Zeszyty Naukowe Wydziału Informatycznych Technik Zarządzania
Wyższej Szkoły Informatyki Stosowanej i Zarządzania
„Współczesne Problemy Zarządzania”
Nr 1/2012

GRADACYJNA ANALIZA DANYCH – IDEA I PRZYKŁAD ZASTOSOWANIA

Stanisław Lenkiewicz
Studium Doktoranckie przy IBS PAN
Newelska 6, 01-447 Warszawa
stan@lenkiewicz.eu

Artykuł prezentuje gradacyjną analizę danych – oryginalną metodę eksploracji danych opracowaną i rozwijaną w Instytucie Podstaw Informatyki PAN – oraz jej podstawowe narzędzie: gradacyjną analizę odpowiedniości. Choć oparta na solidnych podstawach teoretycznych, metoda gradacyjna jest na tyle uniwersalna, że mogą ją stosować badacze zajmujący się różnymi dziedzinami, nawet bardzo odległymi od statystyki i informatyki.

Tekst składa się z trzech części. Pierwsza stanowi omówienie (z racji ograniczonej objętości – dość ogólne) idei analizy gradacyjnej. W drugiej przedstawiono algorytm GCA. Trzecia to przykład wykorzystania GCA: na podstawie danych statystycznych dostępnych w portalu Eurostatu dokonano podziału 27 krajów Unii Europejskiej na kilka grup. Uzyskane wyniki mogą zaskakiwać.

Słowa kluczowe: eksploracja danych, analiza skupień, gradacyjna analiza odpowiedniości, krzywa koncentracji, nadreprezentacja, element odstający.

1. Wprowadzenie

W życiu często mamy do czynienia z dużymi zestawami danych. Zawierają one wartościowe informacje, które jednak nie są łatwo dostępne. Na przykład wykaz kursów akcji spółek notowanych na giełdzie z okresu ostatniego roku jest bardzo cenny dla inwestora, jednak na jego podstawie nie dokona on decyzji zakupu bądź sprzedaży. Akcje których spółek mają najbardziej stabilne notowania, jakie jest ryzyko związane z inwestowaniem w akcje firm należących do poszczególnych sektorów, jaki jest oczekiwany zysk z akcji danej spółki i ile wynosi prawdopodobieństwo jego osiągnięcia – nie sposób odpowiedzieć na te pytania na podstawie surowych danych.

Uzyskanie informacji na podstawie danych wymaga ich odpowiedniego przetworzenia. Proces ten jest określany mianem eksploracji danych (ang. *Data Mining*). Wykorzystuje on szereg narzędzi, m.in. metody statystyczne, metody sztucznej inteligencji (sieci neuronowe, algorytmy ewolucyjne), logikę rozmytą, analizę skupień. To ostatnie stanowi istotny punkt naszych rozważań.

Analiza skupień ma na celu podział zbioru badanych obiektów (np. osób, przedmiotów, zjawisk) na rozdzielne grupy (klasy, skupienia). Obiekty należące do

wyróżnionych grup powinny być, według przyjętych kryteriów pomiaru, podobne do siebie i jednocześnie odmienne od obiektów należących do pozostałych grup. Otrzymany podział daje dwojakie korzyści; z jednej strony umożliwia uzyskanie informacji na temat struktury danych, z drugiej zaś stanowi wskazówkę co do kierunku dalszych badań, pozwalając na skoncentrowanie uwagi na wyłonionych grupach, a tym samym: ograniczenie obszaru poszukiwań. Badania pogrupowanych danych dotyczą wewnętrznej struktury wyłonionych grup (czyli relacji zachodzących pomiędzy ich elementami), a także podobieństw i różnic między grupami.

Wynik badania (czyli stwierdzenie, które obiekty są sobie bliskie, które zaś nie, w sensie przynależności do skupień) w znacznym stopniu zależy od przyjętej metody pomiaru podobieństwa. Jej wybór nie jest zadaniem łatwym; cechy analizowanych obiektów mogą mieć różny charakter (cechy nominalne, porządkowe, interwałowe, ilorazowe), mogą być wyrażone w różnych jednostkach, a odmienność skal ich pomiaru może implikować konieczność przeprowadzenia ich wstępnej normalizacji.

Szczegółowe omówienie idei analizy skupień oraz wykorzystywanych przez nią narzędzi prezentują, m.in., Aldenderfer i Blashfield (1984) oraz Romesburg (2004). Przykłady zastosowań w medycynie podają Jarochowska i in. (2005), Myatt i Johnson (2009), Szromek i Krajewska-Siuda (2008, 2009). Zastosowania w psychologii opisują, m.in., Majchrzyk (2001) i Noworol (1989), w ekonomii i socjologii – Grabowska i Wiech (2009), Johann (2005), Myatt i Johnson (2009), Siedlecka (2006), Wyka (2009). Stosunkowo skromnie na tle innych dziedzin wygląda wykorzystanie analizy skupień w przemyśle. Jest to wciąż obszar do nowych zastosowań. Problem ten szerzej omówił Wyrozumski (2002).

Jak wspomnieliśmy na wstępie, eksploracja danych wykorzystuje wiele różnorodnych metod. W tym artykule skoncentrujemy się na nieco mniej znanej, ale bardzo użytecznej gradacyjnej analizie danych (ang. *Grade Data Analysis, GDA*) – metodzie opracowanej i rozwijanej w Instytucie Podstaw Informatyki PAN. Podstawowym narzędziem tej metody jest algorytm gradacyjnej analizy odpowiedniości (ang. *Grade Correspondence Analysis, GCA*), którego praktyczną realizację stanowi program GradeStat. Pozwala on na wszechstronną analizę danych, w tym – przeprowadzenie procedury analizy skupień i wykrycie elementów odstających.

W niniejszym artykule przedstawimy zarys GCA, ilustrując opis praktycznymi przykładami. Kompletny wykład problematyki zawiera praca Kowalczyk i in. (2004), a interesujące przykłady zastosowań – prace Ciok (2004), Pleszczyńska i in. (2006), Szczęsny i Jarochowska (2006), Wiech (2007).

2. Gradacyjna analiza danych (GDA)

Jak wspomnieliśmy we wprowadzeniu, celem eksploracji danych jest uzyskanie z danych informacji. Dane są zbiorami obiektów posiadających określone cechy. Na przykład w dziennym podsumowaniu notowań giełdowych obiektami są poszczególne spółki, a ich cechami: nazwa spółki, kurs otwarcia i kurs zamknięcia, kurs minimalny i maksymalny, procentowa zmiana kursu, wolumen obrotów. Dane zazwyczaj są przedstawiane w postaci tabelarycznej, przy czym obiektom odpowia-

dają wiersze, a cechom kolumny – tak też wyglądają podsumowania giełdowe, jakie można znaleźć w prasie czy w serwisach internetowych. Aby uzyskać użyteczną informację, należy odpowiednio przetworzyć dane. Ktoś, kogo interesuje informacja, które spółki odnotowały największe wzrosty, a które – największe spadki, uporządkuje tabelę notowań według kolumny procentowej zmiany kursu. Jednak, aby stwierdzić, spółki którego sektora dają szansę największych zysków, nie wystarczy kilka prostych przekształceń tabeli. Ogólnie: im bardziej złożoną informację chcemy uzyskać, tym mocniejszych narzędzi musimy użyć podczas przetwarzania danych.

Punktem wyjścia gradacyjnej analizy danych jest porównanie pomiędzy sobą odpowiednich cech poszczególnych obiektów. Miernikami podobieństwa / odmienności, jakie wykorzystuje GDA, są wskaźniki koncentracji, których ilustrację stanowią krzywe koncentracji. Oba pojęcia zostaną omówione w dalszej części.

2.1. Krzywa koncentracji

Krzywa koncentracji jest narzędziem służącym do graficznego porównywania rozkładów dwu jednowymiarowych zmiennych losowych.

Rozważmy następujący przykład. W Tabeli 1 zamieszczono dane o liczbie miast województw: mazowieckiego i dolnośląskiego z podziałem według liczby mieszkańców. Dane pochodzą ze strony Głównego Urzędu Statystycznego.

Tabela 1. Liczba miast województw: mazowieckiego i dolnośląskiego z podziałem wg liczby mieszkańców

Województwo	Liczba mieszkańców								Ogółem
	Poniżej 2 000	2 000 – 4 999	5 000 – 9 999	10 000 – 19 999	20 000 – 49 999	50 000 – 99 999	100 000 – 199 999	200 000 i więcej	
Mazowieckie	4	21	14	22	17	4	1	2	85
Dolnośląskie	2	23	29	17	13	4	2	1	91

Źródło: GUS.

Wartości w poszczególnych wierszach dzielimy przez sumy wierszy. Uzyskujemy tabelę, której wiersze opisują rozkłady prawdopodobieństwa dwu zmiennych losowych. Wartości w poszczególnych komórkach odpowiadają prawdopodobieństwom tego, że wybrane losowo miasto leżące w danym województwie ma daną liczbę mieszkańców (a dokładniej: ma liczbę mieszkańców z danego zakresu).

Tabela 2. Tabela po podzieleniu wyrazów przez sumy wierszy

Województwo	Liczba mieszkańców								Ogółem
	Poniżej 2 000	2 000 – 4 999	5 000 – 9 999	10 000 – 19 999	20 000 – 49 999	50 000 – 99 999	100 000 – 199 999	200 000 i więcej	
Mazowieckie	0,05	0,25	0,16	0,26	0,20	0,05	0,01	0,02	1,00
Dolnośląskie	0,02	0,25	0,32	0,19	0,14	0,04	0,02	0,01	1,00

Źródło: opracowanie własne na podstawie danych GUS.

