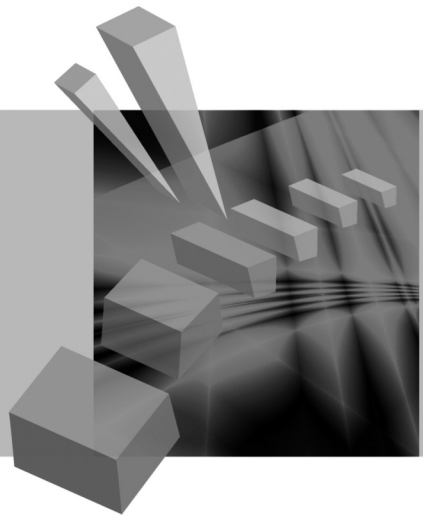
**PRACE NAUKOWE**
Uniwersytetu Ekonomicznego we Wrocławiu
**RESEARCH PAPERS**
of Wrocław University of Economics

**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 Pociecha,
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

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl



**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

## Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

**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 $\operatorname{int} A$ oznacza wnętrze nieskończonego zbioru $A$, zaś $\operatorname{fr} A$ – jego brzeg (krawędź). Jeśli z kolei $A$ jest skończony, to symbolem $|A|$ oznaczać będziemy liczbę jego elementów. Załóżmy, że badany obszar przestrzenny $U$ składa się z $n$ (gdzie $n$ jest liczbą naturalną) jednostek $U_1, U_2, \ldots, U_n$. 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 $\text{fr}\, U_i \setminus \bigcup_{j=1}^{k_i} \left(\text{fr}\, U_{i_j} \cap \text{fr}\, U_i\right) \neq \emptyset$. W przeciwnym razie, tzn. gdy $\text{fr}\, U_i \setminus \bigcup_{j=1}^{k_i} \left(\text{fr}\, U_{i_j} \cap \text{fr}\, U_i\right) = \emptyset$, 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.

$$\delta_{ij} \stackrel{\text{def}}{=} \min_{k=1,2,\dots,n} \left|\Delta_{ijk}\right| - 1, \tag{1}$$

gdzie $\Delta_{ijk} = \{U_{i_1}, U_{i_2}, \dots, U_{i_k}\colon\ U_{i_c} \in U,\ i_c \in \{1, 2, \dots, n\},\ c = 1, 2, \dots, k,.\ U_{i_1} = U_i,\ U_{i_{c-1}} \bowtie U_c, c = 2, \dots, k, U_{i_k} = U_j\}$. Miara zagnieżdżenia minimalizuje tę wielkość: $\delta_i^* \stackrel{\text{def}}{=} \min_{j:U_j \in U_B} \delta_{ij}$. Rozmiar najdłuższej możliwej ścieżki przejścia $\delta_U \stackrel{\text{def}}{=} \max_{i,j:U_i,U_j \subseteq U} \delta_{ij}$ nazywa się średnicą zbioru $U$. Formalnie rzecz ujawszy, 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 diagonali ma zera, liczba elementów równych 1 jest dwukrotnością liczby wszystkich par obiektów sąsiadujących ze sobą (czyli $2s$, przy czym $s = \left|(i,j)\colon i, j = 1,2, \dots,\ n,\ i \neq j, \mathrm{U}_i \bowtie \mathrm{U}_j|\right)$, $\mathbf{S} \cdot \mathbf{1} = \boldsymbol{k}$, gdzie $\boldsymbol{k} = (k_1, k_2, \dots, k_n)$ 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 $\mathrm{tr}(\mathbf{T}) = 2s$ 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 normalizacji[1]. Definicję macierzy sąsiedztwa w tym przypadku oprzemy na zagregowanej odległości obiektów będącej funkcją $d: U \times U \to \mathbb{R}$, 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 $d_{\max} \stackrel{\text{def}}{=} \max_{h,i=1,2,\dots,n} d_{hi}$, $d_{\min} \stackrel{\text{def}}{=} \min_{h,i=1,2,\dots,n, h \neq i} d_{hi}$, $d_{\mathrm{mm}} \stackrel{\text{def}}{=} \min_{h=1,2,\dots,n} \max_{i=1,2,\dots,n} d_{hi}$ oraz

$$\tilde{d} \stackrel{\text{def}}{=} \frac{d_{\max} - d_{\mathrm{mm}}}{d_{\max} - d_{\min}} d_{\max}. \tag{2}$$

