

**PRACE NAUKOWE**

Uniwersytetu Ekonomicznego we Wrocławiu

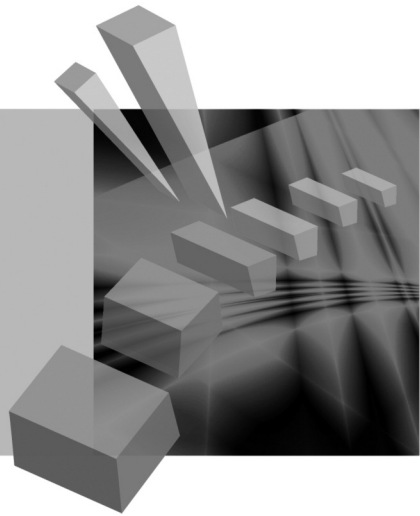
**RESEARCH PAPERS**

of Wrocław University of Economics

**242**

# **Taksonomia 19.**

## **Klasyfikacja i analiza danych – teoria i zastosowania**



Redaktorzy naukowi  
**Krzysztof Jajuga**  
**Marek Walesiak**



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu  
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pocięcha,  
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS  
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie [www.ibuk.pl](http://www.ibuk.pl)

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych  
The Central European Journal of Social Sciences and Humanities <http://cejsh.icm.edu.pl>  
oraz w The Central and Eastern European Online Library [www.ceeol.com](http://www.ceeol.com),  
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon [http://kangur.uek.krakow.pl/  
bazy\\_ae/bazekon/nowy/index.php](http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php)

Informacje o naborze artykułów i zasadach recenzowania znajdują się  
na stronie internetowej Wydawnictwa  
[www.wydawnictwo.ue.wroc.pl](http://www.wydawnictwo.ue.wroc.pl)

Kopowanie i powielanie w jakiegokolwiek formie  
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu  
Wrocław 2012

**ISSN 1899-3192** (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)  
**ISSN 1505-9332** (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM  
Nakład: 320 egz.

## Spis treści

<b>Wstęp</b> .....	13
<b>Stanisława Bartosiewicz</b> , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej .....	17
<b>Andrzej Sokolowski</b> , Q uniwersalna miara odległości .....	22
<b>Eugeniusz Gatnar</b> , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP) .....	31
<b>Marek Walesiak</b> , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
<b>Krzysztof Jajuga, Marek Walesiak</b> , XXV lat konferencji taksonomicznych – fakty i refleksje .....	47
<b>Józef Pocięcha, Barbara Pawelek</b> , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne .....	50
<b>Paweł Lula</b> , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych .....	58
<b>Ewa Roszkowska</b> , Zastosowanie metody TOPSIS do wspomagania procesu negocjacji.....	68
<b>Andrzej Młodak</b> , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne .....	76
<b>Andrzej Bąk</b> , Modele kategorii nieuporządkowanych w badaniach preferencji .....	86
<b>Jacek Kowalewski</b> , Zintegrowany model optymalizacji badań statystycznych.....	96
<b>Jan Paradysz, Karolina Paradysz</b> , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
<b>Tomasz Szubert</b> , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
<b>Izabela Szamrej-Baran</b> , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne .....	126
<b>Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski</b> , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
<b>Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz</b> , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych .....	144
<b>Hanna Dudek</b> , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów .....	153

<b>Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka</b> , Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
<b>Ewa Chodakowska</b> , Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
<b>Bartosz Soliński</b> , Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
<b>Krzysztof Szwarz</b> , Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
<b>Elżbieta Gołata, Grażyna Dehnel</b> , Rejestry administracyjne w analizie przedsiębiorczości.....	202
<b>Katarzyna Chudy, Marek Sobolewski, Kinga Stępień</b> , Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
<b>Katarzyna Dębowska</b> , Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
<b>Alina Bojan</b> , Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
<b>Justyna Brzezińska</b> , Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
<b>Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka</b> , Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
<b>Bartłomiej Jefmański</b> , Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
<b>Julita Stańczuk</b> , Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
<b>Jerzy Krawczuk</b> , Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
<b>Anna Czapkiewicz, Beata Basiura</b> , Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
<b>Radosław Pietrzyk</b> , Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
<b>Aleksandra Witkowska, Marek Witkowski</b> , Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
<b>Marcin Pelka</b> , Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
<b>Justyna Wilk</b> , Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