Na podstawie otrzymanych rozkładów prawdopodobieństwa obliczamy dystrybuanty dla poszczególnych wartości obu zmiennych losowych.

Tabela 3. Dystrybuanty „zmiennych losowych wierszowych"

Województwo	Liczba mieszkańców							
	Poniżej 2 000	2 000 – 4 999	5 000 – 9 999	10 000 – 19 999	20 000 – 49 999	50 000 – 99 999	100 000 – 199 999	200 000 i więcej
Mazowieckie	0,05	0,29	0,46	0,72	0,92	0,96	0,98	1,00
Dolnośląskie	0,02	0,27	0,59	0,78	0,92	0,97	0,99	1,00

Źródło: opracowanie własne na podstawie danych GUS.

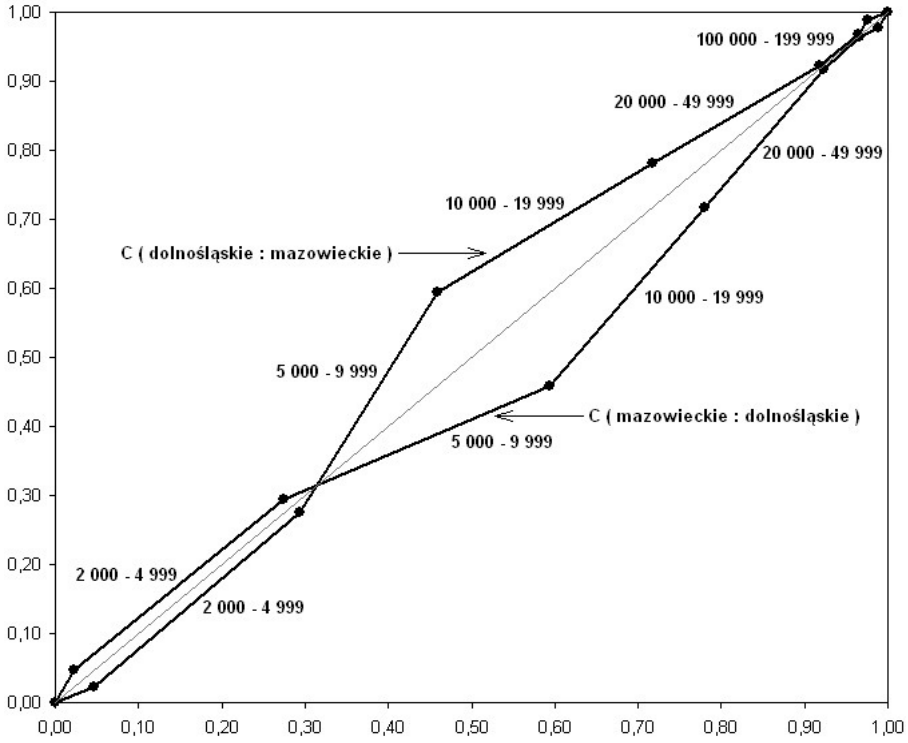
Zaznaczamy w układzie współrzędnych punkty: (0,00, 0,00), (0,05, 0,02), (0,29, 0,27), (0,46, 0,59), (0,72, 0,78), (0,92, 0,92), (0,96, 0.97), (0,98, 0,99), (1,00, 1,00) i łączymy je odcinkami. Otrzymujemy w ten sposób łamaną, którą nazywamy **krzywą koncentracji** wielkości miast województwa dolnośląskiego względem wielkości miast województwa mazowieckiego. Analogicznie można wykreślić krzywą dla zależności odwrotnej (koncentracji wielkości miast województwa mazowieckiego względem wielkości miast województwa dolnośląskiego). Obie krzywe przedstawiono na Rys. 1. Jak łatwo zauważyć, jedna stanowi odbicie drugiej względem przekątnej układu współrzędnych.

Krzywa koncentracji jest łamaną złożoną z tylu odcinków, ile cech mają badane obiekty. Każdy odcinek stanowi ilustrację porównania „koncentracji" danej cechy w obu obiektach; jeśli jest nachylony do osi odciętych pod kątem mniejszym niż 45^o, „koncentracja" cechy w pierwszym obiekcie jest mniejsza niż w drugim; kąt większy niż 45^o oznacza zależność odwrotną, a kąt równy 45^o – identyczną „koncentrację" cechy w obu obiektach. W analizowanych przykładzie krzywa koncentracji ilustruje porównanie „nasycenia" badanych województw miastami o określonej wielkości.

Krzywa koncentracji na niektórych odcinkach biegnie poniżej, na innych zaś powyżej przekątnej układu. Jeśli jednak uporządkujemy odcinki, z których się składa, rosnąco według kąta nachylenia do osi poziomej, to uzyskana z ich połączenia krzywa zawsze będzie położona poniżej przekątnej (i będzie krzywą wypukłą). Tę krzywą nazywamy **krzywą maksymalnej koncentracji**. Na Rys. 2 przedstawiono krzywą maksymalnej koncentracji wielkości miast województwa dolnośląskiego względem wielkości miast województwa mazowieckiego.

Krzywa maksymalnej koncentracji jest tą, która leży najniżej. Każda krzywa koncentracji zawiera się w obszarze ograniczonym od dołu krzywą maksymalnej koncentracji (patrz Rys. 3), a od góry – jej odbiciem względem przekątnej układu współrzędnych.

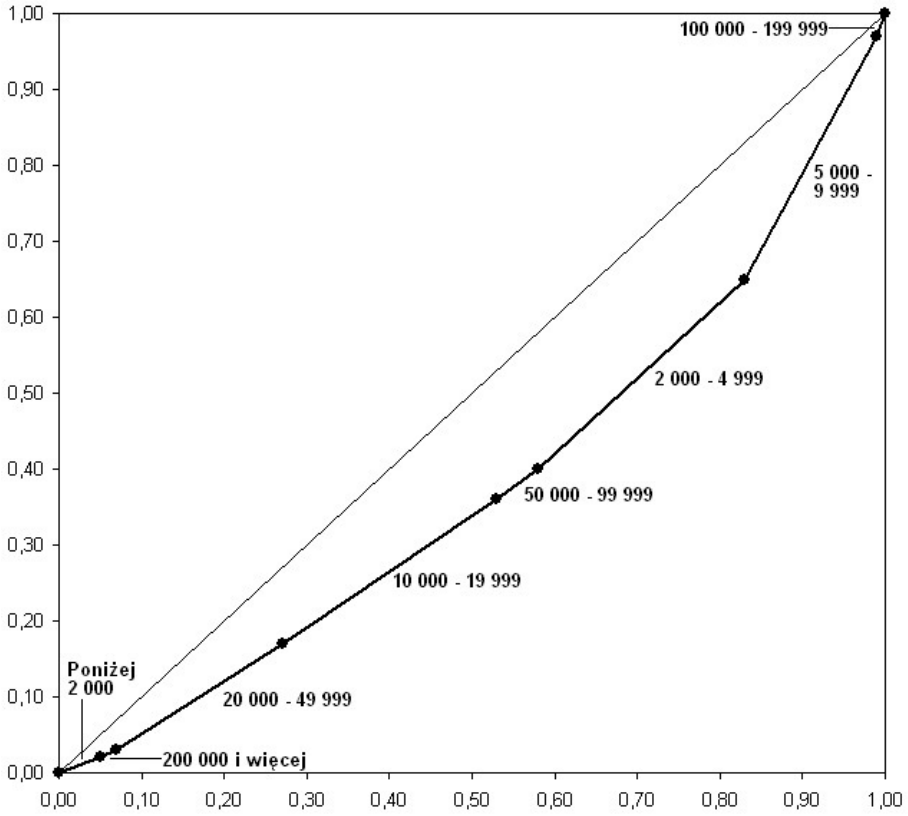
Rysunek 1. Krzywe koncentracji wielkości miast województw: dolnośląskiego i mazowieckiego



Źródło: opracowanie własne na podstawie danych GUS.

Poprzez wykreślanie krzywych maksymalnej koncentracji można porównywać wiele par rozkładów. Można byłoby np. porównać parami liczby miast o określonej wielkości we wszystkich województwach Polski. Zróżnicowanie rozkładów jest większe w tej parze, dla której krzywa maksymalnej koncentracji leży niżej.

Rysunek 2. Krzywa maksymalnej koncentracji wielkości miast województwa dolnośląskiego względem wielkości miast województwa mazowieckiego.



Źródło: opracowanie własne na podstawie danych GUS.

