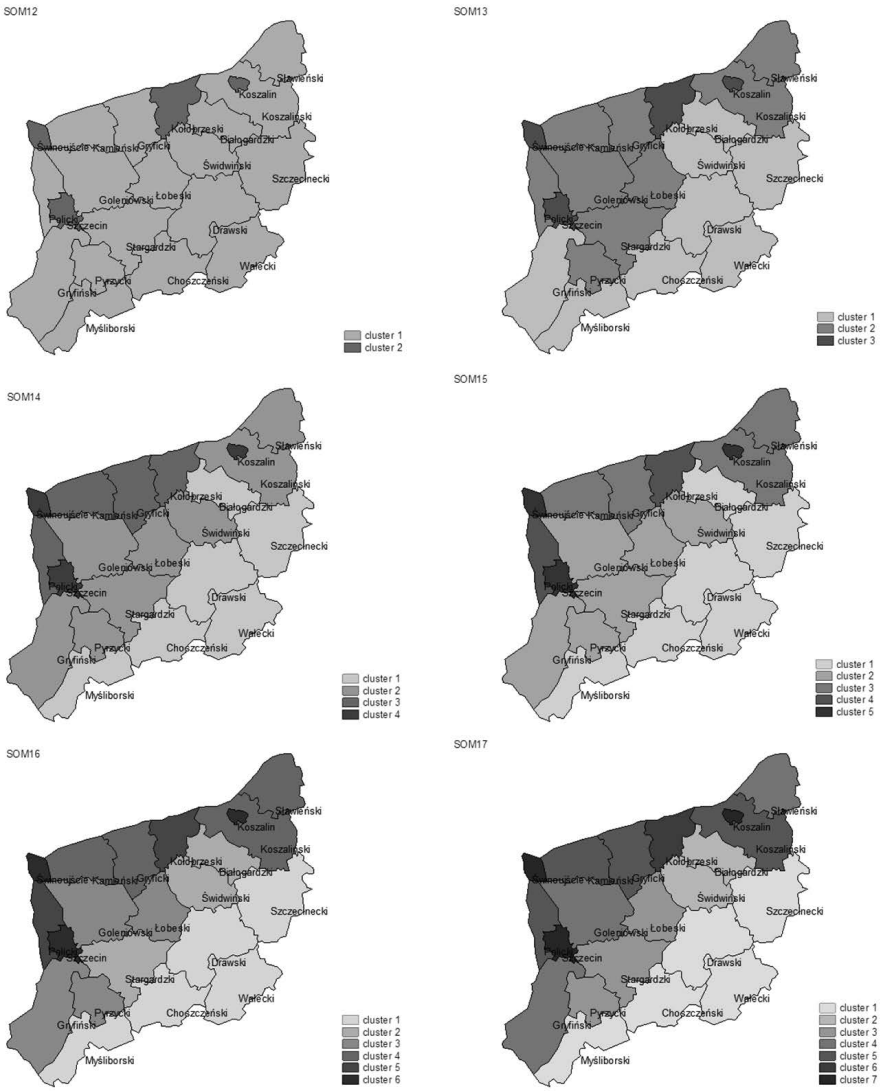

Fig. 1. Spatial distribution of clusters of the Zachodniopomorskie Voivodeship districts obtained by SOM12-SOM17

Source: own study.

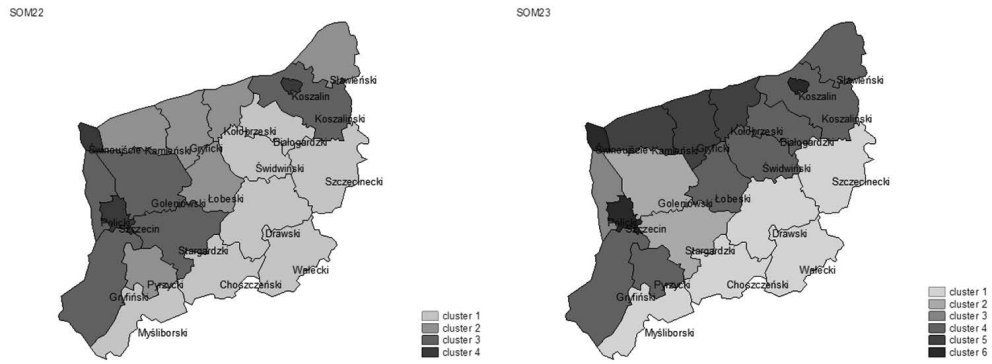

Fig. 2. Spatial distribution of clusters of the Zachodniopomorskie Voivodeship districts obtained by SOM22 and SOM23

Source: own study.

