

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

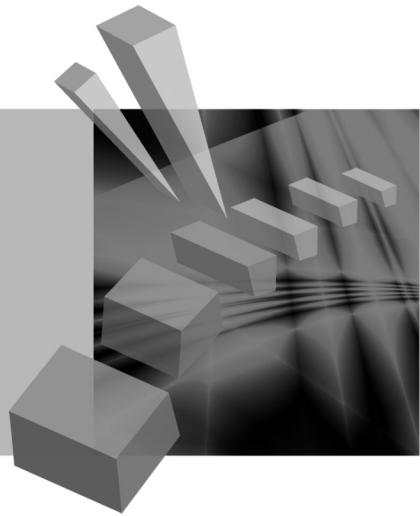
RESEARCH PAPERS

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

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,
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

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

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

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

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

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

Summaries

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

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

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

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

Andrzej Młodak

PWSZ im. Prezydenta Stanisława Wojciechowskiego w Kaliszu
Urząd Statystyczny w Poznaniu – Ośrodek Statystyki Miast

SĄSIEDZTWO OBSZARÓW PRZESTRZENNYCH W UJĘCIU FIZYCZNYM ORAZ SPOŁECZNO-EKONOMICZNYM – PODEJŚCIE TAKSONOMICZNE

Streszczenie: W pracy zaprezentowano porównanie koncepcji sąsiedztwa fizycznego oraz społeczno-gospodarczego obszarów przestrzennych. To pierwsze dotyczy ich położenia na mapie administracyjnej oraz ewentualnej wspólnoty granic, drugie zaś – podobieństwa w zakresie złożonego zjawiska społeczno-ekonomicznego. Praca opisuje klasyczną teorię sąsiedztwa i jego macierzy oraz ukazuje, jak można zaadaptować ją w przypadku wielowymiarowej analizy danych. Na zakończenie zaproponowano efektywną metodę porównywania obu typów sąsiedztwa wykorzystującą znaną z algebry postać normalną Smitha macierzy całkowitoliczbowych. Użyteczność rozpatrywanych modeli zbadano na przykładzie danych o rynku pracy dla Kalisza i gmin powiatu kaliskiego.

Słowa kluczowe: sąsiedztwo, macierz sąsiedztwa, odległość, metryka, postać normalna Smitha.

Pamięci prof. dra hab. Wiesława Wagnera

1. Wstęp

Pojęcie „sąsiedztwo” zazwyczaj kojarzy się z bliskim położeniem dwóch obiektów. Najpopularniejsze rozumienie tego terminu polega na tym, że obiekty owe są określonymi obszarami przestrzennymi, a to bliskie położenie jest najczęściej postrzegane w wymiarze fizycznym, tzn. w kontekście istniejących wspólnych granic wytyczonych na drodze realizacji decyzji administracyjnych lub uzgodnień politycznych. Założenia te stanowią także fundament funkcjonowania systemu badań, analiz i udostępniania danych statystycznych.

W dzisiejszych czasach sąsiedztwo fizyczne traci jednak na znaczeniu na rzecz sąsiedztwa społeczno-gospodarczego. Oznacza to, że przestrzenny kontekst sąsiedztwa ustępuje miejsca bliskości obiektów pod względem określonych zjawisk demo-

graficznych, ekonomicznych i innych. Wspomnijmy w tym miejscu, że odległości pomiędzy miejscowościami coraz częściej mierzy się czasem, jaki jest potrzebny na pokonanie dzielącego je dystansu (np. w kontekście organizacji ważnych imprez międzynarodowych). Tak więc dla kształtowania polityki rozwoju regionalnego sąsiedztwo społeczno-gospodarcze jest równie ważne jak fizyczne (a czasem nawet istotniejsze). W niektórych państwach statystyka publiczna wychodzi naprzeciw temu zapotrzebowaniu, np. w Wielkiej Brytanii istnieje system statystyki sąsiedztw (zob. <http://www.neighbourhood.statistics.gov.uk>), warto wspomnieć także o szerszych strefach miejskich (*Larger Urban Zone* – LUZ) obrazujących obszar funkcjonalnego oddziaływania miasta, a wyznaczanych przez grupowanie gmin pod względem odsetka osób dojeżdżających z nich do pracy w owym mieście (zob. np. [Rogalińska 2007; Młodak 2008]).