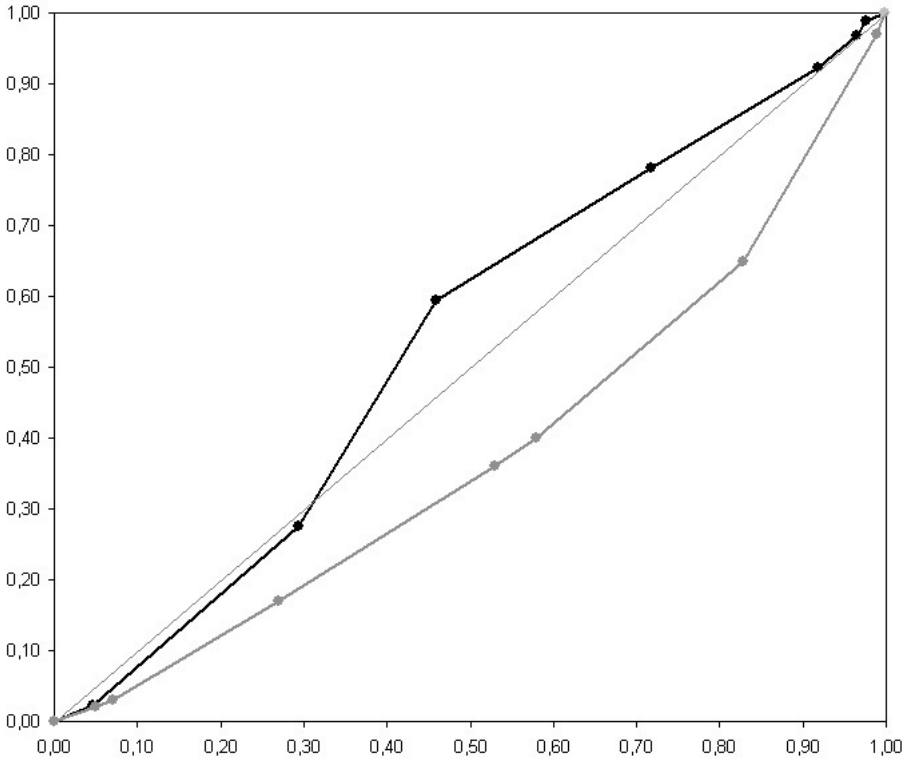
2.2. Krzywa Lorenza

Krzywa maksymalnej koncentracji jest narzędziem specyficznym dla GDA, jednak pomysł graficznego porównywania rozkładów dwu zmiennych losowych nie jest nowy. Krzywa koncentracji stanowi adaptację stosowanej w statystyce i ekonometrii krzywej Lorenza.

Rozważmy rozkład dyskretny złożony z n elementów posiadających interesującą nas cechę y. Ustawmy elementy rozkładu niemalejąco ze względu na cechę y, tzn. tak, że dla każdego $i = 1, \ldots, n$ zachodzi zależność $y_i \leq y_{i+1}$. Krzywą Lorenza będzie łamana złożona z odcinków o końcach w punktach (F_i, L_i), gdzie $F_0=0$, $L_0=0$,

$$F_i = \frac{i}{n},\ S_i = \sum_{j=1}^{i} y_j,\ L_i = \frac{S_i}{S_n}.$$

Rysunek 3. Jedna z krzywych koncentracji (u góry),
krzywa maksymalnej koncentracji (u dołu) i przekątna układu współrzędnych



Źródło: opracowanie własne na podstawie danych GUS.

Krzywa Lorenza jest wykorzystywana przede wszystkim do pomiaru nierównomierności zarobków w badanym kraju. Skonstruujemy ją dla zarobków Polaków. W Tabeli 4 przedstawiono wyniki Ogólnopolskiego Badania Wynagrodzeń przeprowadzonego przez firmę Sedlak & Sedlak w 2011 roku.

Dla skonstruowania krzywej Lorenza potrzebujemy dokładnych danych na temat zarobków każdej badanej osoby. W Tabeli 4 znajdują się przedziały, w jakich mieszczą się zarobki respondentów, oraz odsetki uczestników badania, którzy zadeklarowali dochody z danego przedziału. Przed wykreśleniem krzywej musimy więc dokonać przekształcenia danych.

W badaniu wzięły udział 92 262 osoby, więc zarobki poniżej 2 000 PLN zadeklarowało 92 262 × 14,30% = 13 193,466 osób. Ten wynik zaokrąglamy do wartości całkowitej 13 193. Zarobki z zakresu 2 001 – 3 000 zadeklarowało 92 262 × 22,50% = 20 758,95 ≈ 20 759 osób itd.

Tabela 4. Wyniki Ogólnopolskiego Badania Wynagrodzeń w roku 2011

Wysokość miesięcznych zarobków w PLN	Odsetek uczestników badania
Poniżej 2 000	14,30%
2 001 – 3 000	22,50%
3 001 – 4 000	18,10%
4 001 – 5 000	12,80%
5 001 – 6 000	8,30%
6 001 – 7 000	5,70%
7 001 – 8 000	4,20%
8 001 – 9 000	2,90%
9 001 – 10 000	2,40%
10 001 – 11 000	1,50%
11 001 – 12 000	1,30%
12 001 – 13 000	0,90%
13 001 – 14 000	0,70%
14 001 – 15 000	0,70%
15 001 – 16 000	0,50%
16 001 – 17 000	0,40%
17 001 – 18 000	0,40%
Powyżej 18 000	2,40%

Źródło: portal pracuj.pl (dostęp 4 lipca 2012).

Nie znamy dokładnej wielkości zarobków poszczególnych osób, tylko przedziały, z jakich one pochodzą. Przyjmiemy więc, że wszystkie osoby z pierwszego przedziału zarabiają 1 500 PLN, z drugiego – 2 500 PLN itd. W przypadku osób z ostatniego (otwartego z prawej strony) przedziału uznamy, że ich zarobki wynoszą 18 500 PLN.

Tabela 5 zawiera część tabeli danych po dokonaniu opisanych modyfikacji. Oczywiście skutkiem dokonanych przeliczeń zawarte w niej wartości są przybliżone, jednak na tyle dokładne, że powstała na ich podstawie krzywa Lorenza wiernie oddaje strukturę zarobków respondentów. Przedstawiono ją na Rys. 4.

Krzywa Lorenza jest krzywą wypukłą, którą przedstawia się w kwadracie jednostkowym. Stanowi ona ilustrację nierównomierności rozkładu dochodów w społeczeństwie. Z Rys. 4 można odczytać, że 50% mieszkańców naszego kraju dysponuje zaledwie 25% ogółu dochodów, a 76% mieszkańców – zaledwie połową. Połowa ogólnych dochodów jest skupiona w rękach zaledwie 24% Polaków.

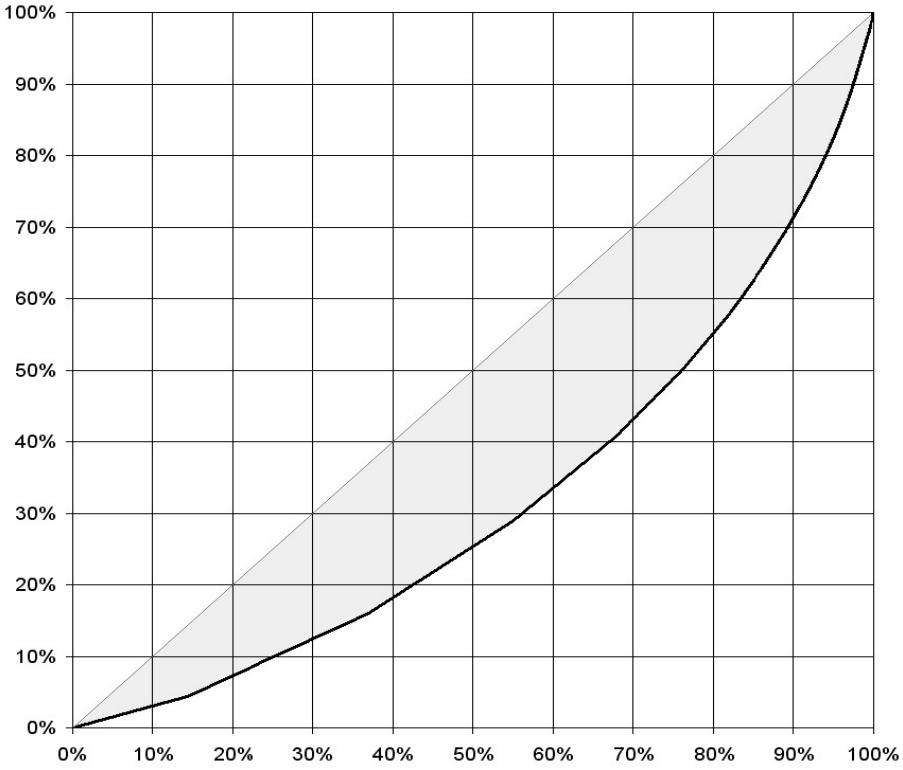
Jeśli krzywa Lorenza przechodzi w prostą stanowiącą przekątną układu, świadczy to o idealnie proporcjonalnym podziale dochodów. W praktyce taka sytuacja nigdy nie zachodzi. Wielkość pola figury ograniczonej od dołu krzywą Lorenza, a od góry przekątną kwadratu stanowi miernik nierównomierności rozkładu dochodów w społeczeństwie i nosi nazwę wskaźnika Giniego.

Tabela 5. Wyniki Ogólnopolskiego Badania Wynagrodzeń w roku 2011 w postaci umożliwiającej wykreślenie krzywej Lorenza (fragment)

i	y_i	F_i	S_i	L_i
1	1 500	0,00%	1 500	0,00%
13 193	1 500	14,30%	19 789 500	4,40%
13 194	2 500	14,30%	19 792 000	4,40%
33 952	2 500	36,80%	71 687 000	15,94%
33 953	3 500	36,80%	71 690 500	15,94%
50 652	3 500	54,90%	130 137 000	28,93%
50 653	4 500	54,90%	130 141 500	28,93%
62 461	4 500	67,70%	183 277 500	40,74%
62 462	5 500	67,70%	183 283 000	40,74%
70 119	5 500	76,00%	225 396 500	50,10%
70 120	6 500	76,00%	225 403 000	50,10%
75 378	6 500	81,70%	259 580 000	57,70%
75 379	7 500	81,70%	259 587 500	57,70%
79 253	7 500	85,90%	288 642 500	64,16%
79 254	8 500	85,90%	288 651 000	64,16%
81 929	8 500	88,80%	311 388 500	69,22%
81 930	9 500	88,80%	311 398 000	69,22%
84 143	9 500	91,20%	332 421 500	73,89%
84 144	10 500	91,20%	332 432 000	73,90%
85 527	10 500	92,70%	346 953 500	77,12%
85 528	11 500	92,70%	346 965 000	77,13%
86 726	11 500	94,00%	360 742 000	80,19%
86 727	12 500	94,00%	360 754 500	80,19%
87 557	12 500	94,90%	371 129 500	82,50%
87 558	13 500	94,90%	371 143 000	82,50%
88 202	13 500	95,60%	379 837 000	84,43%
88 203	14 500	95,60%	379 851 500	84,44%
88 848	14 500	96,30%	389 204 000	86,51%
88 849	15 500	96,30%	389 219 500	86,52%
89 310	15 500	96,80%	396 365 000	88,11%
89 311	16 500	96,80%	396 381 500	88,11%
89 679	16 500	97,20%	402 453 500	89,46%
89 680	17 500	97,20%	402 471 000	89,46%
90 048	17 500	97,60%	408 911 000	90,90%
90 049	18 500	97,60%	408 929 500	90,90%
92 262	18 500	100,00%	449 870 000	100,00%

Źródło: opracowanie własne na podstawie danych z portalu pracuj.pl.