The information contained in Table 4 shows that when dividing districts into two clusters using the SOM12 network, collections with a very diverse number of districts were obtained – cluster 1 contains the vast majority of the districts (seventeen), while in cluster 2 there are only four: Koszalin, Szczecin, Świnoujście and the kołobrzeski district. The first three districts are large cities, in addition, Szczecin is the capital of the voivodeship.

It can be further seen that as the number of clusters increases to three, the most numerous cluster (cluster 1 determined using SOM12) is divided into smaller subsets, while the previous cluster 2 does not change (in the SOM13 network it is cluster 3). When the number of output neurons increases to four (SOM14 network), the kołobrzeski district is excluded from the last cluster and only three cities remain in it: Koszalin, Szczecin and Świnoujście. Further increasing the number of output neurons no longer changes the objects in the last cluster – it always contains the three listed cities.

In the case of the division of districts into four typological groups using the SOM14 network, it can be seen that on the map of the voivodeship (Figure 1) that these clusters (apart from the cluster containing the three cities) are arranged in clear diagonal stripes towards NE-SW and only the białogardzki district disturbs this order – it belongs to the cluster 1 located lower on the map, and disconnects from cluster 1 after using the SOM16 network.

When increasing the number of groups to five (SOM15) and six (SOM16), it can be seen that the kołobrzeski district, which initially belonged to a cluster of cities, now belongs to a different cluster only together with the policki district. When the number of clusters is increased to seven (SOM17), the kołobrzeski district already creates a separate one-element cluster.

The difference in the change of groups' elements is most visible when increasing the number of clusters from two to three – then nine districts are separated from cluster 1 (i.e. more than half). Further increasing the number of typological groups already has a slight impact on cluster 1 – for six groups (SOM16) the białogardzki district detaches from it and finally for seven groups (SOM17) cluster 1 does not change any more. Cluster 1 includes the following districts: choszczeński, drawski, myśliborski, szczecinecki and wałecki. They are located at the southern and south-eastern edge of the Zachodniopomorskie Voivodeship, at the border with neighbouring voivodeships. The identical composition of cluster 1 also appears for six groups determined using the SOM23 network.

The distribution of districts in clusters designated using SOM22 and SOM23 differs from that determined using SOM14 and SOM16 respectively. This is due to the structure of the output layer of these networks, i.e. the location of the winning neurons on the topological map. In cases of the SOM12-SOM17 networks, the output neurons are linearly arranged, their distances are analogous to the groups distances, hence cluster 1 is then more and more distant from the next ones. For the SOM22 and SOM23 networks, the neurons are arranged in a compact group and the distances between them are no longer so obvious in interpretation – e.g. in the SOM23 network the first output neuron is directly adjacent to the second and fourth neurons. Looking at the map of the voivodeship (Figure 2), the diagonal stripes are not so clearly visible anymore – the three largest cities form a separate cluster again, and the remaining clusters seem to be concentrated around them.

The sets of designated clusters of the voivodeship districts were subjected to additional analysis. For this purpose, for all the clusters the study determined their centers of gravity (i.e. the average values of normalised variables), as well as their weighted averages.