<b>Tomasz Bartłomowicz, Justyna Wilk</b> , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
<b>Kamila Migdał-Najman</b> , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących .....	342
<b>Dorota Rozmus</b> , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i> .....	352
<b>Krzysztof Najman</b> , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG .....	361
<b>Małgorzata Misztal</b> , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna .....	370
<b>Mariusz Kubus</b> , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
<b>Barbara Batóg, Jacek Batóg</b> , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym .....	387
<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej .....	396
<b>Iwona Staniec</b> , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach .....	406
<b>Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawelczyk, Jerzy Kołodziej, Jerzy Błaszczyk</b> , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami .....	416
<b>Iwona Foryś</b> , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
<b>Ewa Genge</b> , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
<b>Jerzy Korzeniewski</b> , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień .....	444
<b>Andrzej Dudek</b> , SMS – propozycja nowego algorytmu analizy skupień .....	451
<b>Artur Mikulec</b> , Metody oceny wyniku grupowania w analizie skupień.....	460
<b>Małgorzata Machowska-Szewczyk</b> , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych .....	469
<b>Artur Zaborski</b> , Analiza PROFIT i jej wykorzystanie w badaniu preferencji .....	479
<b>Karolina Bartos</b> , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena .....	488

<b>Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak</b> , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
<b>Izabela Kurzawa</b> , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś .....	505
<b>Aleksandra Łuczak, Feliks Wysocki</b> , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych .....	513
<b>Agnieszka Sompolska-Rzechuła</b> , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim .....	523
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk</b> , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego .....	532
<b>Iwona Bąk</b> , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę .....	541
<b>Aneta Becker</b> , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
<b>Katarzyna Dębowska</b> , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej .....	562
<b>Anna Domagała</b> , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
<b>Henryk Gierszal, Karina Pawlina, Maria Urbańska</b> , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej .....	580
<b>Hanna Gruchociak</b> , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
<b>Tomasz Klimanek, Marcin Szymkowiak</b> , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy .....	601
<b>Jarosław Lira</b> , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce .....	610
<b>Christian Lis</b> , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku .....	619
<b>Beata Bieszk-Stolorz, Iwona Markowicz</b> , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
<b>Lucyna Przezbórska-Skobiej, Jarosław Lira</b> , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
<b>Paweł Ulman</b> , Model rozkładu wydatków a funkcje popytu.....	646
<b>Maria Urbańska, Tadeusz Mizera, Henryk Gierszal</b> , Zastosowanie metod analizy statystycznej w badaniach mięczaków .....	655

## Summaries

<b>Stanisława Bartosiewicz</b> , The effects of subjectivism in multivariate analysis revisited.....	21
<b>Andrzej Sokółowski</b> , Q universal distance measure .....	30
<b>Eugeniusz Gatnar</b> , Data quality in central banks' statistical systems (NBP example) .....	38
<b>Marek Walesiak</b> , Distance measures for ordinal data – strategies of proceedings.....	46
<b>Krzysztof Jajuga, Marek Walesiak</b> , XXV years of taxonomic conferences – some facts and remarks.....	49
<b>Józef Pocięcha, Barbara Pawelek</b> , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
<b>Paweł Lula</b> , Learning-based systems of information extraction from textual resources .....	67
<b>Ewa Roszkowska</b> , The application of the TOPSIS method to support the negotiation process .....	75
<b>Andrzej Młodak</b> , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
<b>Andrzej Bąk</b> , Models for unordered categories in preference analysis.....	95
<b>Kowalewski Jacek</b> , An integrated model of optimizing statistical surveys ....	105
<b>Jan Paradysz, Karolina Paradysz</b> , Areas of unemployment in Poland – benchmark problem .....	115
<b>Tomasz Szubert</b> , How to play to lose the least? Classification of systems in sports bets .....	125
<b>Izabela Szamrej-Baran</b> , Classification of EU member states in view of fuel poverty .....	134
<b>Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski</b> , An attempt to use the gravity model in the analysis of commuters.....	143
<b>Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz</b> , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households .....	152
<b>Hanna Dudek</b> , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
<b>Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka</b> , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study .....	172
<b>Ewa Chodakowska</b> , Selected methods of classification in schools' rating.....	181
<b>Bartosz Soliński</b> , Renewable energy sector in the European Union – classification in the light of change management strategy .....	191
<b>Krzysztof Szwarc</b> , Classification of Wielkopolska voivodeship due to the demographic situation .....	201

<b>Elżbieta Gołata, Grażyna Dehnel</b> , Administrative registers in business analysis.....	211
<b>Katarzyna Chudy, Marek Sobolewski, Kinga Stępień</b> , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
<b>Katarzyna Dębowska</b> , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
<b>Alina Bojan</b> , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
<b>Justyna Brzezińska</b> , Log-linear analysis in the study of mortality in EU.....	246
<b>Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka</b> , Latent class analysis in student satisfaction surveys.....	254
<b>Bartłomiej Jefmański</b> , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
<b>Julita Stańczuk</b> , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
<b>Jerzy Krawczuk</b> , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
<b>Anna Czapkiewicz, Beata Basiura</b> , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
<b>Radosław Pietrzyk</b> , Timing and selectivity in mutual funds performance measurement.....	305
<b>Aleksandra Witkowska, Marek Witkowski</b> , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
<b>Marcin Pelka</b> , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
<b>Justyna Wilk</b> , Comparative study of symbolic data classification software.....	332
<b>Tomasz Bartłomowicz, Justyna Wilk</b> , Application of symbolic data analysis methods for domain database searching.....	341
<b>Kamila Migdał-Najman</b> , A proposal of hybrid clustering method based on self-learning networks.....	351
<b>Dorota Rozmus</b> , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
<b>Krzysztof Najman</b> , A dynamic grouping based on self-learning GNG networks.....	369
<b>Małgorzata Misztal</b> , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
<b>Mariusz Kubus</b> , The application of pre-conditioning of explanatory variable for feature selection.....	386
<b>Barbara Batóg, Jacek Batóg</b> , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395