Rysunek 4. Krzywa Lorenza dla zarobków Polaków w roku 2011.



Źródło: opracowanie własne na podstawie danych z portalu pracuj.pl.

2.3. Wskaźnik Giniego

Wskaźnik (współczynnik, indeks) Giniego stanowi miernik nierównomierności rozkładu zmiennej losowej. W przypadku rozkładu dyskretnego, którego wyrazy uporządkowano rosnąco (a taki właśnie rozważamy), wskaźnik ten wyraża się wzorem:

$$G(y) = \frac{\sum_{i=1}^{n} (2i - n - 1) y_i}{n^2 \bar{y}}, \tag{1}$$

gdzie y_i to wartość i-tego wyrazu rozkładu (w naszym przypadku wartość dochodu i-tej osoby objętej badaniem), a $\bar{y}$ to średnia wszystkich wyrazów (tutaj: średnia dochodów całej badanej grupy).

Wskaźnik Giniego dla analizowanych danych wynosi 36,62%. Nie należy go jednak interpretować jako wskaźnika Giniego dla Polski w 2011 roku, bowiem dane, jakimi dysponujemy, nie są reprezentatywne. W badaniu uczestniczyły głównie

osoby młode (68,1% respondentów ma nie więcej niż 35 lat) i dobrze wykształcone (77,2% badanych ma wyższe wykształcenie). Tymczasem w skali kraju osoby w wieku do 35 lat stanowią ok. 38% ludności w wieku produkcyjnym, a wśród wszystkich aktywnych zawodowo Polaków wyższe wykształcenie ma ok. 29%.

Na stronach Eurostatu można znaleźć wskaźnik Giniego dla Polski w roku 2010; wynosi on 31,1%. W roku 2011 mógł się nieznacznie zmienić, jednak nie aż o 5,52 punktu procentowego.

Szczegółowe omówienie krzywej Lorenza oraz wskaźnika Giniego (wraz z dowodami wzajemnych powiązań) podaje Biernacki (2006).

2.4. Porównywanie „rozkładów kolumnowych"

Porównywanie rozkładów zmiennych stanowi narzędzie pomiaru ich podobieństwa / odmienności. Do tej pory porównywaliśmy „rozkłady wierszowe" (czyli obiekty); można jednak w podobny sposób porównywać „rozkłady kolumnowe" (czyli cechy obiektów – ilustruje to Rys. 5).

Krzywe koncentracji rozpatrywane w punkcie 2.1 stanowiły ilustrację porównania „nasycenia" **dwu województw** miastami o poszczególnych liczbach mieszkańców. Pozwalały odpowiedzieć na pytanie, w którym z nich jest więcej miast małych, średnich, dużych itd. Krzywa pokazana na Rys. 5 ilustruje porównanie „nasycenia" miastami liczącymi 100 000 – 199 999 mieszkańców względem „nasycenia" miastami liczących 5 000 – 9 999 **we wszystkich województwach**. Pozwala stwierdzić, w którym z nich duże miasta przeważają nad małymi, w którym zaś jest odwrotnie.

2.5. Wskaźniki koncentracji

Krzywe koncentracji można traktować jak wykresy dystrybuanty $F(X)$ pewnej ciągłej zmiennej losowej X określonej na przedziale [0; 1]. Średnia rozkładu zmiennej X mówi o zróżnicowaniu rozkładów, których porównanie ilustruje krzywa. Średnia ta wynosi

$$E(X)=\int_0^1 xf(x)dx\,, \tag{2}$$

gdzie $f(x)$ to funkcja gęstości. Przekształcenie wzoru (2)

$$E(X)=\int_0^1 x\frac{dF(x)}{dx}dx=\int_0^1 xdF(x)=1-\int_0^1 F(x)dx \tag{3}$$

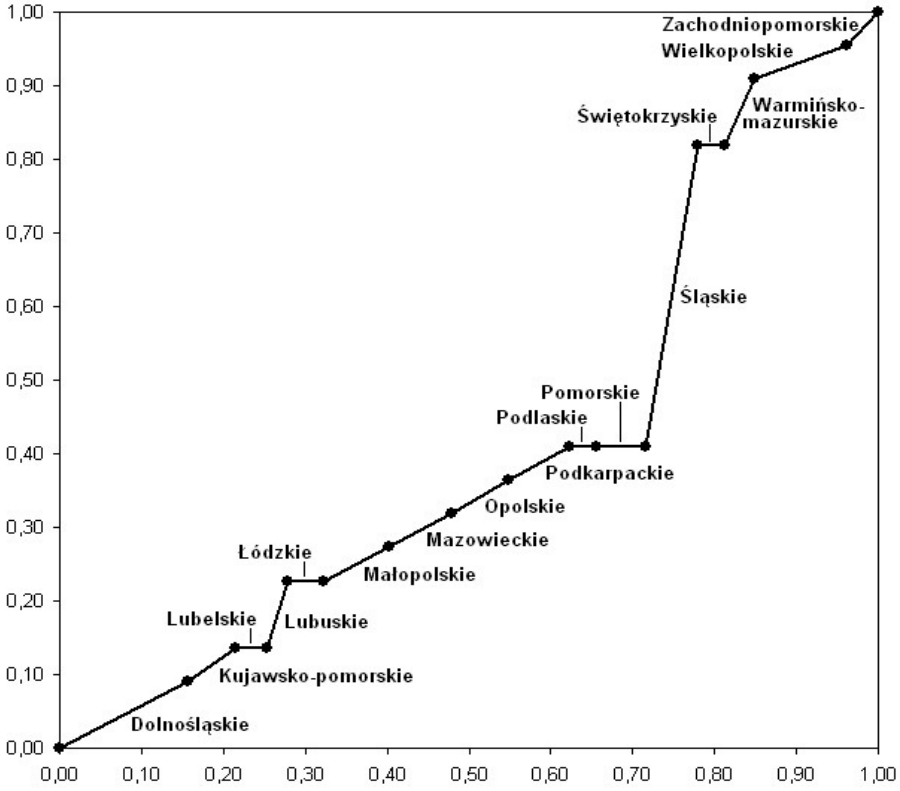
pokazuje, że **średnia jest równa polu nad krzywą koncentracji**.

Średnia przyjmuje wartości z zakresu 0 – 1. W przypadku gdy porównywane rozkłady są identyczne, krzywa koncentracji pokrywa się z przekątną układu i wówczas średnia wynosi ½. Wygodnym miernikiem zróżnicowania rozkładów byłby

taki, który dla rozkładów identycznych przyjmowałby wartość 0, a dla skrajnie różnych – wartość 1 lub -1. Uzyskujemy to przez przekształcenie średniej:

$$ar = 2\left(E(X) - \frac{1}{2}\right). \tag{4}$$

Rysunek 5. Krzywa koncentracji miast liczących 100 000 – 199 999 mieszkańców względem miast liczących 5 000 – 9 999 mieszkańców.



Źródło: opracowanie własne na podstawie danych GUS.

Otrzymujemy w ten sposób wzór na **wskaźnik koncentracji** (jednego rozkładu względem drugiego). Jak nietrudno zauważyć, wskaźnik ten przyjmuje największą wartość w przypadku krzywej maksymalnej koncentracji, bowiem wówczas pole nad krzywą (równe $E(X)$) jest największe.

Przyjrzyjmy się Rys. 6. Przedstawia on znaną nam z Rys. 1 krzywą koncentracji wielkości miast województwa dolnośląskiego względem wielkości miast województwa mazowieckiego. Zaznaczono na nim przekątną układu i zacieniowano dwa pola: A – ograniczone od dołu krzywą koncentracji, a od góry przekątną układu oraz B – ograniczone od dołu przekątną układu, a od góry krzywą koncentracji.