The weighted averages for clusters 1 and 2 determined using SOM12 equalled 0.4170 and 0.5282, respectively. After increasing the number of typological groups to four (SOM14), the weighted averages of the clusters were: 0.3904, 0.4477, 0.4296, and 0.5283, respectively. It can be seen that the separation of the kołobrzeski district from cluster 2 (determined by SOM12) resulted in a slight increase in the weighted average of the cities. In addition, cluster 1 based on SOM14 has the lowest weighted average of all four clusters. In the case of four groups determined using SOM22, their weighted averages (0.4174, 0.4145, 0.44492, 0.5283) no longer had such varied values as for SOM14. The analysis of the weighted averages also shows that their diversification decreases with the increase in the number of clusters from four to seven (from SOM14 to SOM17).

Figure 3 presents the average values of selected variables for four clusters determined using SOM14.

The selected variables DLM_Z4, NE_Z5, RC_Z10, STI_Z15, E_Z20 (normalised values of registered unemployment rate, gaseous emissions, tourist density indicator, population per one pharmacy, revenue of districts budgets per capita) are examples of

five analysed areas of socio-economic development: demography and labour market, natural environment, recreation and culture, social and technical infrastructure, and the economy. Their impact on assigning districts into clusters seems to be most visible among other variables. The variables were normalised using the zero unitarisation method, therefore while interpreting their averages it should be remembered that they are transformed into the form of stimulant.

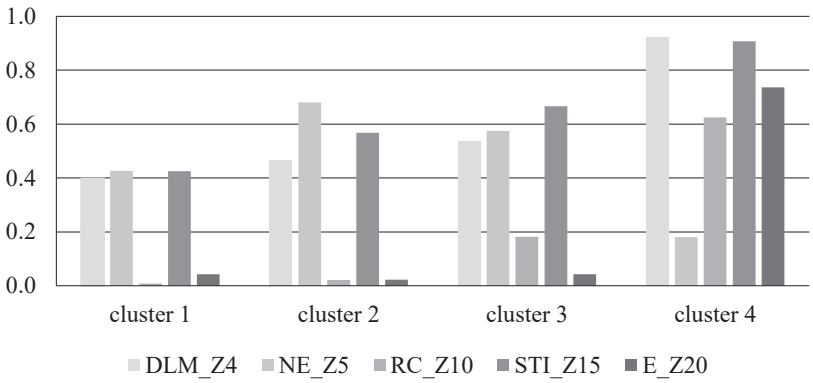

Fig. 3. Averages of selected variables for clusters determined using SOM14

Source: own study.

Figure 3 shows that the districts belonging to cluster 4 (the cities of Koszalin, Szczecin, Świnoujście) are characterised by the highest average values of variables DLM_Z4, RC_Z10, STI_Z15, E_Z20 and the lowest average of NE_Z5. The averages of DLM_Z4 and STI_Z15 greater than 0.9 mean that cities have a similar high level of development due to the values of registered unemployment rate and population per pharmacy. However, the low average of NE_Z5 means that cities are distinguished from other districts by their similar unfavourable high gaseous emissions.

Among all the clusters, the lowest average values of variables DLM_Z4, RC_Z10, STI_Z15, E_Z20 are for cluster 1 (the districts: białogardzki, choszczeński, drawski, myśliborski, szczecinecki, wałecki) which means that they are characterised by a similar low level of development due to selected features describing demography and labour market, recreation and culture, social and technical infrastructure, and the economy.

Figure 4 presents the average values of the selected variables for the seven clusters determined using SOM17.

Figure 4 shows that the cities belonging now to cluster 7, have, as before, the highest averages of DLM_Z4, RC_Z10, STI_Z15, E_Z20 and the lowest average of NE_Z5. A very similar high level of development due to the values of registered

unemployment rate, tourist density indicator, population per one pharmacy, revenue of districts budgets per capita is distinguished for cluster 6 comprising only the kołobrzeski district. When divided into two clusters (SOM12), the kołobrzeski district belonged to the same cluster as the cities which indicates that increasing the division to seven clusters is too detailed and did not improve the results in the mentioned districts.