Konsekwencją tych tendencji stało się zainteresowanie badaczy metodologią tworzenia i analizy strukturalnej sąsiedztw. W ostatnich latach niektórzy specjaliści z zakresu wielowymiarowej analizy danych (jak np. zmarły w 2010 r. prof. dr hab. Wiesław Wagner) sporo uwagi poświęcali macierzy sąsiedztwa, zwanej także macierzą bliskości (*contiguity matrix*) obrazującej układ geograficznego sąsiedztwa obszarów przestrzennych wchodzących w skład określonego szerszego terytorium (np. gmin tworzących powiat), a także identyfikującej jednostki wewnętrzne i brzegowe takiego układu oraz ścieżki sąsiedztwa, tj. sposoby „przejścia” od jednej jednostki do drugiej przez kolejne jednostki sąsiadujące. Tak skonstruowaną macierz wykorzystywano pomocniczo, np. w analizie skupień (zob. np. [Wipperman 2004]).

W pracy zostanie przedstawiona podstawowa konstrukcja macierzy sąsiedztwa fizycznego (część 2) oraz sposób jej adaptacji dla oceny złożonych zjawisk społeczno-gospodarczych, opisywanych za pomocą wielu cech statystycznych (część 3). Dzięki temu możemy otrzymać dwie macierze sąsiedztwa: jedną bardziej naturalną, drugą – ekonomiczną. Zaprezentujemy efektywną metodę oceny podobieństwa obu macierzy z wykorzystaniem specyficznego narzędzia algebry macierzy całkowitoliczbowych, jaką jest postać normalna Smitha (w części 4). Odniesiemy się też do przykładu empirycznego – danych o rynku pracy dla Kalisza i gmin powiatu kaliskiego. Całość zwieńczy część 5 zawierająca podsumowanie przeprowadzonych rozważań.

2. Koncepcja macierzy sąsiedztwa fizycznego

Przedstawimy obecnie podstawowe założenia macierzy sąsiedztwa w ujęciu fizycznym. Wykorzystamy w tym celu idee zawarte w pracach W. Wagnera i A. Mantaja [2010], B.H. Wippermana [2004] oraz – opisowo – w książce B.S. Everitta i in. [2011].

Niech $\text{int } A$ oznacza wewnątrz nieskończonego zbioru A , zaś $\text{fr } A$ – jego brzeg (krawędź). Jeśli z kolei A jest skończony, to symbolem $|A|$ oznaczamy będziemy liczbę jego elementów. Założmy, że badany obszar przestrzenny U składa się z n (gdzie n jest liczbą naturalną) jednostek U_1, U_2, \dots, U_n . Zatem $\bigcup_{i=1}^n U_i = U$ i

$\text{int } U_i \cap \text{int } U_j = \emptyset$ dla każdych $i, j = 1, 2, \dots, n$, $i \neq j$. Jeśli $\text{fr } U_i \cap \text{fr } U_j \neq \emptyset$, to mówimy, że jednostka U_i sąsiaduje z jednostką U_j , co oznaczamy $U_i \bowtie U_j$, $i, j = 1, 2, \dots, n$, $i \neq j$. Zakłada się tu, że obszar U jest wypukły, tzn. od każdej jednostki U_i można przejść do innej, nie wykraczając poza U , $i = 1, 2, \dots, n$. Niech $\Gamma_{ij} = \text{fr } U_i \cap \text{fr } U_j$ będzie wspólną granicą jednostek U_i i U_j , a $\gamma_{ij} \geq 0$ – jej długością (jeśli $\Gamma_{ij} = \emptyset$, czyli jednostki U_i i U_j nie sąsiadują ze sobą, to $\gamma_{ij} = 0$), $i, j = 1, 2, \dots, n$, $i \neq j$.

Niech $U_{i_1}, U_{i_2}, \dots, U_{i_{k_i}}$ będą wszystkimi jednostkami sąsiadującymi z jednostką U_i , tzn. $U_i \bowtie U_j$ dla każdego $j \in \{i_1, i_2, \dots, i_{k_i}\}$ oraz $\text{fr } U_i \cap \text{fr } U_j = \emptyset$ dla każdego $j \notin \{i_1, i_2, \dots, i_{k_i}\}$ (gdzie $1 \leq k_i \leq n - 1$ jest liczbą naturalną). Jednostka U_i nazywana jest brzegową, gdy $\text{fr } U_i \setminus \bigcup_{j=1}^{k_i} (\text{fr } U_{i_j} \cap \text{fr } U_i) \neq \emptyset$. W przeciwnym razie, tzn. gdy $\text{fr } U_i \setminus \bigcup_{j=1}^{k_i} (\text{fr } U_{i_j} \cap \text{fr } U_i) = \emptyset$, jednostkę U_i określa się jako wewnętrzną, $i = 1, 2, \dots, n$. Zbiór wszystkich jednostek brzegowych oznaczać będziemy przez U_B , a wewnętrznych przez U_W . Stąd $U_B \cap U_W = \emptyset$ oraz $U_B \cup U_W = U$.