Ponieważ średnia $E(X)$ jest równa polu nad krzywą koncentracji, a pole nad przekątną układu wynosi ½, to

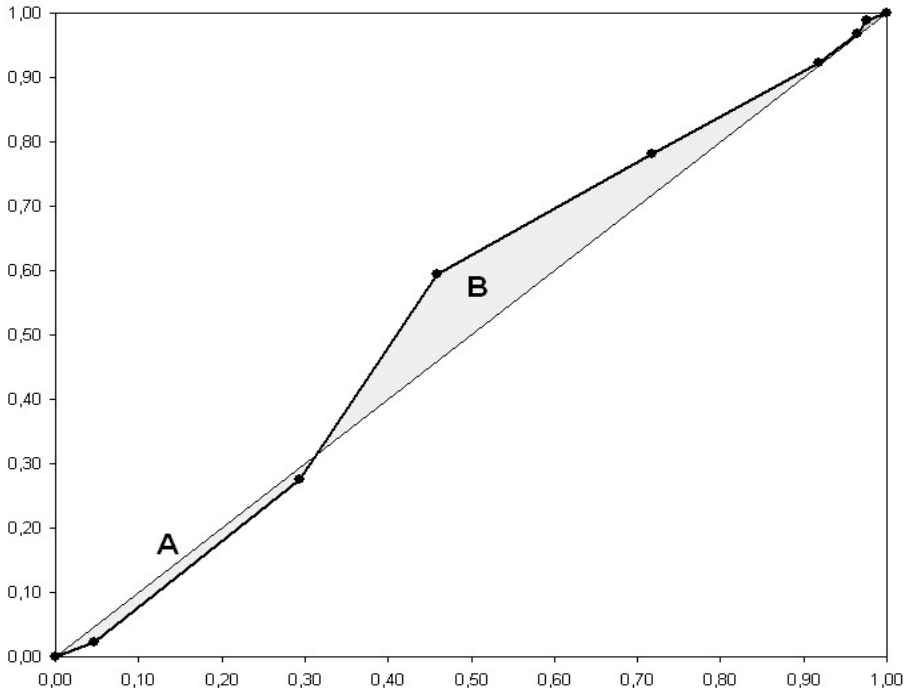
$$ar = 2\left(E(X) - \frac{1}{2}\right) = 2\left(A + \frac{1}{2} - B - \frac{1}{2}\right) = 2(A - B). \tag{5}$$

Jak widać, wskaźnik koncentracji jest równy różnicy dwóch pól ograniczonych krzywą koncentracji i przekątną układu: położonego poniżej i powyżej przekątnej. Stąd też oznaczenie wskaźnika: *ar* od angielskiego *area*.

Ponieważ krzywa maksymalnej koncentracji zawsze leży poniżej przekątnej układu (w wyjątkowym przypadku: pokrywa się z nią), to wskaźnik maksymalnej koncentracji $ar_{\max} = 2A \geq 0$.

Jak można zauważyć, wskaźnik maksymalnej koncentracji jest analogią do omówionego w p. 2.3 wskaźnika Giniego.

Rysunek 6. Krzywa koncentracji wielkości miast województwa dolnośląskiego względem mazowieckiego



Źródło: opracowanie własne na podstawie danych GUS.

2.6. Wskaźnik ρ^*. Zasada działania algorytmu GCA

Dla każdej pary rozkładów (pary wierszy lub pary kolumn tabeli) można znaleźć takie ustawienie ich elementów, by wskaźnik koncentracji jednego względem drugiego osiągnął wartość maksymalną ($ar = ar_{\max}$). Dokonanie tego dla całej tabeli

pozwoliłoby na zbadanie struktury całej zbiorowości, tzn. porównanie rozkładu wszystkich cech we wszystkich obiektach. Zazwyczaj jest to jednak niewykonalne – maksymalizacja *ar* dla jednej pary zmiennych (poprzez zmianę uporządkowania wierszy bądź kolumn tabeli) powoduje zmniejszenie wartości *ar* dla innych par zmiennych. Toteż algorytm GCA dąży do takiego ustawienia elementów tabeli, by osiągnięte wskaźniki *ar* dla wszystkich zmiennych były możliwie najbliższe (choć niekoniecznie równe) $ar_{\max}$. W tym celu GCA w każdym kroku zmienia uporządkowanie wierszy i kolumn, dążąc do maksymalizacji współczynnika korelacji rang Spearmana ρ^*:

$$\rho^* = 3\sum_{i=1}^{m}\sum_{s=1}^{k}\left(p_{is}\left(2S_{row}(i)-1\right)\left(2S_{col}(s)-1\right)\right), \tag{6}$$

gdzie

$$S_{row}(i) = \left(\sum_{j=1}^{i-1} p_{j+}\right) + \frac{1}{2}p_{i+}, \tag{7}$$

$$S_{col}(s) = \left(\sum_{t=1}^{s-1} p_{+t}\right) + \frac{1}{2}p_{+s}, \tag{8}$$

$$p_{j+} = \sum_{s=1}^{k} p_{js} \quad \text{– suma } j\text{-tego wiersza}, \tag{9}$$

$$p_{+t} = \sum_{t=1}^{m} p_{ts} \quad \text{– suma } t\text{-tej kolumny}, \tag{10}$$

p_{ij} to wyraz w i-tym wierszu i j-tej kolumnie, m – liczba wierszy tabeli, k – liczba jej kolumn.

GCA dokonuje permutacji wierszy i kolumn tabeli zgodnie z tzw. regresją gradacyjną, którą opisują wzory:

$$r_{col}(s) = \frac{\sum_{i=1}^{m}\left(p_{is}S_{row}(i)\right)}{p_{+s}} \quad \text{dla kolumn} \tag{11}$$

oraz

$$r_{row}(i) = \frac{\sum_{s=1}^{k}\left(p_{is}S_{col}(s)\right)}{p_{i+}} \quad \text{dla wierszy.} \tag{12}$$

Jak dowiedli Ciok i in. (1995), w każdym kroku algorytmu współczynnik ρ^* wzrasta. Ponieważ liczba możliwych uporządkowań wierszy i kolumn jest skończona i wynosi $m! \times k!$, algorytm musi się zakończyć.

2.7. Wskaźniki nadreprezentacji

W Tabeli 6 przedstawiono przykład tzw. rozkładu proporcjonalnego. Charakteryzuje się on tym, że dla każdego wyrazu p_{ij} zachodzi zależność:

$$p_{ij} = p_{i+} \times p_{+j}\,, \tag{13}$$

gdzie p_{i+} i p_{+j} to, odpowiednio, suma i-tego wiersza i j-tej kolumny.

Tabela 6. Rozkład proporcjonalny.

	$j = 1$	$j = 2$	$j = 3$	$j = 4$	Suma p_{i+}
$i = 1$	0,12	0,10	0,14	0,04	0,40
$i = 2$	0,06	0,05	0,07	0,02	0,20
$i = 3$	0,12	0,10	0,14	0,04	0,40
Suma p_{+j}	0,30	0,25	0,35	0,10	1,00

Źródło: opracowanie własne.

Przez **wskaźnik nadreprezentacji** rozumiemy iloraz

$$c_{ij} = \frac{p_{ij}}{p_{i+} \times p_{+j}}\,. \tag{14}$$

Ponieważ dla rozkładu proporcjonalnego zachodzi zależność (13), wskaźniki nadreprezentacji dla wszystkich komórek Tabeli 6 są równe 1.

W Tabeli 7 pokazano przykład rozkładu nieproporcjonalnego, a w Tabeli 8 – wskaźniki nadreprezentacji dla tego rozkładu.

Tabela 7. Rozkład nieproporcjonalny

	$j = 1$	$j = 2$	$j = 3$	$j = 4$	Suma p_{i+}
$i = 1$	0,25	0,05	0,04	0,06	0,40
$i = 2$	0,10	0,20	0,03	0,02	0,35
$i = 3$	0,10	0,05	0,03	0,07	0,25
Suma p_{+j}	0,45	0,30	0,10	0,15	1,00

Źródło: opracowanie własne.

Tabela 8. Wskaźniki nadreprezentacji dla rozkładu z Tabeli 7

	$j = 1$	$j = 2$	$j = 3$	$j = 4$
$i = 1$	1,39	0,42	1,00	1,00
$i = 2$	0,63	1,90	0,86	0,38
$i = 3$	0,89	0,67	1,20	1,87

Źródło: opracowanie własne.

Uwaga: wartości w Tabeli 8 zaokrąglono; wskaźnik dla $i = 3$, $j = 2$ wynosi dokładnie 2/3.

Wskaźnik nadreprezentacji informuje, na ile wartość zaobserwowana odbiega od tej, jakiej należałoby oczekiwać przy idealnej proporcjonalności rozkładu.

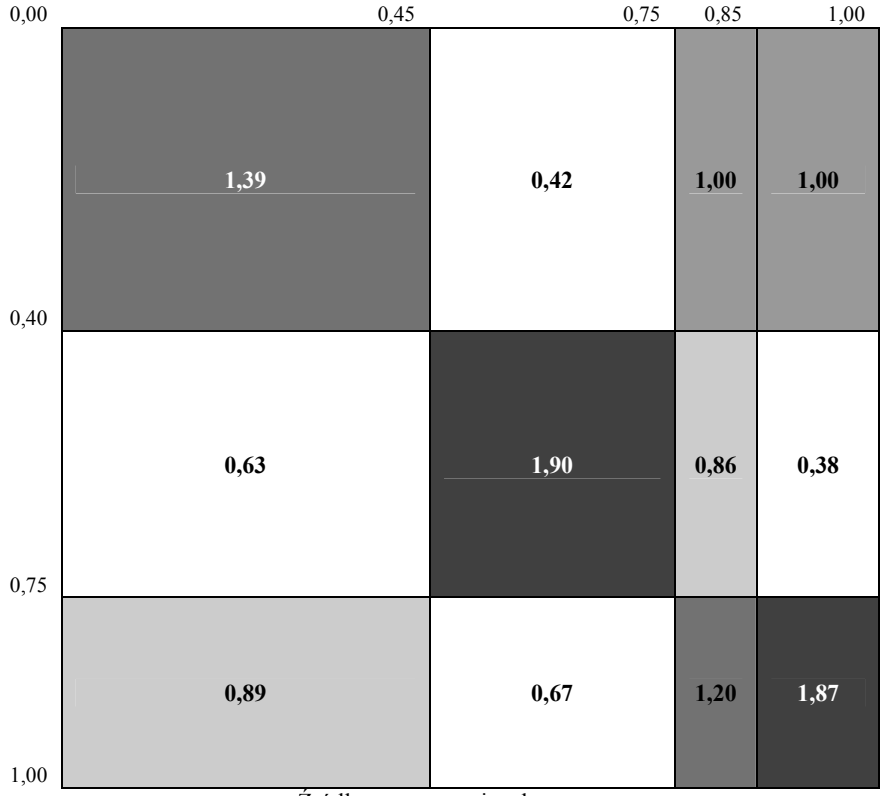
2.8. Mapy nadreprezentacji

Dla lepszego zobrazowania, czym są wskaźniki nadreprezentacji, dzielimy kwadrat jednostkowy na kolumny o szerokościach proporcjonalnych do sum kolumn Tabeli 7 oraz na wiersze o wysokościach proporcjonalnych do sum wierszy Tabeli 7. Komórki cieniujemy, stosując następujący kod kolorów:

- dla $c_{ij} \leq 2/3$,
- dla $2/3 < c_{ij} < 0{,}99$,
- dla $0{,}99 < c_{ij} \leq 1/0{,}99$,
- dla $1/0{,}99 < c_{ij} \leq 3/2$,
- dla $c_{ij} > 3/2$.