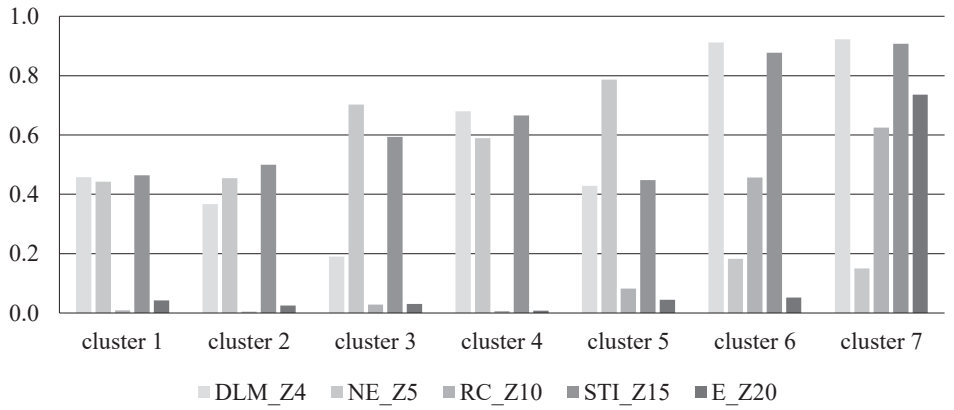

Fig. 4. Averages of selected variables for clusters determined using SOM17

Source: own study.

Cluster 1 (the districts: choszczeński, drawski, myślborski, szczecinecki, wałecki), to which the białogardzki district no longer belongs, is again characterised by the lowest level of development due to selected features describing recreation and culture, social and technical infrastructure, and the economy (RC_Z10, STI_Z15, E_Z20).

Figure 5 presents the average values of the selected variables for the four clusters determined using SOM22.

Using the SOM14 and SOM22 networks differing in the structure of the output layer, four clusters were obtained. Cluster 4 designated using SOM22 and SOM14 includes the same districts: Koszalin, Szczecin, Świnoujście. Figure 5 shows that these cities are again characterised by the highest averages of variables DLM_Z4, RC_Z10, STI_Z15, E_Z20 and the lowest average of NE_Z5.

Cluster 1 determined using SOM22 differs from the analogous cluster obtained using SOM14 in that it additionally contains the świdwiński district. This cluster has the lowest averages of the variables DLM_Z4, RC_Z10, STI_Z15 (for SOM14 this also applied to the variable E_Z20), however, it should be noted that the differences between this and other clusters are not so clear.

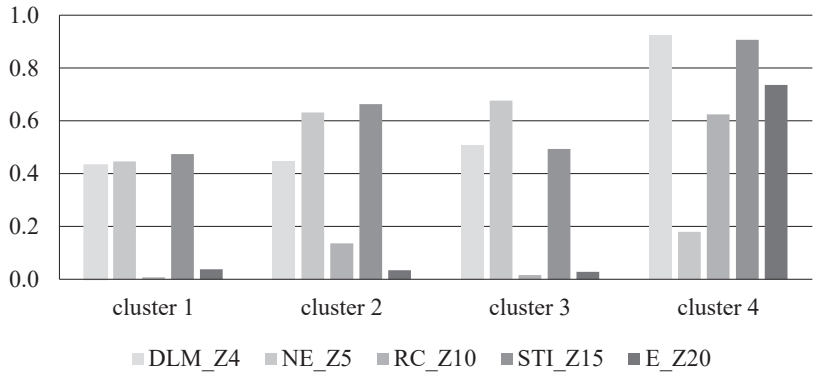

Fig. 5. Averages of selected variables for clusters determined using SOM22

Source: own study.

Oesterreich, Perzyńska and Barej-Kaczmarek (2019) used synthetic taxonomic measures to evaluate the level of socio-economic development of the Zachodniopomorskie Voivodeship districts and their grouping. They used a dynamic approach, but the main conclusion is identical – there is a large diversity of districts with respect to the level of socio-economic development. Wawrzyniak and Batóg (2014) noted a strong polarization of these districts in four areas (demography, urbanization, structure of economic entities, labour market) in relation to Szczecin (which was assumed as the main centre of regional development). The greatest similarity to Szczecin (i.e. the smallest polarization) in all areas was characteristic for other cities with district status, which coincides with the conclusions of this research.