<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
<b>Iwona Staniec</b> , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
<b>Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk</b> , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
<b>Iwona Foryś</b> , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
<b>Ewa Genge</b> , Trimming approach to the mixtures of normal distributions.....	443
<b>Jerzy Korzeniewski</b> , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
<b>Andrzej Dudek</b> , SMS – proposal of new clustering algorithm.....	459
<b>Artur Mikulec</b> , Evaluation methods for the grouping result in cluster analysis.....	468
<b>Małgorzata Machowska-Szewczyk</b> , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
<b>Artur Zaborski</b> , PROFIT analysis and its using in the research of preferences.....	487
<b>Karolina Bartos</b> , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
<b>Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak</b> , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
<b>Izabela Kurzawa</b> , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
<b>Aleksandra Luczak, Feliks Wysocki</b> , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
<b>Agnieszka Sompolska-Rzechuła</b> , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk</b> , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
<b>Iwona Bąk</b> , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
<b>Aneta Becker</b> , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
<b>Katarzyna Dębowska</b> , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

---

<b>Anna Domagała</b> , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
<b>Henryk Gierszal, Karina Pawlina, Maria Urbańska</b> , Statistical analysis in demand research of ICT services in mobile networks.....	589
<b>Hanna Gruchociak</b> , Construction of regression estimator for two-level data	600
<b>Tomasz Klimanek, Marcin Szymkowiak</b> , Application of spatial models in indirect estimation of some labor market characteristics .....	609
<b>Jarosław Lira</b> , Forecasting of hog livestock production profitability in Poland .....	618
<b>Christian Lis</b> , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports .....	627
<b>Beata Bieszk-Stolorz, Iwona Markowicz</b> , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers .....	636
<b>Lucyna Przezbórska-Skobiej, Jarosław Lira</b> , Agritourism space of Poland and its valuation.....	645
<b>Paweł Ulman</b> , Model of expenses distribution and demand functions.....	654
<b>Maria Urbańska, Tadeusz Mizera, Henryk Gierszal</b> , Methods of statistical analysis in research of molluscs .....	663

**Kamila Migdał-Najman**

Uniwersytet Gdański

---

## PROPOZYCJA HYBRYDOWEJ METODY GRUPOWANIA OPARTEJ NA SIECIACH SAMOUCZĄCYCH

---

**Streszczenie:** W artykule autorka dokonuje prezentacji hybrydowej metody grupowania opartej na sieciach neuronowych samouczących typu SOM i GNG. Autorka weryfikuje potencjał proponowanej hybrydowej metody grupowania typu: sieć SOM + metoda  $k$ -średnich. Proponowane podejście weryfikuje na przykładzie badania preferencji i zachowań komunikacyjnych mieszkańców Gdyni w 2010 r.

**Słowa kluczowe:** sieć samoorganizująca się Kohonena (SOM), sieć typu gaz neuronowy (GNG), wskaźniki jakości grupowania, badanie preferencji, segmentacja.

### 1. Wstęp

Sztuczna sieć neuronowa typu SOM (*Self Organizing Map*), nazywana również siecią samouczącą, siecią lub mapą Kohonena, samoorganizującym się odwzorowaniem lub mapą cech (por. [Berthold, Hand 1999; Kohonen 1995;1997; 2001; 2006; Fort, Pagès 1996; Yin 2002; Fausett 1994; Deboeck, Kohonen 1998]), zaproponowana została w 1982 r. przez fińskiego profesora Teuvo Kohonena. Sieć SOM jest jedną z bardziej popularnych i efektywnych aplikacji *data mining*, która znajduje zastosowanie w zagadnieniach, takich jak: klasyfikacja, grupowanie, redukcja wymiarowości, wyszukiwanie anomalii i odchyłeń od wartości typowych, wizualizacja wielowymiarowych zbiorów danych i badanie dynamiki zjawisk (por. [Papadimitriou i in. 2002; Delgado 2000; Fessant, Midenet 2002; Deventer, Moolman, Aldrich 1996; Migdał-Najman, Najman 2004]). W wyjściowej przestrzeni cech umieszczone zostają neurony, które lokalnie aproksymują analizowane obiekty. Neurony te uporządkowane są w pewną strukturę nazywaną siecią, w której są ze sobą w określonych związkach. Każdy wektor wejściowy połączony jest równolegle z wszystkimi neuronami na sieci przez wektor wag (współrzędnych). Wagi początkowo są liczbami losowymi z przedziału od zera do jednego. Dopasowanie tych wag jest istotą mechanizmu uczenia sieci. Uczenie sieci można nazwać adaptacyjnym procesem doboru wag sieci. Neurony, które są sąsiadami w przestrzeni, wykazują skłonność

do rozpoznawania podobnych (bliskich) do siebie obiektów wejściowych. Neurony sąsiadujące ze sobą na mapie mają podobne wektory wag. Każdy neuron, a dokładnie jego wagi, staje się pewnym wzorcem grupy bliskich sobie sygnałów (obiektów) wejściowych.