Ważną rolę w teorii sąsiedztw odgrywa *miara zagnieżdżenia jednostki wewnętrznej* $U_i \in U_W$. Opiera się ona na odległości obiektu U_i od obiektu $U_j \in U_B$ definiowanej jako najmniejsza liczba krawędzi oddzielających oba obiekty (najkrótsza *ścieżka przejścia*), tzn.

$$\delta_{ij} \stackrel{\text{def}}{=} \min_{k=1,2,\dots,n} |\Delta_{ijk}| - 1, \quad (1)$$

gdzie $\Delta_{ijk} = \{U_{i_1}, U_{i_2}, \dots, U_{i_k}: U_{i_c} \in U, i_c \in \{1, 2, \dots, n\}, c = 1, 2, \dots, k, U_{i_1} = U_i, U_{i_{c-1}} \bowtie U_{i_c}, c = 2, \dots, k, U_{i_k} = U_j\}$. Miara zagnieżdżenia minimalizuje tę wielkość: $\delta_i^* \stackrel{\text{def}}{=} \min_{j: U_j \in U_B} \delta_{ij}$. Rozmiar najdłuższej możliwej ścieżki przejścia $\delta_U \stackrel{\text{def}}{=} \max_{i,j: U_i, U_j \in U} \delta_{ij}$ nazywa się średnicą zbioru U . Formalnie rzecz ująwszy, obszar U jest wypukły, jeśli dla każdych $i, j \in \{1, 2, \dots, n\}$, $i \neq j$ istnieje takie $k \in \{2, 3, \dots, n\}$, że $\Delta_{ijk} > 0$.

Macierz sąsiedztwa $\mathbf{S} = [s_{ij}]$ rozmiaru $n \times n$ definiujemy jako

$$s_{ij} = \begin{cases} 1 & \text{gdy } i \neq j \text{ oraz } U_i \bowtie U_j, \\ 0 & \text{w przeciwnym razie,} \end{cases}$$

dla każdych $i, j = 1, 2, \dots, n$. A zatem jest to macierz symetryczna, na diagonalu ma zera, liczba elementów równych 1 jest dwukrotnością liczby wszystkich par obiektów sąsiadujących ze sobą (czyli $2s$, przy czym $s = |(i, j): i, j = 1, 2, \dots, n, i \neq j, U_i \bowtie U_j|$), $\mathbf{S} \cdot \mathbf{1} = \mathbf{k}$, gdzie $\mathbf{k} = (k_1, k_2, \dots, k_n)$ jest wektorem rozmiaru $1 \times n$ zawierającym liczby jednostek sąsiadujących z danymi jednostkami; w macierzy $\mathbf{T} \stackrel{\text{def}}{=} \mathbf{S} \cdot \mathbf{S}$ rozmiaru $n \times n$, $t_{ii} = k_i$, t_{ij} oznacza liczbę wspólnych sąsiadów dla pary jednostek U_i, U_j , jeśli nie są one oddzielone przez więcej niż jedną inną jed-

nostkę, oraz 0 w pozostałych przypadkach ($i, j = 1, 2, \dots, n, i \neq j$), przy czym $\text{tr}(\mathbf{T}) = 2s$ i $\det(\mathbf{T}) = \det^2(\mathbf{S})$.

3. Sąsiedztwo w sensie podobieństwa społeczno-gospodarczego

Załóżmy, że zjawisko społeczno-gospodarcze opisane jest za pomocą m (gdzie m jest liczbą naturalną) zmiennych X_1, X_2, \dots, X_m . Niech $\mathbf{X} = [x_{ij}]$ (x_{ij} – wartość zmiennej X_j dla jednostki U_i), $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$ rozmiaru $n \times m$ będzie macierzą danych dla naszych obiektów. Zakładamy przy tym milcząco, że zmienne te są zmiennymi diagnostycznymi i zostały poddane normalizacji¹. Definicję macierzy sąsiedztwa w tym przypadku oprzemy na zagregowanej odległości obiektów będącej funkcją $d: U \times U \rightarrow \mathbb{R}$, taką, że $d(U_i, U_i) = 0$ (zwrotność), $d(U_h, U_i) = d(U_i, U_h)$ (symetria) oraz $d(U_h, U_i) \geq 0$ (nieujemność) dla każdych $h, i = 1, 2, \dots, n$. Jeśli ostatni warunek zastąpimy silniejszą nierównością trójkąta ($d(U_h, U_i) + d(U_i, U_l) \geq d(U_h, U_l)$ dla każdych $h, i, l = 1, 2, \dots, n$), to funkcja d jest *metryką*. Można tu użyć np. odległości euklidesowej, odległości medianowej itp. (zob. [Młodak 2006]). Niech $d_{hi} \stackrel{\text{def}}{=} d(U_h, U_i)$, $h, i = 1, 2, \dots, n$ i $d_{\max} \stackrel{\text{def}}{=} \max_{h,i=1,2,\dots,n} d_{hi}$, $d_{\min} \stackrel{\text{def}}{=} \min_{h,i=1,2,\dots,n, h \neq i} d_{hi}$, $d_{\text{mm}} \stackrel{\text{def}}{=} \min_{h=1,2,\dots,n} \max_{i=1,2,\dots,n} d_{hi}$ oraz