5. Conclusion

The empirical research confirmed the usefulness of the Kohonen network to divide the districts of the Zachodniopomorskie Voivodeship into clusters. In the research, the hypothesis that the separated clusters include districts with a similar level of socio-economic development, was confirmed.

Socio-economic development is characterised by the use of many different features, and Kohonen networks cope well with this multidimensionality. When using Kohonen networks to divide multi-feature objects into typological groups, just like in statistical cluster analysis, an important step is to properly prepare the collected data so that they do not duplicate information and have high diagnostic value. The comparability of the variables should also be ensured and their nature taken into consideration – the variables should be normalised and the destimulants transposed into stimulants.

The constructed and trained networks were positively verified on the basis of the universal measure of the network quality – mean absolute error. Designated networks differ in the structure of the output layer – a different number of neurons means a different number of clusters determined using a given network. With an increase in the number of output neurons, the values of network errors are reduced, which indicates an improved distribution of the districts in the clusters. On the other hand, smaller and smaller changes in MAE values and the decreasing diversity of the averages of the variables for the clusters and their weighted averages, indicate that there is no need to include too many clusters in the analysis. A comparison of the averages of the variables for these clusters also shows that they have more diverse values when neurons of the output layer are arranged linearly in one column or row.

The application of the Kohonen networks allowed the division of the voivodeship into disjointed groups of similar districts. In this division, regardless of the maximum number of clusters, the cluster of the cities: Szczecin (the capital of the voivodeship), Koszalin and Świnoujście, is particularly clear. They are characterised by a higher level of socio-economic development than other districts, which is confirmed by the analysis of the averages of the variables for the clusters. The second clear cluster, which emerged when dividing the districts into four groups, includes the following: białogardzki, choszczeński, drawski, myśliborski, szczecinecki, wałecki. The analysis of the averages of the variables indicates that these districts have the lowest level of development. Interestingly, the districts from most clusters (excluding cities) also form clear clusters on the map. Districts with the lowest level of development are located at the southern and south-eastern edge of the voivodeship, at the border with neighbouring voivodeships, distant from Szczecin. Information about the existing disproportions in development (also related to the geographical location) may be an important feedback for the authorities of the voivodeship and the districts (e.g. when formulating their development strategies) and stimulating for residents and local entrepreneurs.

References

Azoff, E. (1994). *Neural Network Time Series Forecasting of Financial Markets*. Chichester: John Wiley & Sons.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford: University Press.

Brol, R. (1998). *Zarządzanie rozwojem lokalnym – studium przypadków*. Wrocław: Wydawnictwo Akademii Ekonomicznej we Wrocławiu.

Gately, E. (1999). *Sieci neuronowe. Prognozowanie finansowe i projektowanie systemów transakcyjnych*. Warszawa: WIG-Press.

Gibas, P., and Heffner, K. (2007). *Analiza ekonomiczno-przestrzenna*. Katowice: Wydawnictwo Akademii Ekonomicznej w Katowicach.

Gorynia, M., and Łaźniewska, E. (2012). *Konkurencyjność regionalna. Koncepcje – strategie – przykłady*. Warszawa: PWN.

Gorzelak, G., and Smętkowski, M. (2019). *Rozwój regionalny, Polityka regionalna*. Forum Obywatelskiego Rozwoju. Retrieved November 10, 2019 from: https://for.org.pl/pl/publikacje/raporty-for/raport-for-rozwoj-regionalny-polityka-regionalna

Grabiński, T. (1992). *Metody taksonometrii*. Kraków: Akademia Ekonomiczna w Krakowie.

Kogut-Jaworska, M., Szewczuk, A., and Zioło, M. (2011). *Rozwój lokalny i regionalny. Teoria i praktyka*. Warszawa: C.H. Beck.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59-69.

Kukuła, K. (2000). *Metoda unitaryzacji zerowanej*. Warszawa: PWN.

Lula, P., and Tadeusiewicz, R. (2001). *Wprowadzenie do sieci neuronowych*. Kraków: StatSoft.

Lula, P., Tadeusiewicz, R., and Wójtowicz, P. (1999). *Sieci neuronowe. Materiały na seminarium*. Kraków: StatSoft.