Jednym ze sposobów wizualizacji wyników sieci SOM jest macierz  $U$  (*unified distance matrix*, *U-matrix*), nazywana macierzą ujednoczonych odległości. Odległość między sąsiadującymi neuronami prezentowana jest różną kolorystyką. Ciemna kolorystyka<sup>1</sup> między neuronami odpowiada dużym odległościom, a zatem luce między wektorami wag w przestrzeni wejściowej. Jasna kolorystyka między neuronami oznacza, że wektory wag są blisko innych w przestrzeni wejściowej (por. [Fessant, Midenet 2002]). Technika ta umożliwia poszukiwanie skupień w danych wejściowych bez posiadania *a priori* żadnych informacji o tych klasach, ujawniając na mapie „pasma gór” i „wąwozów”. Pierwsze są często strefą nieregularnie ukształtowaną z wysoką tendencją do tworzenia skupień, podczas gdy drugie rozdzielają zbiór danych na obszary, które mają odmienne właściwości. Wizualna eksploracja macierzy ujednoczonych odległości w celu poszukiwania skupień, którą zaproponował Kohonen, jest jednak subiektywna. Zależy od doboru kolorów i umiejętności „czytania” mapy SOM. Jest ona również często niepraktyczna, szczególnie kiedy prowadzonych jest wiele badań i budowanych jest wiele sieci SOM. Aby zobiektywizować i zautomatyzować proces wyróżniania skupień w literaturze, proponuje się zastosowanie metody dwustopniowej. Na pierwszym stopniu buduje się sieć SOM, a na drugim uzyskane neurony klasyfikuje się metodą  $k$ -średnich. Dla uzyskanych skupień neuronów możliwa jest identyfikacja jednostek, które one odwzorowują. Podejście to ma jednak wady wynikające z zastosowanej na drugim stopniu metody  $k$ -średnich. Jest ona wrażliwa na początkowe centra skupień, wyróżnia jedynie skupienia sferyczne, może prowadzić do otrzymania skupień o zerowej liczbie jednostek. Wymaga także ustalenia *a priori* liczby skupień – aby proces ten zautomatyzować, trzeba testować różne konfiguracje skupień, a następnie uzyskaną strukturę trzeba sprawdzić jednym z kilkudziesięciu wskaźników jakości grupowania. Dla uniknięcia tych niedogodności proponuje się, aby na drugim stopniu użyć innej metody grupowania, która nie ma wad metody  $k$ -średnich. Celem artykułu jest propozycja zautomatyzowanej metody wyróżniania skupień na neuronach sieci SOM na bazie sieci neuronowej o zmiennej strukturze typu GNG (*Growing Neural Gas*) B. Fritzkego (por. [Fritzke 1994; 1995; Kohonen 1995; 1999; 2001; 2006; Vesanto 1997; Deboeck, Kohonen 1998; Migdał-Najman 2009]). Uzyskana klasyfikacja dla zaproponowanej hybrydy: sieć SOM + sieć GNG, porównana zostanie z podejściem: sieć SOM + metoda  $k$ -średnich. Propozycja zautomatyzowanej metody zaprezentowana zostanie na zbiorze danych przedstawiających preferencje i zachowania komunikacyjne mieszkańców Gdyni.

---

<sup>1</sup> Jest to sprawa umowna, kolory w macierzy  $U$  można oznaczyć całkowicie odmiennie.

## 2. System transportu publicznego w Gdyni

Od 1998 r. Zarząd Komunikacji Miejskiej (ZKM) w Gdyni przy współpracy z Uniwersytetem Gdańskim co dwa lata prowadzi badanie preferencji i zachowań komunikacyjnych mieszkańców Gdyni. Ocenia różne aspekty komunikacji miejskiej oraz poglądy użytkowników dotyczące określonych rozwiązań polityki transportowej realizowanej przez władze samorządowe. Badania są prowadzone na podstawie reprezentatywnej próby mieszkańców Gdyni, biorąc pod uwagę liczbę ludności w poszczególnych dzielnicach miasta, proporcje kobiet i mężczyzn, a także ich wykształcenie. Liczebność próby w kolejnych edycjach badania różni się nieznacznie i oscyluje na poziomie ok. 2000 osób. Struktura kwestionariusza jest stabilna i liczy do 45 pytań.