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. $d_{\mathrm{mm}}$). 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 zmienności (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 ujednolicenia 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 $U_{i_1}, U_{i_2}, \ldots, U_{i_q}, i_c \in \{1,2,\ldots,n\}, \; c = 1,2,\ldots,q$ tworzą *cykl sąsiedztwa* długości $q$, jeśli $U_{i_c} \bowtie U_{i_{c+1}}$ dla każdego $c = 1,2,\ldots,q$, 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 $A = [a_{lj}], \; l,j = 1,2,\ldots,p,$ gdzie $a_{l(l+1)} = a_{(l+1)l} = a_{p1} = a_{1p} = 1$ dla $l = 1,2,\ldots,p-1$, a pozostałe elementy są zerami. Na przykład dla $p = 5$ wygląda to tak:

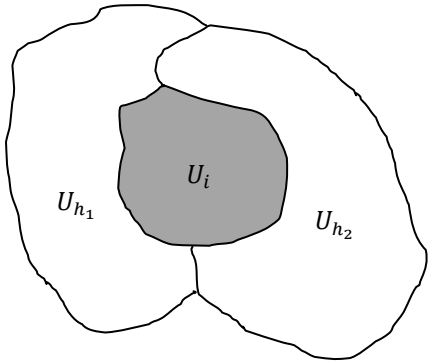
$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Z algebraicznego punktu widzenia jest to szczególny przypadek *macierzy Toeplitza* (zob. np. [Eberle, Maciel 2003]). Można wykazać[2], że gdy $n > 2$, wartości własne takiej macierzy są postaci $\lambda_h = 2\cos(2\,h\pi/n)$, $h = 1,2,\ldots,n$, a zatem jej wyznacznik jest równy $\det(A) = \prod_{h=1}^{n} \lambda_h = 2^n \prod_{h=1}^{n} \cos(2\,h\pi/n)$. 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 jednostce[3]. 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 $\hbar(a,\varepsilon)$ będzie kulą o środku $a$ i promieniu $\varepsilon > 0$ w tejże przestrzeni. Wówczas (zob. np. [Palmgren 2009]) $\hbar(a,\varepsilon) \subseteq \hbar(b,\eta) \Leftrightarrow d(a,b) + \varepsilon \leq \eta$. 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 $d_{k,h_1+h_2} \stackrel{\text{def}}{=} d\left(U_k, U_{h_1} \cup U_{h_2}\right) = \max\left\{d_{kh_1}, d_{kh_2}\right\}$ – aby spełniona była nierówność postaci: $d_{i,h_1+h_2} < \text{med}_{k=1,2,\dots,n,k\neq i,h_1,h_2}\left|d_{ik} - d_{k,h_1+h_2}\right|$, $\quad i,h_1,h_2,k \in \{1,2\dots,n\}, i \neq h_1,h_2,h_1 \neq h_2$.

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 $\tilde{\gamma}_{ii_r} = \left(\left(1/\left(d_{ii_r} + 0{,}01\right)\right)/\left(\sum_{h=1}^{q}\left(1/\left(d_{ii_h} + 0{,}01\right)\right)\right)\right)\gamma_{ii_r}$, 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 możliwości 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 $\mathbf{A}^* = \mathrm{diag}(a_1^*, a_2^*, \dots, a_n^*)$ 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, że[4] $a_i^* \in \mathbb{Z}$ dla każdego $i = 1,2,\dots,n$, $a_1^* > 0, a_2^* > 0, \dots, a_r^* > 0$ dla $r = \mathrm{rz}(\mathbf{A})$, $a_{r+1}^* = a_{r+2}^* = \dots = a_n^* = 0$ 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 przestawiania 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

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

[4] $\mathbb{Z}$ oznacza zbiór liczb całkowitych.

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

$$d = \frac{1}{n}\sum_{i=1}^{n} d_i. \tag{4}$$

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 $\tilde{\mathbf{S}}$ wyznaczono z użyciem progu (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 cząstkowo, tzn. dla każdej jednostki odrębnie według odpowiednich wierszy i kolumn macierzy sąsiedztw[5]. 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 taxonometric 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.

Wipperman 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.