Tak opracowaną tablicę nazywamy **mapą nadreprezentacji**. Jak łatwo zauważyć, mapa nadreprezentacji dla rozkładu proporcjonalnego byłaby jednostajnie szara. Na Rys. 7 pokazano mapę nadreprezentacji dla rozkładu z Tabeli 7. Dodatkowo w poszczególnych komórkach podano wskaźniki nadreprezentacji.

Rysunek 7. Mapa nadreprezentacji dla rozkładu z Tabeli 7.

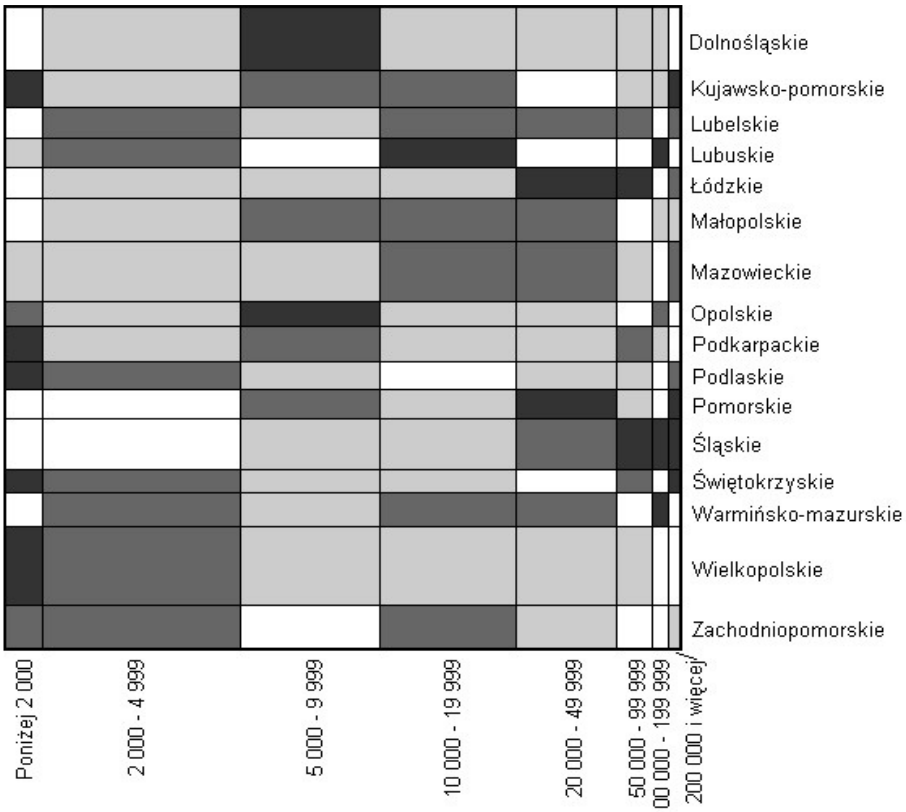


Źródło: opracowanie własne.

Uwaga na pole w 3 wierszu i 2 kolumnie – patrz notka do Tabeli 8.

Na Rys. 8 przedstawiono mapę nadreprezentacji dla zestawienia miast w poszczególnych województwach Polski sporządzoną w programie GradeStat. Można z niej odczytać, że w Polsce dominują miasta małe (2 000 – 20 000 mieszkańców); najwięcej miast jest w województwie wielkopolskim, najmniej – w świętokrzyskim; skupiskiem największej liczby dużych miast jest województwo śląskie.

Rysunek 8. Mapa nadreprezentacji dla rozkładu wielkości miast w poszczególnych województwach Polski.



Źródło: opracowanie własne na podstawie danych GUS.

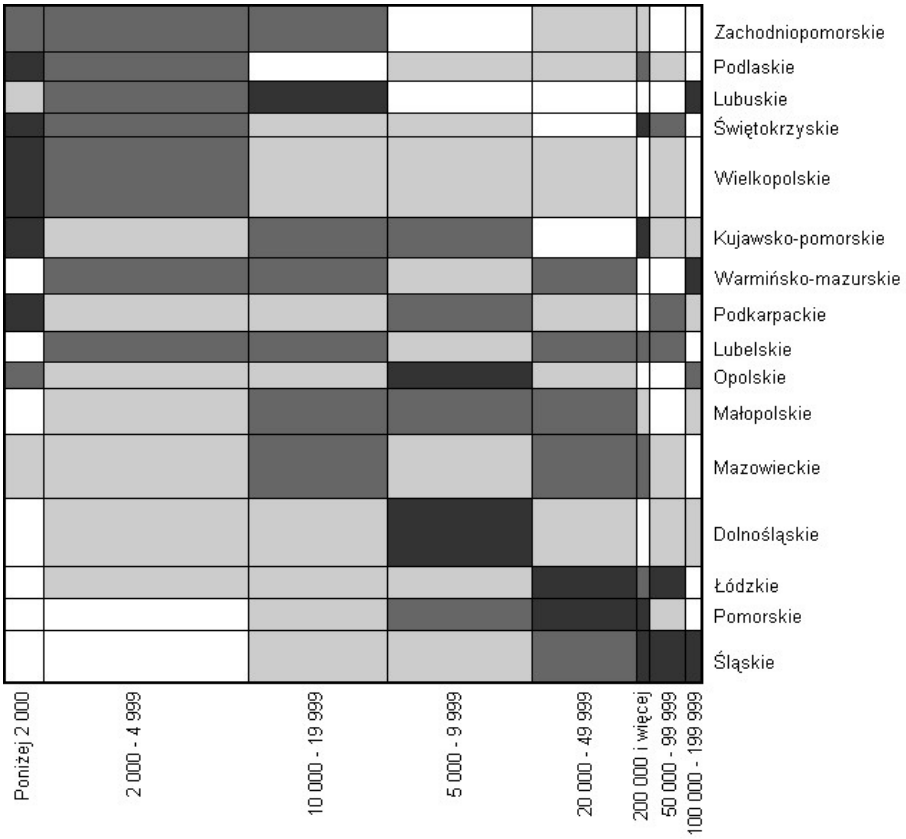
3. Analiza danych przy pomocy GCA

Jak zaznaczyliśmy w punkcie 2.6, algorytm GCA dąży do takiej zmiany ustawienia wierszy i kolumn tabeli danych, by wartość współczynnika korelacji rang Spearmana ρ^* była możliwie bliska maksymalnej. GCA jest metodą Monte Carlo, toteż nie gwarantuje uzyskania maksymalnej wartości ρ^*, jednak daje wyniki bardzo jej bliskie.

Tabela danych po wykonaniu na niej GCA staje się bardziej regularna: obszary o tym samym stopniu szarości na jej mapie nadreprezentacji tworzą w miarę

zwarte skupienia, przy czym obszary najciemniejsze układają się możliwie blisko przekątnej. Na Rys. 9 pokazano wynik zastosowania GCA do danych nt. wielkości miast w poszczególnych województwach Polski.

Rysunek 9. Mapa nadreprezentacji dla rozkładu wielkości miast w poszczególnych województwach Polski po przeprowadzeniu GCA.



Źródło: opracowanie własne na podstawie danych GUS.

Uwaga: po wykonaniu algorytmu GCA można uzyskać wykres, w którym porządek wierszy i kolumn będzie odwrotny: u góry województwo śląskie, następnie pomorskie, łódzkie itd. aż do zachodniopomorskiego, miasta kolejno: „100 000 – 199 999", „50 000 – 99 999" itd. aż do „Poniżej 2 000". Oba wykresy niosą dokładnie tę samą informację.