System transportu publicznego w Gdyni tworzą dwa podsystemy: sieci linii drogowej transportu zbiorowego (trolejbusy i autobusy) i linia szybkiej kolei miejskiej (SKM). Podsystemy te nie są zintegrowane. Drogowy transport zbiorowy jest organizowany przez Zarząd Komunikacji Miejskiej (ZKM) w Gdyni – zakład budżetowy gminy Gdynia. SKM jest organizowana i obsługiwana przez PKP Szybką Kolej Miejską w Trójmieście sp. z o.o. Spółka ta funkcjonuje w ramach holdingu PKP SA. SKM na obszarze Gdyni obejmuje 16 km zelektryfikowanej dwutorowej linii z 9 przystankami: Orłowo, Redłowo, Wzgórze św. Maksymiliana, Główna, Stocznia, Grabówek, Leszczynki, Chylonia i Cisowa. Na rysunku 1 przedstawiono przebieg traktacji SKM na terenie miasta Gdynia.



**Rys. 1.** Przebieg traktacji SKM w Gdyni

Źródło: opracowanie własne.

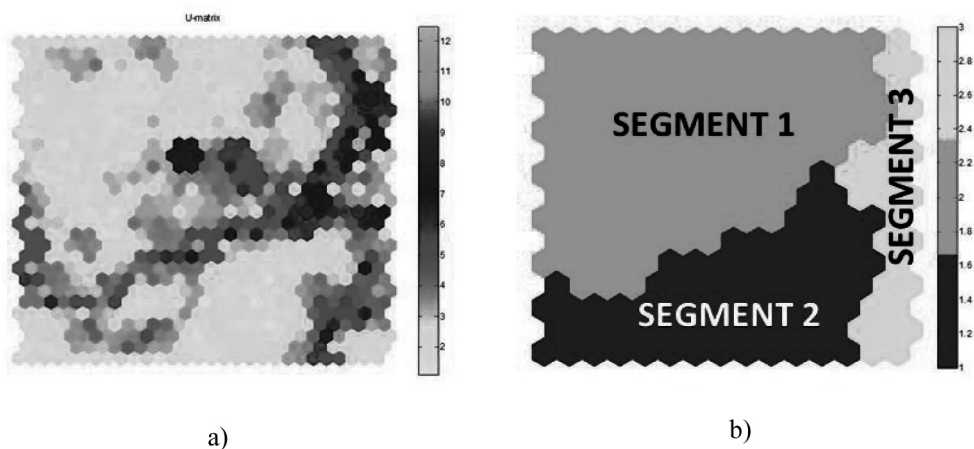
ZKM w Gdyni oferuje przewozy na obszarach: Gdyni, Sopotu, Rumii, Kosakowa, Żukowa, Wejherowa. Realizuje obsługę przewozową na 90 liniach, w tym:

12 trolejbusowych, 78 autobusowych, w tym: 64 zwykłych, 6 nocnych, 3 bezpłatnych, 3 specjalnych i 2 wodnych.

### 3. Eksperyment badawczy i wyniki analizy porównawczej

W celu wyróżnienia jednorodnych grup użytkowników komunikacji miejskiej w Gdyni ze względu na ich opinie dotyczące podstawowych aspektów podróżowania autobusem, trolejbusem i SKM wykorzystano samouczącą się sieć neuronową typu SOM. Wybór metody wynikał z jej własności, a także struktury zbioru danych. Zbiór składał się z 1975 jednostek, którymi są badane osoby i 18 cech wyrażonych na skali porządkowej (skala 5-stopniowa) opisujących ich opinie w zakresie: cen biletów, szybkości, wygody, czystości, punktualności i bezpieczeństwa podróżowania (np. 1 – bardzo niebezpieczny, 2 – niebezpieczny, 3 – ani niebezpieczny, ani bezpieczny, 4 – bezpieczny, 5 – bardzo bezpieczny). W zbiorze znajdowała się niewielka liczba braków danych.

**Pierwsza hybryda.** Przy budowie sieci SOM testowano różne jej topologie. Budowano sieci o heksagonalnej strukturze połączeń o rozmiarach od  $10 \times 10$  do  $15 \times 15$ . Uwzględniano cztery funkcje sąsiedztwa: gaussowską, uciętą gaussowską, prostokątną i wykładniczą. W każdym wariantcie uczono sieć od 1000 do 5000 iteracji. Dla każdej symulacji wyznaczano trzy miary jakości sieci: błąd kwantyzacji, topograficzny i dystorsji. Optymalną w sensie minimalizacji miar błędów siecią jest sieć SOM o następujących parametrach: typ połączeń neuronów: heksagonalny, rozmiar sieci:  $15 \times 15$  neuronów, funkcja sąsiedztwa: uciętą gaussowska, błąd kwantyzacji: 1,54, błąd topograficzny: 0,26, błąd dystorsji: 3,44, liczba iteracji uczących: 5000. Macierz U i uzyskaną klasyfikację metodą  $k$ -średnich na neuronach sieci SOM przedstawiono na rys. 2.



**Rys. 2.** a) macierz U i b) uzyskana klasyfikacja metodą  $k$ -średnich na neuronach sieci SOM

Źródło: opracowanie własne.