$$\tilde{d} \stackrel{\text{def}}{=} \frac{d_{\max} - d_{\text{mm}}}{d_{\max} - d_{\min}} d_{\max}. \quad (2)$$

Macierz sąsiedztwa społeczno-gospodarczego $\tilde{\mathbf{S}} = [\tilde{s}_{ij}]$ rozmiaru $n \times n$ jest postaci

$$\tilde{s}_{hi} = \begin{cases} 1 & \text{gdy } h \neq i \text{ oraz } d_{hi} < \tilde{d}, \\ 0 & \text{w przeciwnym razie,} \end{cases}$$

dla każdych $i, h = 1, 2, \dots, n$. W klasycznej analizie skupień jako próg podobieństwa (lub jego braku) przyjmuje się czasem średnią arytmetyczną, medianę odległości lub minimaks (tzn. d_{mm}). Średnia arytmetyczna jest nazbyt wrażliwa na występowanie obserwacji odstających, mediana zakłada niejako „z góry” liczbę sąsiedztw, próg minimaksowy zaś na ogół wymusza posiadanie przez każdy obiekt przynajmniej jednego sąsiada. Rozwiązanie (2) jest zatem o wiele bardziej elastyczne. Zamiast miary odległości można by tutaj użyć też siatki kwantylowej o dostatecznie gęstych okach (zob. [Młodak 2011]). Opcja taka uniemożliwi jednak identyfikację pewnych cech sąsiedztwa, o których niżej.

¹ Oznacza to, że wybrano zestaw zmiennych wskaźnikowych opisujących dane zjawisko, a następnie podano go weryfikacji zmiennościowej (eliminacji zmiennych o zbyt niskiej wartości współczynnika zmienności, co sygnalizuje znikomą wartość analityczną z taksonomicznego punktu widzenia), korelacyjnej (eliminacja zmiennych nadmiernie skorelowanych z innymi, a zatem będącymi nośnikami podobnej informacji) oraz wybranej procedurze normalizacyjnej dla ujednoczenia mian (zob. np. A. Młodak [2006]).

Zauważmy, że w macierzy $\tilde{\mathbf{S}}$ mogą występować jednostki (lub podgrupy jednostek) niemające żadnych sąsiadów, czyli obszar U w tym ujęciu może nie być wypukły. Mamy też tutaj większy kłopot z ustaleniem, które jednostki są wewnętrzne, a które brzegowe. W tym celu należy analizować cykle sąsiedztwa obserwowane w macierzy $\tilde{\mathbf{S}}$. Niech q będzie liczbą naturalną, $3 \leq q \leq n$. Powiemy, że jednostki $U_{i_1}, U_{i_2}, \dots, U_{i_q}, i_c \in \{1, 2, \dots, n\}$, $c = 1, 2, \dots, q$ tworzą cykl sąsiedztwa długości q , jeśli $U_{i_c} \bowtie U_{i_{c+1}}$ dla każdego $c = 1, 2, \dots, q$, przy czym $i_{q+1} = i_1$. Jednostka U_i jest zatem wewnętrzna, jeśli jednostki z nią sąsiadujące tworzą cykl sąsiedztwa, czyli gdy submacierz sąsiedztwa dla pewnego podzbioru jej sąsiadów o liczebności $p \in \mathbb{N}$, $p \leq k_i$ ma postać $A = [a_{lj}]$, $l, j = 1, 2, \dots, p$, gdzie $a_{l(l+1)} = a_{(l+1)l} = a_{p1} = a_{1p} = 1$ dla $l = 1, 2, \dots, p-1$, a pozostałe elementy są zerami. Na przykład dla $p = 5$ wygląda to tak:

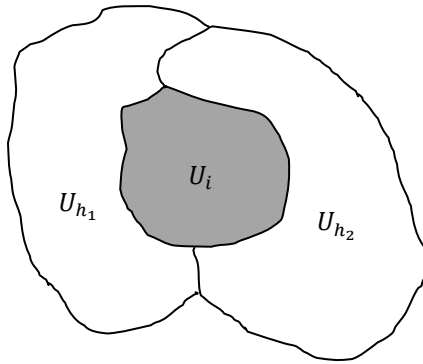
$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Z algebraicznego punktu widzenia jest to szczególnie przypadek macierzy Toeplitza (zob. np. [Eberle, Maciel 2003]). Można wykazać², że gdy $n > 2$, wartości własne takiej macierzy są postaci $\lambda_h = 2 \cos(2h\pi/n)$, $h = 1, 2, \dots, n$, a zatem jej wyznacznik jest równy $\det(A) = \prod_{h=1}^n \lambda_h = 2^n \prod_{h=1}^n \cos(2h\pi/n)$. Jednostka może być też wewnętrzna, gdy sąsiaduje jedynie z dwiema innymi jednostkami, które ją w całości otaczają, lub jeśli jest enklawą, tzn. zawiera się w całości w innej jednostce³. Sytuację ową łatwo zidentyfikować, gdy każda z tych jednostek sąsiadujących spełnia wyżej wskazany warunek jednostki wewnętrznej. Gorzej, gdy tak nie jest. Wtedy, zakładając, że odległość d jest metryką, dostrzegamy, iż jest ona również dobrze określona jako miara odległości podzbiorów płaszczyzny reprezentowanych przez badane obiekty (np. według formuły Hausdorffa). Wykorzystamy więc definicję zawierania się kuli w przestrzeni metrycznej o metryce d . Niech $\mathcal{K}(a, \varepsilon)$ będzie kulą o środku a i promieniu $\varepsilon > 0$ w tejże przestrzeni. Wówczas (zob. np. [Palmgren 2009]) $\mathcal{K}(a, \varepsilon) \subseteq \mathcal{K}(b, \eta) \Leftrightarrow d(a, b) + \varepsilon \leq \eta$. W naszym kontekście można użyć najmniejszych kul zawierających dane podzbiory, a te podzbiory potraktować jako ich „środki”. Jednostkę U_i będziemy więc uważać za zawartą w całości w jednostce U_h wtedy i tylko wtedy, gdy U_h jest jej jedynym są-

² Dowód – z wykorzystaniem równań różnicowych jednorodnych rzędu drugiego i pierwiastków z jedności – jest bardzo podobny jak w przypadku niecyklicznej macierzy Toeplitza, tzn. macierzy Toeplitza z niezerowymi elementami na głównej przekątnej i dwóch skośnych z nią sąsiadujących oraz zerami w pozostałych miejscach (zob. [Meyer 2000]).

³ W wymiarze fizycznym ma to często miejsce w przypadku miasta na prawach powiatu, które otaczane jest w całości przez odpowiedni powiat ziemski – tak jest np. w przypadku Konina czy Leszna w Wielkopolsce.

siadem oraz $d_{ih} < \text{med}_{k \in \{1,2,\dots,n\}, k \neq i, h} |d_{ik} - d_{hk}|$. Po prawej stronie mamy tu szacunek bezwzględnej różnicy promieni minimalnych kul zawierających podzbiory U_i i U_h , $i, h \in \{1,2, \dots, n\}$, $i \neq h$.



Rys. 1. Otaczanie jednostki przez dwie inne

Źródło: opracowanie własne.

Pozostaje jeszcze sytuacja, gdy jednostka U_i sąsiaduje w pełni z dwiema innymi jednostkami, które ją otaczają (zob. rys. 1). Wtedy macierz $\tilde{\mathbf{S}}$ wskazuje, że jednostka ta ma tylko dwóch sąsiadów, U_{h_1} i U_{h_2} . Na to, aby w istocie oni otaczali ją w całości, potrzeba i wystarcza, by minimalna kula otaczająca U_i była zawarta w analogicznej kuli dla $U_{h_1} \cup U_{h_2}$, a zatem – oznaczając $d_{k,h_1+h_2} \stackrel{\text{def}}{=} d(U_k, U_{h_1} \cup U_{h_2}) = \max\{d_{kh_1}, d_{kh_2}\}$ – aby spełniona była nierówność postaci: $d_{i,h_1+h_2} < \text{med}_{k=1,2,\dots,n, k \neq i, h_1, h_2} |d_{ik} - d_{k,h_1+h_2}|$, $i, h_1, h_2, k \in \{1,2, \dots, n\}$, $i \neq h_1, h_2, h_1 \neq h_2$.