Masters, T. (1996). *Sieci neuronowe w praktyce*. Warszawa: Wydawnictwa Naukowo-Techniczne.

Nizam, M. (2010). Kohonen neural network clustering for voltage control in power systems. *Telkomnika: Indonesian Journal of Electrical Engineering*, *8*(2), 115-122.

Nowak, E. (1990). *Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych*. Warszawa: PWE.

Oesterreich, M., Perzyńska, J., and Barej-Kaczmarek, E. (2019). Application of the TOPSIS procedure for the evaluation of socio-economic development of the West Pomeranian Voivodeship districts in 2004-2017. *Zeszyty Naukowe Uniwersytetu Przyrodniczo-Humanistycznego w Siedlcach. Seria: Administracja i Zarządzanie*, *49*(122), 79-88.

Panek, T., and Zwierzchowski J. (2013). *Statystyczne metody wielowymiarowej analizy porównawczej. Teoria i zastosowania*. Warszawa: Oficyna Wydawnicza SGH.

Sobiechowska-Ziegert, A., and Mikulska, A. (2013). Measure of the level of socio-economic development in provinces. *Quantitative methods in Economics*, *14*(2), 200-209.

Statistical Office in Szczecin. (2019). *Zachodniopomorskie Voivodeship. Subregions, Powiats, Gminas*. Retrieved July 19, 2019 from: https://szczecin.stat.gov.pl/publikacje-i-foldery/roczniki-statystyczne/wojewodztwo-zachodniopomorskie-podregiony-powiaty-gminy-2019,7,19.html

Szirmai, A. (2015) Socio-economic development. Cambridge: Cambridge University Press.

Tarka, D. (2011). Własności cech diagnostycznych w badaniach typu taksonomicznego. *Ekonomia i zarządzanie*, *2*(4), 194-205.

Timm N. H. (eds). (2002). *Cluster analysis and multidimensional scaling*. *Applied multivariate analysis. Springer texts in statistics*. New York: Springer.

Wawrzyniak, K., and Batóg, B. (2014). Polaryzacja powiatów województwa zachodniopomorskiego według wybranych kategorii ekonomicznych. *Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania*, *36*(2), 399-414.

Witkowska, D. (2002). *Sztuczne sieci neuronowe i metody statystyczne. Wybrane zagadnienia finansowe*. Warszawa: C.H. Beck.

ZASTOSOWANIE SIECI KOHONENA DO GRUPOWANIA POWIATÓW WOJEWÓDZTWA ZACHODNIOPOMORSKIEGO POD WZGLĘDEM POZIOMU ROZWOJU SPOŁECZNO-GOSPODARCZEGO

Streszczenie: W artykule przedstawiono możliwości wykorzystania sztucznych sieci neuronowych w wielowymiarowej analizie – analizie skupień. Ilustracją rozważań teoretycznych jest badanie empiryczne, w którym obiektami badań są powiaty województwa zachodniopomorskiego. W badaniu wykorzystano dane statystyczne z wielu obszarów dotyczących rozwoju społeczno-gospodarczego,

takich jak: demografia, rynek pracy, środowisko naturalne, kultura i rekreacja, infrastruktura społeczna i techniczna, gospodarka. Celem pracy był podział województwa zachodniopomorskiego na rozłączne grupy typologiczne powiatów za pomocą sieci Kohonena (map samoorganizujących). Skonstruowano i nauczono kilkanaście sieci różniących się strukturą warstwy wyjściowej. Ich wartościami wejściowymi były wybrane charakterystyki rozwoju społeczno-gospodarczego powiatów. Przy użyciu zweryfikowanych sieci utworzono różne zestawy grup badanych obiektów. Przeprowadzone badanie potwierdziło, że sieci Kohonena są użytecznym narzędziem wyodrębniania skupień powiatów podobnych do siebie pod względem poziomu rozwoju społeczno-gospodarczego.

Słowa kluczowe: rozwój społeczno-gospodarczy, powiat, sieć Kohonena, skupienie, województwo zachodniopomorskie.