W wyniku zastosowania metody  $k$ -średnich na neuronach sieci SOM wyodrębniono trzy segmenty podróżujących komunikacją zbiorową. Liczbę badanych w wyróżnionych segmentach przedstawiono w tab. 1.

**Tabela 1.** Liczba badanych w wyróżnionych segmentach dla hybrydy: sieć SOM + metoda  $k$ -średnich

Segment	Segment 1 – S1	Segment 2 – S2	Segment 3 – S3	Razem
Liczba badanych w segmentach	834	606	535	1975

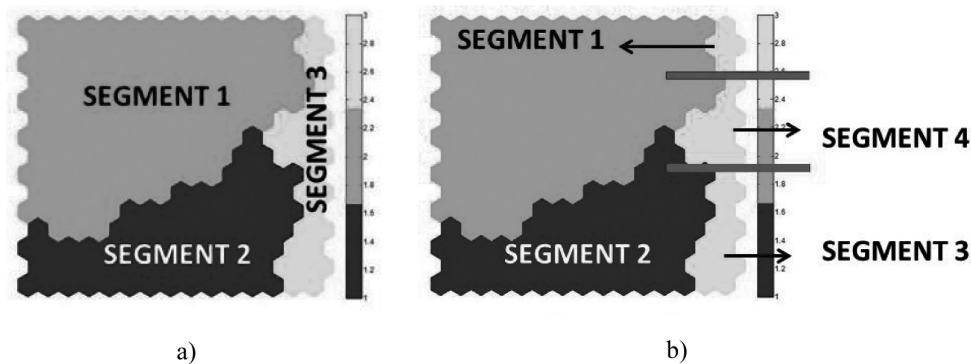
Źródło: opracowanie własne.

**Druga hybryda.** Drugą klasyfikację podróżujących komunikacją zbiorową w Gdyni uzyskano dla hybrydy: sieć SOM + sieć GNG. Parametry optymalnej w sensie minimalizacji miar błędów sieci SOM przedstawiono powyżej (por. hybryda pierwsza). Na etapie drugim na neuronach sieci SOM przeprowadzono grupowanie za pomocą sieci GNG. Przy budowie sieci GNG testowano: parametr lambda: 90 (liczba iteracji, po których jest wstawiany nowy neuron), maksymalny wiek połączenia: 89 (ile iteracji pozostaje neuron, który się nie uczy), krok uczenia neuronu zwycięzcy: 0,05, krok uczenia neuronów połączonych ze zwycięzcą: 0,006. Sieć GNG składała się z 35 neuronów. Ze względu na brak możliwości wizualizacji uzyskanej klasyfikacji na sieci GNG wynik tego podejścia przedstawiono na uzyskanej w kla-

**Tabela 2.** Liczba badanych w wyróżnionych segmentach dla hybrydy: sieć SOM + sieć GNG

Segment	Segment 1 – S1	Segment 2 – S2	Segment 3 – S3	Segment 4 – S4	Razem
Liczba badanych w segmentach	878	606	136	535	1975

Źródło: opracowanie własne.



**Rys. 3.** Klasyfikacja a) metodą  $k$ -średnich na neuronach sieci SOM i b) siecią GNG na neuronach sieci SOM

Źródło: opracowanie własne.



syfikacji metodą  $k$ -średnich na neuronach sieci SOM. W wyniku zastosowania sieci GNG na neuronach sieci SOM wyróżniono 4 segmenty podróżujących komunikacją zbiorową. Liczbę badanych w wyróżnionych segmentach przedstawiono w tab. 2. Uzyskane klasyfikacje dla dwóch proponowanych podejść przedstawiono na rys. 3.

**Analiza porównawcza.** Porównując dwie uzyskane klasyfikacje, można zauważyć, że w wyniku zastosowania hybrydy: sieć SOM + sieć GNG wyróżniono o jeden segment więcej niż w hybrydzie: sieć SOM + metoda  $k$ -średnich. Co ciekawe, sieć GNG wyróżniła dwa skupienia (segment 1 i segment 2) identyczne jak metoda  $k$ -średnich. Natomiast segment 3, który wyróżniono w metodzie  $k$ -średnich, został przez sieć GNG podzielony na trzy części. 44 osoby tego segmentu sieć GNG zaklasyfikowała do skupienia 1, a pozostałe zostały rozdzielone na dwie grupy i przydzielone do segmentu 3 i segmentu 4. W wyniku zastosowania sieci GNG na neuronach sieci SOM udało się zauważyć ważne różnice między segmentami, które zostały „zatarłe” w wyróżnionym segmencie 3 w metodzie  $k$ -średnich. Klasyfikację krzyżową dla dwóch proponowanych hybryd przedstawiono w tab. 3.

W tabeli 4 przedstawiono wskaźniki oceny podobieństwa dwóch porównywanych klasyfikacji: wskaźnik Jaccarda, Randa, korygowany Randa i Fowlkesa Mallowsa (por. [Jaccard 1908; Rand 1971; Fowlkes, Mallows 1983]).