Jednostkę, która nie jest wewnętrzna, należy uważać za brzegową. W przeciwieństwie do sąsiedztwa fizycznego wśród jednostek brzegowych mogą jednakże wystąpić jednostki izolowane. Otóż jednostka będzie izolowana, jeśli odpowiadający jej rząd i kolumna macierzy $\tilde{\mathbf{S}}$ są zerowe (co oznacza, że jednostka ta nie sąsiaduje z żadną inną jednostką). Może tutaj wystąpić także izolowane skupisko jednostek, tzn. właściwy podzbiór zbioru U składający się z takich jednostek, które sąsiadują ze sobą, ale żadna z nich nie sąsiaduje z jakąkolwiek jednostką spoza tego skupiska.

Na podstawie macierzy $\tilde{\mathbf{S}}$ można obliczać miary zagnieżdżenia obiektów (1), ale już idea długości granicy nie ma tutaj zastosowania. Jedynie dla jednostki wewnętrznej można wykorzystać strukturę bliskości (tj. odwrotność odległości powiększonej o 0,01 – by uniknąć zerowania się mianownika) w stosunku do jej tworzących cykl sąsiadów, np. jeśli dla U_i są to $U_{i_1}, U_{i_2}, \dots, U_{i_q}$, wtedy $\tilde{\gamma}_{ii_r} = ((1/(d_{ii_r} + 0,01)))/(\sum_{h=1}^q (1/(d_{ii_h} + 0,01))))\gamma_{ii_r}$, przy czym $r = 1,2, \dots, q$, $i, q, i_r \in \{1,2, \dots, n\}$.

4. Porównywanie sąsiedztwa fizycznego i społeczno-gospodarczego

Wyznaczenie macierzy \mathbf{S} i $\tilde{\mathbf{S}}$ pociąga za sobą w naturalny sposób pytanie o możliwość porównywania obiektów i struktury sąsiedztw w obu rozpatrywanych aspektach. Najprostsze, bezpośrednie podejście nie oddaje w pełni właściwości modelu, tzn. nie wskazuje, które obiekty wykazują największą, a które najmniejszą strukturalną różnicę sąsiedztwa w kontekście traktowania modelu jako integralnej całości oraz czy występuje wpływ identyczności sąsiadów (a więc swego rodzaju dublowania informacji) na kształt tychże różnic.

Dlatego też lepszym rozwiązaniem wydaje się tzw. postać normalna Smitha (*Smith normal form*) macierzy całkowitoliczbowej. Ujmując rzecz od strony formalnej, jeśli $\mathbf{A} = [a_{ij}]$, $i, j = 1, 2, \dots, n$ jest macierzą o elementach całkowitych, to istnieje macierz $\mathbf{A}^* = \text{diag}(a_1^*, a_2^*, \dots, a_n^*)$ równoważna macierzy \mathbf{A} (co oznacza, że macierz \mathbf{A}^* można otrzymać z \mathbf{A} przez wykonanie na niej ciągu operacji elementarnych – dodawania wierszy/kolumn, mnożenia ich przez stałą bądź permutacji) taka, że⁴ $a_i^* \in \mathbb{Z}$ dla każdego $i = 1, 2, \dots, n$, $a_1^* > 0, a_2^* > 0, \dots, a_r^* > 0$ dla $r = \text{rz}(\mathbf{A})$, $a_{r+1}^* = a_{r+2}^* = \dots = a_n^* = 0$ i $a_i^* | a_{i+1}^*$, $i = 1, 2, \dots, r - 1$. Liczby $a_1^*, a_2^*, \dots, a_r^*$ zwane są czynnikami niezmiennymi (*invariant factors*) lub dzielnikami elementarnymi (*elementary divisors*).

Dla macierzy całkowitoliczbowych jej postać normalna Smitha jest wyznaczona jednoznacznie (w odróżnieniu od np. diagonalizacji wartościami własnymi), ponadto zerowe wiersze/kolumny macierzy \mathbf{A}^* wskazują obiekty, które mają dokładnie takich samych sąsiadów jak inne – a więc niewnoszące żadnej dodatkowej informacji standaryzacyjnej. Konstrukcja macierzy \mathbf{A}^* wyzyskuje też kompleksowe własności modelu i zagregowane strukturyzacyjne różnice pomiędzy poszczególnymi obiektami. Ponadto konstrukcja postaci normalnej Smitha redukuje wpływ zasobu wspólnej informacji występującego między parami różnych obiektów. Wynika to z typowych algorytmów wyznaczania takiej macierzy opartych właśnie na ciągu operacji elementarnych (zob. np. [Havas, Majewski 1997]). Jedyną trudność stanowi fakt, że podczas wyznaczania postaci normalnej Smitha dokonuje się przedstawiania wierszy/kolumn, dlatego też trzeba rejestrować te operacje.