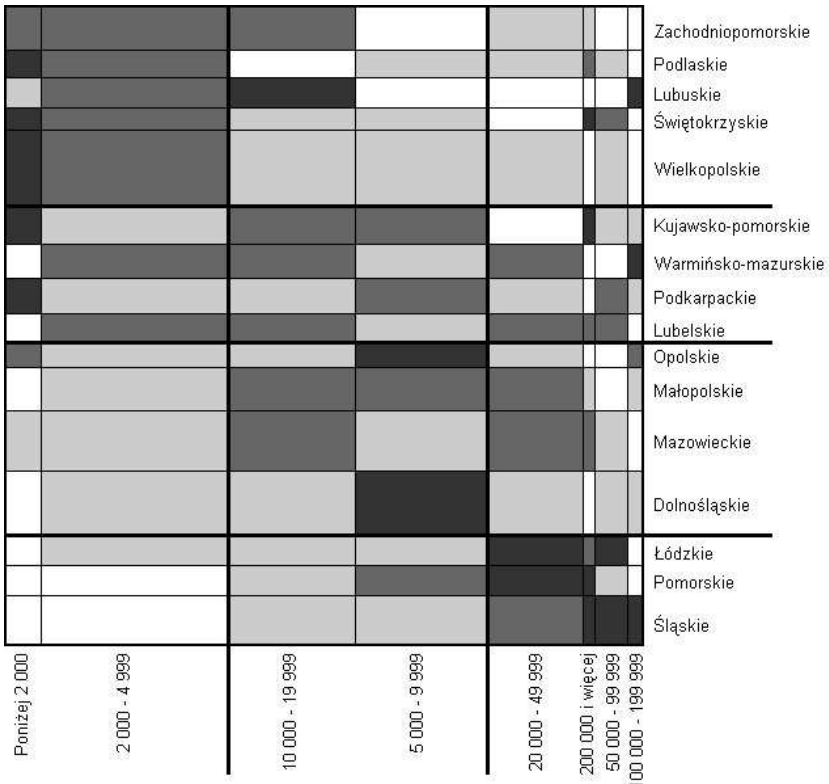
3.1. Analiza skupień

Po GCA wiersze i kolumny bliskie sobie (czyli takie, których rozkłady są do siebie w miarę zbliżone) sąsiadują ze sobą w tabeli. Dzięki temu jest możliwe przeprowadzenie analizy skupień, zarówno dla wierszy (obiektów), jak i kolumn (cech obiektów). Nie ma jednoznacznego kryterium mówiącego, ile skupień należy wy-

różnić. Zależy to od preferencji badacza. Zazwyczaj trzeba przeprowadzić kilka prób, by zauważyć strukturę danych.

Na Rys. 10 pokazano wynik analizy skupień dla wielkości miast w poszczególnych województwach Polski. Przyjęto 4 skupienia dla obiektów (województw) i 3 skupienia dla cech (wielkości miast).

Rysunek 10. Analiza skupień dla wielkości miast w poszczególnych województwach Polski.



Źródło: opracowanie własne na podstawie danych GUS.

Pierwsze skupienie obiektów obejmuje województwa: zachodniopomorskie, podlaskie, lubuskie, świętokrzyskie i wielkopolskie. Są to województwa, które na tle kraju charakteryzują się dużą liczbą małych miast (poniżej 5 000).

Drugie skupienie tworzą województwa: kujawsko-pomorskie, warmińsko-mazurskie, podkarpackie i lubelskie. Są to województwa wyróżniające się dużym odsetkiem miast liczących od 2 000 do 19 999 mieszkańców. Co do miast liczących mniej niż 2 000 lub co najmniej 20 000 mieszkańców – to skupienie stanowi prawdziwą mozaikę.

Trzecie skupienie, złożone z województw: opolskiego, małopolskiego, mazowieckiego i dolnośląskiego, to skupienie, w którym miasta małe ustępują liczeb-

nie miastom średnim. Nie oznacza to, że miast małych jest bardzo mało; są – o czym świadczy szary kolor odpowiednich komórek – jednak nie dominują jak w skupieniu 1. W skupieniu 3 pojawiają się również w większej liczbie miasta duże (20 000 mieszkańców i więcej).

Skupienie czwarte obejmuje województwa o silnej „krajowej nadreprezentacji" miast dużych. Należą do niego województwa: łódzkie, pomorskie i śląskie. Co ciekawe, mają one również silną „nadreprezentację lokalną" miast dużych – proporcje pomiędzy liczebnością tych miast oraz miast mniejszych są w nich zdecydowanie inne niż w innych województwach. Co więcej – w tym skupieniu występuje również silna niedoreprezentacja miast małych i nieznaczna miast średnich.

Warto zauważyć, że choć wykonanie GCA zmieniło kolejność kolumn, to jednak wyróżnione skupienia mają charakter naturalny: miasta małe, średnie i duże. Nie doszło do połączenia miast największych z najmniejszymi.

3.2. Wykrywanie elementów odstających

Jak wspomnieliśmy w punkcie 2.6, zasadą działania GCA jest takie uporządkowanie kolumn i wierszy tabeli danych, by wskaźniki koncentracji *ar* dla każdej pary rozkładów (wierszy lub kolumn) były możliwie bliskie $ar_{\max}$, czyli wartości maksymalnej. GCA dąży do minimalizacji wartości średniej różnic *ar* i $ar_{\max}$, oznaczanej $AvgDistA_{row}$ dla wierszy oraz $AvgDistA_{col}$ dla kolumn. Jak wykazali Kowalczyk i in. (2004), średnie te wyrażają się wzorami:

$$AvgDistA_{row}(i;P)=\sum_{s=1}^{i-1}\frac{ar_{\max}(i:s;row(P)-ar(i:s;row(P)))}{(m-1)\sqrt{2}}+ \\ +\sum_{s=i+1}^{m}\frac{ar_{\max}(s:i;row(P)-ar(s:i;row(P)))}{(m-1)\sqrt{2}},\qquad i=1,...,m, \tag{14}$$

$$AvgDistA_{col}(j;P)=\sum_{t=1}^{j}\frac{ar_{\max}(j:t;col(P)-ar(j:t;col(P)))}{(k-1)\sqrt{2}}+ \\ +\sum_{t=j+1}^{k}\frac{ar_{\max}(t:j;col(P)-ar(t:j;col(P)))}{(k-1)\sqrt{2}},\qquad j=1,...,k. \tag{15}$$

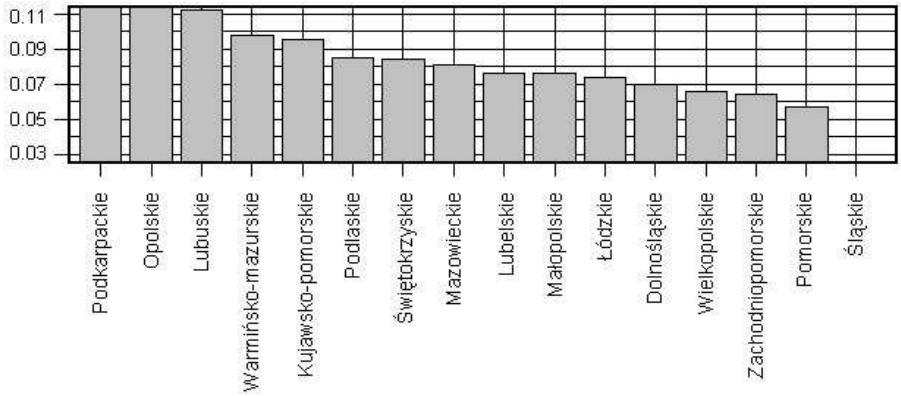
gdzie *P* to tabela danych licząca *m* wierszy i *k* kolumn.

Z tej zasady można wysnuć wniosek, że te elementy (wiersze – obiekty lub kolumny – cechy), dla których średnia wartość *AvgDistA* jest znacznie większa niż dla pozostałych, są elementami odstającymi, tj. takimi, które znacząco dobiegają od ogólnego trendu.

Na Rys. 11 pokazano wykres owych średnich dla województw (wierszy), a na Rys. 12 – dla wielkości miast (kolumn). Daje się zauważyć, że wartości *AvgDistA* dla województw są w miarę wyrównane. Można byłoby co najwyżej wskazać województwa: podkarpackie, opolskie i lubelskie jako nieco odmienne, mają bowiem

najwyższe (prawie jednakowe) wartości *AvgDistA*. Przyczynę tego można wyczytać z Rys. 10: w odpowiadających im wierszach mapy nadreprezentacji znajdują się „przeplatanki" kolorów, podczas gdy w pozostałych kolory tworzą bardziej zwarte bloki.

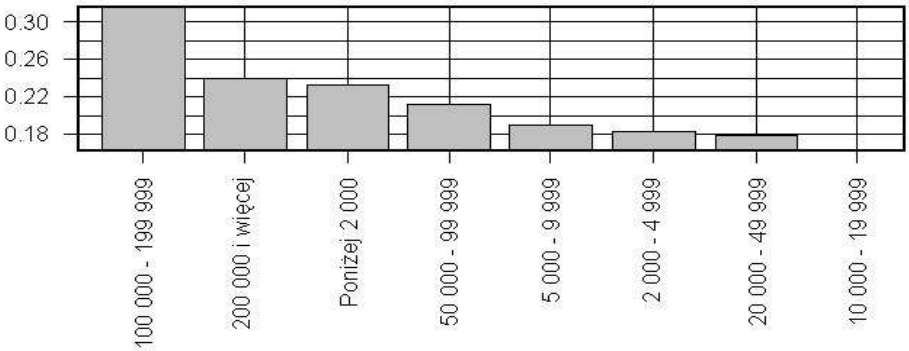
Rysunek 11. Wartości *AvgDistA* dla województw.



Źródło: opracowanie własne na podstawie danych GUS.

Inna sytuacja ma miejsce w przypadku wielkości miast. Tutaj wyraźnie widać jeden element odstający – miasta o liczbie mieszkańców od 100 000 do 199 999. Warto zauważyć, że trudno byłoby je wskazać intuicyjnie.

Rysunek 12. Wartości *AvgDistA* dla wielkości miast.



Źródło: opracowanie własne na podstawie danych GUS.