Uzyskane wyniki wskazują na istotne różnice w uzyskanych podziałach. Ich ocena będzie wynikała z możliwości interpretacji uzyskanych różnic.

**Tabela 3.** Klasyfikacja krzyżowa dla dwóch hybryd

Hybrydy		SOM + GNG				Razem
		S1	S2	S3	S4	
<b>SOM + metoda <math>k</math>-średnich</b>	S1	<b>834</b>	0	0	0	834
	S2	0	<b>606</b>	0	0	606
	S3	<b>44</b>	0	<b>136</b>	<b>355</b>	535
Razem		878	606	136	355	1975

Źródło: opracowanie własne.

**Tabela 4.** Wskaźniki oceny podobieństwa wyników grupowania

Porównywane hybrydy	Jaccard	Rand	korygowany Rand	Fowlkes Mallows
sieć SOM + $k$ -średnich/sieć SOM + sieć GNG	0,85	0,95	0,88	0,92

Źródło: opracowanie własne.

#### 4. Segmentacja podróżujących komunikacją zbiorową

W badaniu preferencji i zachowań komunikacyjnych mieszkańców Gdyni zastosowano hybrydę: sieć SOM + sieć GNG. W badaniu wyróżniono cztery segmenty



umożliwiający analizę preferencji i zachowań komunikacyjnych mieszkańców Gdyni. W tabeli 5 przedstawiono profile wyróżnionych segmentów.

**Tabela 5.** Charakterystyka wyróżnionych segmentów

Profil	Segment 1	Segment 2	Segment 3	Segment 4
Preferowany środek transportu	autobus, trolejbus, SKM	autobus trolejbus	autobus	nie mam zdania
Dzielnica	Chylonia, Cisowa, Grabówek, Działki Leśne, Karwiny, Leszczynki, Wzgórze św. Maksymiliana, Śródmieście, Redłowo, Orłowo, Witomino	Babie Doły, Cisowa, Dąbrowa, Karwiny, Leszczynki, Śródmieście	Oksywie, Obłuże, Pogórze	Mały Kack, Wielki Kack, Chwarzno-Wiczlino
Wiek	16-20 21-30 31-40	51-60 61-70 71-75	31-40 51-60 61-70	31-40 41-50 51-60
Płeć	54% kobiet	59% kobiet	57% kobiet	37% kobiet
Prawo jazdy	53%	46%	48%	93%
Status zawodowy	pracuje, pracuje i uczy się/ studiuje, uczy się / studiuje	pracuje, jest na emeryturze	pracuje, jest na emeryturze	pracuje
Realizacja podróży	zawsze komunikacją zbiorową, przeważnie komunikacją zbiorową	zawsze komunikacją zbiorową	zawsze komunikacją zbiorową, w równym stopniu komunikacją zbiorową i samochodem osobowym	zawsze samochodem osobowym
Cechy realizowane najlepiej	dostępność, częstotliwość, punktualność	dostępność, punktualność, częstotliwość	punktualność, dostępność, bezpośredniość	nie mogę ocenić najlepszej
Standard wygody	miejsca stojące w nieuciążliwych warunkach, przeważnie miejsca siedzące	przeważnie miejsca siedzące, miejsca stojące w nieuciążliwych warunkach	przeważnie miejsca siedzące, miejsca stojące w nieuciążliwych warunkach	nie mam zdania, przeważnie miejsca siedzące
Ocena gdyńskiej komunikacji miejskiej	dobra	dobra	dobra	dobra, nie jestem w stanie ocenić (42% respondentów)

Źródło: opracowanie własne.

Środki komunikacji zbiorowej, takie jak: autobus, trolejbus i SKM, zostały ocenione przez mieszkańców Gdyni „dobrze”. Większość badanych oceniła autobusy i trolejbusy jako neutralnie drogie, szybkie, wygodne, czyste, punktualne i bezpieczne. To, co różniło SKM od dwóch pozostałych środków transportu, to neutralna ocena czystości SKM i bezpieczeństwa. Oceny trzech analizowanych środków transportu przedstawiono w tab. 6.

**Tabela 6.** Ocena środków transportu komunikacji zbiorowej wśród mieszkańców Gdyni

Środek transportu	Autobus	Trolejbus	SKM
Cena	neutralna	neutralna	neutralna
Szybkość	szybki	szybki	szybki
Wygoda	wygodny	wygodny	wygodny
Czystość	czysty	czysty	neutralnie czysty
Punktualność	punktualny	punktualny	punktualny
Bezpieczeństwo	bezpieczny	bezpieczny	neutralnie bezpieczny

Źródło: opracowanie własne.