Niech wobec tego \mathbf{S}^* oraz $\tilde{\mathbf{S}}^*$ będą postaciami normalnymi Smitha macierzy \mathbf{S} i $\tilde{\mathbf{S}}$, zaś $\sigma_{\mathbf{S}}^{(r)}$, $\sigma_{\mathbf{S}}^{(c)}$, $\sigma_{\tilde{\mathbf{S}}}^{(r)}$, $\sigma_{\tilde{\mathbf{S}}}^{(c)}$ oznaczają odpowiednio ostateczne permutacje ich wierszy i kolumn. Wobec tego wkład jednostki U_i do odległości pomiędzy strukturami sąsiedztwa fizycznego i społeczno-gospodarczego definiujemy jako

$$d_i = \frac{d_i^{(h)} + d_i^{(v)}}{2}, \quad (3)$$

⁴ \mathbb{Z} oznacza zbiór liczb całkowitych.

gdzie $d_i^{(h)} = \left| s_{\sigma_S^*(t)}^* - \tilde{s}_{\sigma_S^*(t)}^* \right|$ oraz $d_i^{(v)} = \left| s_{\sigma_S^{(c)}(t)}^* - \tilde{s}_{\sigma_S^{(c)}(t)}^* \right|$ to poziome i pionowe składowe różnicy kompleksowej dla każdego $i = 1, 2, \dots, n$. Odległość kompleksową da się zatem obliczyć jako średnią arytmetyczną odległości (3), czyli

$$d = \frac{1}{n} \sum_{i=1}^n d_i. \quad (4)$$

Jako praktyczny przykład zastosowania proponowanych metod przeanalizujemy różnorakie aspekty sąsiedztwa gmin tworzących rejon Kalisza, tzn. miasto Kalisz oraz 11 gmin wchodzących w skład powiatu ziemskiego kaliskiego.

Okazuje się, że jedynie dwie gminy (Żelazków oraz Opatówek) są wewnętrzne, reszta zaś jest brzegowa. Głębiej zagnieżdżone gminy mają tutaj wyższe wartości współczynnika zagnieżdżenia (np. Żelazków i Opatówek po 0,5455) niż brzegowe (od 0,1818 dla Brzezin i Liskowa do 0,4545 dla Cekowa – Kolonii), ale nie jest to reguła: jeśli jednostka jest otoczona przez małą liczbę innych jednostek brzegowych, a te mają wielu sąsiadów, relacja ta może być odwrotna. Średnica rozpatrywanego zbioru jest niska i wynosi 5.

Aby przeanalizować problem sąsiedztwa badanych gmin w kontekście społeczno-ekonomicznym, wybraliśmy cztery zmienne statystyczne charakteryzujące sytuację na rynku pracy. Są to: średnia roczna zmiana liczby zarejestrowanych bezrobotnych w latach 2006-2010 (w %), liczba osób dojeżdżających do pracy w danej gminie przypadająca na liczbę osób wyjeżdżających do pracy poza daną gminę w roku 2006, liczba osób w wieku produkcyjnym na 100 osób w wieku nieprodukcyjnym w roku 2006 oraz udział zarejestrowanych bezrobotnych w liczbie osób w wieku produkcyjnym w 2006 r. (w %). Macierz sąsiedztwa społeczno-gospodarczego \tilde{S} wyznaczono z użyciem prognozy (2) (wyniósł on 2,9553). Okazało się, że Kalisz jest w tym ujęciu jednostką izolowaną, ponieważ nie ma żadnego sąsiada. Wszystkie pozostałe gminy są wewnętrzne, tak więc każda gmina jest zagnieżdżona nieskończenie głęboko. Strukturę taką można określić jako *doskonałą*.

Porównania obu macierzy dokonano za pomocą specjalnej procedury napisanej w pakiecie SAS Enterprise Guide 4.2. Indywidualna miara odległości (3) przyjmuje wartości niezerowe dla miasta Kalisza (4,0) oraz gmin: Koźminek, Żelazków (każda po 3,5), Szczytniki (2,0) i Stawiszyn (1,0). Kompleksowa odległość (4) osiągnęła poziom 1,1667. Struktura sąsiedztwa jest wobec tego całkiem odmienna, niż gdyby różnice te rozpatrywać częściowo, tzn. dla każdej jednostki odrębnie według odpowiednich wierszy i kolumn macierzy sąsiedztwa⁵. Największy udział w tej różnicy mają gminy: Kalisz, Koźminek i Żelazków. Rola Kalisza jest oczywista, gmina Koźminek ma aż czterech fizycznych sąsiadów, ale najwyższe bezrobocie, natomiast

⁵ Podobne różnice (w zakresie korelacji) można zaobserwować, analizując tradycyjną postać macierzy korelacji i jej wiersze oraz elementy diagonalne macierzy do niej odwrotnej (zob. np. [Malina, Zeliaś 1998]).

gmina Żelazków przyciąga potencjalnych pracowników dzięki licznym miejscom pracy.

5. Podsumowanie

Przeprowadzone rozważania teoretyczne i przykładowe analizy empiryczne ukazały pewne wspólne cechy i wyraźne odmienności pomiędzy ideami sąsiedztwa fizycznego i społeczno-gospodarczego jednostek (w tym przypadku przestrzennych). Istniejące podobieństwa obu koncepcji dotyczą zasadniczych idei macierzy sąsiedztwa, możliwości wyznaczenia postaci normalnej Smitha oraz istoty jednostek wewnętrznych i brzegowych.

Najistotniejsze różnice w stosowaniu tych podejść związane są z brakiem wypukłości analizowanego obszaru, istnieniem jednostek izolowanych oraz możliwością nieskończonego zagnieżdżenia. Może to mieć ważne znaczenie z punktu widzenia praktycznej interpretowalności wyników, a co za tym idzie – obserwacji określonych zjawisk złożonych oraz oceny wpływu podejmowanych działań na sytuację regionu. Wyznaczanie form normalnych Smitha jest szybkie i efektywne, stanowią one więc dobre narzędzie do porównywania obu struktur sąsiedztwa.

Badania te otwierają pole do dalszych analiz z zakresu modelowania zjawisk społeczno-ekonomicznych, a szczególnie studiów dotyczących własności ekonometrycznych modeli regresji przestrzennej z wykorzystaniem przedstawionych tutaj macierzy sąsiedztwa.

Literatura

- Eberle M.G., Maciel M.C., *Finding the closest Toeplitz matrix*, „Computational and Applied Mathematics” 2003, vol. 22.
- Everitt B.S., Landau S., Leese M., Stahl D., *Cluster Analysis*, 5th Edition, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., Chichester, UK 2011.
- Havas G., Majewski B. S., *Integer matrix diagonalization*, „Journal of Symbolic Computation” 1997, vol. 24.
- Malina A., Zeliaś A., *On building taxonomic measures on living conditions*, “Statistics in Transition” 1998, vol. 3.
- Meyer C.D., *Matrix Analysis and Applied Linear Algebra*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, USA 2000.
- Młodak A., *Polish Experiences and Possibilities in Realisation of the URBAN AUDIT programme*, [w:] J. Dziechciarz (red.), *Globalization Impact on Regional and Urban Statistics*, Proceedings from the 25th SCORUS Conference on Regional and Urban Statistics and Research, Publishing House of the Wrocław University of Economics, Wrocław, tekst dostępny również w Internecie: <http://www.scorus2006.ae.wroc.pl>, 2008.
- Młodak A., *Analiza taksonomiczna w statystyce regionalnej*, Centrum Doradztwa i Informacji DIFIN, Warszawa 2006.
- Młodak A., *Classification of multivariate objects using interval quantile classes*, “Journal of Classification” 2011, vol. 28.

- Palmgren E., *Open sublocales of localic completions*, U.U.D.M. Report 2009:1, Department of Mathematics, Uppsala University, Uppsala, Szwecja, tekst dostępny na stronie <http://www2.math.uu.se/research/pub/Palmgren18.pdf>, 2009.
- Rogalińska D., *Wykorzystanie danych ze źródeł administracyjnych w statystyce miast*, „Wiadomości Statystyczne” 2007, R. LII, nr 2.
- Wagner W., Mantaj A., *Contiguity matrix of spatial units and its properties on example of land districts of Podkarpackie voivodship*, *Statistics in Transition – new series*, 2010, vol. 11.
- Wiperman B.H., *Hierarchical agglomerative cluster analysis with a contiguity constraint*, A project submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby – Surrey – Vancouver, Kanada. Tekst dostępny także na stronie <http://ir.lib.sfu.ca/bitstream/1892/9005/1/b3861439x.pdf>, 2004.

NEIGHBORHOOD OF SPATIAL AREAS IN THE PHYSICAL AND SOCIO-ECONOMIC CONTEXT – A TAXONOMIC APPROACH

Summary: In the paper a comparison of two concepts of neighborhood of spatial areas: physical and socio-economic is presented. The former one concerns their location on administrative map and potential community of borders, the latter one – a similarity in terms of composite social or economical phenomenon. The paper describes the classical theory of neighborhood and its matrix and shows how it can be used in the case of multivariate data analysis. At the end, an effective method of comparison of both types of neighborhood matrices using – known from algebra – the Smith normal form of an integer matrix is proposed. The utility of investigated model was studied upon the example of data on labor market in Kalisz and districts of kaliski second level of local government administration.

Keywords: neighborhood, neighborhood matrix, distance, metrics, Smith normal form.