## 5. Wnioski

Liczne badania poświęcone sztucznej sieci neuronowej SOM wskazują na jej duży potencjał w zagadnieniach związanych z grupowaniem. Propozycja zautomatyzowanej metody grupowania neuronów sieci SOM oparta na sieci GNG wydaje się posiadać większy potencjał w rozpoznawaniu szczegółów struktury grupowej niż metoda  $k$ -średnich. Sieć GNG jest zaawansowanym narzędziem analizy skupień i w tym zakresie w większości przypadków jest skuteczniejsza od metody  $k$ -średnich. Zastosowanie sieci GNG zamiast metody  $k$ -średnich na neuronach sieci SOM pozwoliło uniknąć niedogodności metody  $k$ -średnich. Uczenie się sieci GNG jest szybsze od metody  $k$ -średnich i nie wymaga poświęcania dodatkowego czasu związanego z analizowaniem różnych konfiguracji skupień i stosowaniem wskaźników ustalania liczby skupień. Sieć GNG sama modyfikuje swoją strukturę, automatycznie ustala optymalną liczbę neuronów i automatycznie ustala liczbę skupień w zbiorze danych. Trudnością w zastosowaniu sieci samouczących się jest brak standardowego oprogramowania.

Zastosowanie proponowanego podejścia wyróżniania skupień opartego na sieciach samoorganizujących: sieci SOM i GNG pozwoliło zaobserwować interesujące prawidłowości w analizie preferencji i zachowań komunikacyjnych mieszkańców Gdyni. Szczegóły zaobserwowane dzięki sieci GNG, a pominięte przez metodę  $k$ -średnich są istotne. Wydaje się, że hybryda sieć SOM + sieć GNG jest wartościowym narzędziem analizy skupień, wartym dalszych badań.

## Literatura

- Berthold M., Hand D.J., *Intelligent Data Analysis* Springer-Verlag, Berlin Heidelberg 1999.
- Deboeck G., Kohonen T., *Visual Explorations in Finance with Self-Organizing Maps*, Springer-Verlag, London 1998.
- Delgado A., *Control of nonlinear systems using a self-organizing neural network*, „Neural Computing&Applications” 2000, no 9.
- Deventer J.S.J., Moolman D.W., Aldrich C., *Visualisation of plant disturbances using Self-Organizing Maps*, „Computers Chemical Engineering” 1996, no 20 .
- Fessant F., Midenet S., *Self-Organizing map for data imputation and correction in surveys*, „Neural Computing&Applications” 2002, no 10.
- Fausett L., *Fundamentals of Neural Networks, Architectures, Algorithms, and Applications*, Florida Institute of Technology, Prentice Hall International, Inc., 1994.
- Fowlkes E.B., Mallows C.L., *A method for comparing two hierarchical clusterings*, „Journal of the American Statistical Association” 1983, no 78, 383.
- Fort J.C., Pagès G., *About the Kohonen algorithm: strong or weak self-organization?*, „Neural Networks” 1996, no 9, 5.
- Fritzke B., *Growing cell structures – a self-organizing network for unsupervised and supervised learning*, „Neural Networks” 1994, no 7, 9 .
- Fritzke B., *A Growing Neural Gas Network Learns Topologies*, Advances in Neural Information Processing Systems, 7<sup>th</sup> edn., MIT Press, Redmond, Washington 1995.
- Jaccard P., *Nouvelles recherches sur la distribution florale*, [w:] Bulletin de la Société Vaudoise des Sciences Naturelles, 44, 1908.
- Kohonen T., *Self-Organizing Maps*, Springer Series in Information Sciences, Springer-Verlag, Berlin Heidelberg 1995; 1999; 2001.
- Kohonen T., *Self-organizing neural projections*, „Neural Networks” 2006, no 19, 6.
- Migdał-Najman K., Najman K., *Diagnozowanie kondycji finansowej spółek notowanych na GPW w Warszawie w oparciu o sieć SOM*, Zeszyty Naukowe nr 389, Rynek Kapitałowy. Skuteczne inwestowanie, część I, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin 2004.
- Migdał-Najman K., *Zastosowanie nienadzorowanych sieci neuronowych typu Growing Neural Gas w analizie skupień*, [w:] Taksonomia 16, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 47, Wydawnictwo UE, Wrocław 2009.
- Papadimitriou S., Mavroudi S., Vladutu L., Pavlides G., Bezerianos A., *The supervised network Self-Organizing Map for classification of large data sets*, „Applied Intelligence” 2002, no 16 .
- Rand W.M., *Objective criteria for the evaluation of clustering methods*, „Journal of the American Statistical Association” 1971, no 66, 336.
- Vesanto J., *Data Mining Techniques Based on the Self-Organizing Map*, Thesis for the degree of Master of Science in Engineering, Helsinki University of Technology, Department of Engineering Physics and Mathematics, Espoo, Finland 1997.
- Yin H., *Data visualization and manifold mapping using the ViSOM*, „Neural Network” 2002, no 15.

## **A PROPOSAL OF THE HYBRID CLUSTERING METHOD BASED ON SELF-LEARNING NETWORKS**

**Summary:** In the article a hybrid clustering method based on a self-learning neural networks, SOM and GNG, is presented. The author verified the potential of the proposed hybrid clustering method such as: network SOM + k-means method. The proposed approach is verified on the example of behavioral research and communication preferences of the inhabitants of Gdynia in 2010.

**Keywords:** Self Organizing Map (SOM), Growing Neural Gas (GNG), validity indexes, preferences research, segmentation.