4. Zastosowanie GCA. Analiza integracji krajów Unii Europejskiej

4.1. Dane surowe

Dane wykorzystane w niniejszym przykładzie pochodzą ze strony Eurostatu (http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/). Eurostat publi-

kuje tabele zawierające dane zebrane w ciągu kilku kolejnych lat. Na ich podstawie na potrzeby niniejszej analizy utworzono tabelę z danymi za rok 2010:

- PKB na głowę mieszkańca – produkt krajowy brutto na głowę mieszkańca w jednostkach siły nabywczej (ang. *Purchasing Power Standard, PPS*). Jednostka siły nabywczej jest „sztuczną walutą" wykorzystywaną przez Eurostat. Odzwierciedla różnice w krajowych poziomach cen, których nie uwzględniają kursy wymiany walut. Jednostka ta pozwala na znaczące wielkościowe porównania wskaźników ekonomicznych.
- Stopa wzrostu PKB – stopa wzrostu PKB w stosunku do roku poprzedniego (w procentach).
- Stopa inflacji – roczna stopa inflacji (w procentach).
- Dług publiczny – dług publiczny w stosunku do PKB (w procentach).
- Stopa bezrobocia – stopa bezrobocia wśród osób w wieku od 15 do 74 lat zdolnych do podjęcia pracy (w procentach).
- Średnia długość życia kobiet – średnia długość życia kobiet (w latach).
- Średnia długość życia mężczyzn – średnia długość życia mężczyzn (w latach).
- Szerokopasmowy Internet – liczba łączy internetowych o przepustowości co najmniej 144 Kbits/s przypadająca na 100 mieszkańców.

Ponadto przy każdym kraju podano informację, czy należy on do strefy euro.

Oczywiście wybór tych właśnie cech opisujących badane kraje był arbitralny; dokonano go wyłącznie na potrzeby tego opracowania. Pełna analiza siły integracji krajów Unii Europejskiej wymagałaby objęcia analizą co najmniej kilkunastu cech zmierzonych w okresie co najmniej kilku lat.

4.2. Dane wejściowe GCA

Dane surowe w postaci takiej jak w Tabeli 9 nie mogą stanowić wejścia GCA z dwóch powodów.

Po pierwsze: występują w nich braki oznaczone w odpowiednich komórkach tabeli dwukropkiem (średnia długość życia kobiet i mężczyzn na Cyprze, w Rumunii i we Włoszech). W takim przypadku należy albo wyłączyć z badania dane niekompletne (4 wymienione kraje lub 2 kolumny z danymi o długości życia), albo uzupełnić braki. Dane brakujące można uzupełniać na wiele sposobów, np. poprzez wyliczenie średniej z otoczenia (w tym przypadku byłaby to średnia z odpowiedniej kolumny). Takie uzupełnienie mogłoby jednak zbyt zaburzyć wynik badania, postanowiono więc skorzystać z dostępności na stronie Eurostatu danych zebranych w latach 2004-2009. Braki zastąpiono średnimi wyliczonymi dla każdego kraju (i każdej płci) w tym okresie.

Po drugie: tabela zawiera wartości ujemne (w kolumnach „stopa wzrostu PKB" i „stopa inflacji"). Dane te są bardzo istotne, toteż powinny zostać zachowane. Problem ujemnych wartości rozwiązano, dodając do tabeli dwie kolumny: stopa spadku PKB oraz stopa deflacji. Wartości ujemne w oryginalnych kolumnach zastą-

piono zerami, a w ich miejsce w nowych kolumnach wpisano ich moduły. Tak przygotowane dane zawiera Tabela 10.

Tabela 9. Dane nt. 27 krajów Unii Europejskiej w 2010 roku

Kraj	Strefa euro	PKB na głowę mieszkańca	Stopa wzrostu PKB	Stopa inflacji	Dług publiczny	Stopa bezrobocia	Śr. długość życia kobiet	Śr. długość życia mężczyzn	Szerokopasmowy Internet
Austria	Tak	30 800	2,3	1,7	71,9	4,4	83,00	77,60	23,5
Belgia	Tak	29 000	2,2	2,3	96,0	8,3	83,50	77,90	30,0
Bułgaria	Nie	10 700	0,4	3,0	16,3	10,2	77,40	70,30	13,9
Cypr	Tak	24 200	1,1	2,6	61,5	6,2	:	:	23,2
Czechy	Nie	19 400	2,7	1,2	38,1	7,3	80,90	74,50	20,4
Dania	Nie	31 000	1,3	2,2	42,9	7,5	81,40	77,20	38,2
Estonia	Tak	15 700	2,3	2,7	6,7	16,9	80,80	70,60	26,0
Finlandia	Tak	28 100	3,7	1,7	48,4	8,4	83,50	76,90	29,1
Francja	Tak	26 300	1,5	1,7	82,3	9,8	85,30	78,30	31,5
Grecja	Tak	21 900	-3,5	4,7	145,0	12,6	82,80	78,40	18,6
Hiszpania	Tak	24 500	-0,1	2,0	61,2	20,1	85,30	79,10	22,5
Holandia	Tak	32 500	1,7	0,9	62,9	4,5	83,00	78,90	38,4
Irlandia	Tak	31 100	-0,4	-1,6	92,5	13,7	83,20	78,70	22,9
Litwa	Nie	14 000	1,4	1,2	38,0	17,8	83,50	77,90	19,6
Luksemburg	Tak	66 300	2,7	2,8	19,1	4,6	78,90	68,00	33,2
Łotwa	Nie	12 500	-0,4	-1,2	44,7	18,7	78,40	68,60	18,8
Malta	Tak	20 100	2,3	2,0	69,4	6,9	83,60	79,20	28,5
Niemcy	Tak	28 800	3,7	1,2	83,0	7,1	83,00	78,00	31,3
Polska	Nie	15 300	3,9	2,7	54,8	9,6	80,70	72,10	14,9
Portugalia	Tak	19 600	1,4	1,4	93,3	12	82,80	76,70	19,6
Rumunia	Nie	11 400	-1,6	6,1	30,5	7,3	:	:	13,7
Słowacja	Tak	17 900	4,2	0,7	41,1	14,4	79,30	71,70	15,5
Słowenia	Tak	20 700	1,4	2,1	38,8	7,3	83,10	76,40	23,6
Szwecja	Nie	30 300	6,1	1,9	39,4	8,4	83,60	79,60	31,9
Węgry	Nie	15 800	1,3	4,7	81,4	11,2	78,60	70,70	19,7
W. Brytania	Nie	27 400	2,1	3,3	79,6	7,8	82,60	78,60	30,6
Włochy	Tak	24 600	1,8	1,6	118,6	8,4	:	:	21,3

Źródło: Eurostat.

Przed wykonaniem algorytmu GCA dokonuje się normalizacji danych. Pierwszym krokiem jest normalizacja kolumn. Jej celem jest eliminacja „zdominowania" analizy przez jedną cechę (bądź kilka cech) o wartościach znacznie przewyższających te, jakie przyjmują pozostałe cechy (w naszym przypadku taką „dominującą" cechą byłby produkt krajowy brutto na głowę mieszkańca). Normalizacja kolumn polega na podzieleniu każdego wyrazu przez sumę danej kolumny. Cechy

powiązane można połączyć w grupę i normalizować wspólnie; wówczas każda wartość zostanie podzielona przez sumę wartości we wszystkich kolumnach danej grupy. W grupach łączy się cechy o podobnym charakterze i mierzone w tych samych jednostkach. Na przykład może być rozsądne połączenie w jednej grupie średniej długości życia kobiet i mężczyzn, ale nie ma sensu połączenie stopy inflacji i liczby szerokopasmowych łączy internetowych. W wyjątkowym przypadku można umieścić wszystkie kolumny w jednej grupie (jak miało to miejsce w opisanej wcześniej analizie wielkości miast).

Tabela 10. Dane z Tabeli 9 przygotowane do GCA.

Kraj	Strefa euro	PKB na głowę mieszkańca	Stopa wzrostu PKB	Stopa spadku PKB	Stopa inflacji	Stopa deflacji	Dług publiczny	Stopa bezrobocia	Śr. długość życia kobiet	Śr. długość życia mężczyzn	Szerokopasmowy Internet
Austria	Tak	30 800	2,3	0,0	1,7	0,0	71,9	4,4	83,00	77,60	23,5
Belgia	Tak	29 000	2,2	0,0	2,3	0,0	96,0	8,3	83,50	77,90	30,0
Bułgaria	Nie	10 700	0,4	0,0	3,0	0,0	16,3	10,2	77,40	70,30	13,9
Cypr	Tak	24 200	1,1	0,0	2,6	0,0	61,5	6,2	82,32	77,80	23,2
Czechy	Nie	19 400	2,7	0,0	1,2	0,0	38,1	7,3	80,90	74,50	20,4
Dania	Nie	31 000	1,3	0,0	2,2	0,0	42,9	7,5	81,40	77,20	38,2
Estonia	Tak	15 700	2,3	0,0	2,7	0,0	6,7	16,9	80,80	70,60	26,0
Finlandia	Tak	28 100	3,7	0,0	1,7	0,0	48,4	8,4	83,50	76,90	29,1
Francja	Tak	26 300	1,5	0,0	1,7	0,0	82,3	9,8	85,30	78,30	31,5
Grecja	Tak	21 900	0,0	3,5	4,7	0,0	145,0	12,6	82,80	78,40	18,6
Hiszpania	Tak	24 500	0,0	0,1	2,0	0,0	61,2	20,1	85,30	79,10	22,5
Holandia	Tak	32 500	1,7	0,0	0,9	0,0	62,9	4,5	83,00	78,90	38,4
Irlandia	Tak	31 100	0,0	0,4	0,0	1,6	92,5	13,7	83,20	78,70	22,9
Litwa	Nie	14 000	1,4	0,0	1,2	0,0	38,0	17,8	83,50	77,90	19,6
Luksemburg	Tak	66 300	2,7	0,0	2,8	0,0	19,1	4,6	78,90	68,00	33,2
Łotwa	Nie	12 500	0,0	0,4	0,0	1,2	44,7	18,7	78,40	68,60	18,8
Malta	Tak	20 100	2,3	0,0	2,0	0,0	69,4	6,9	83,60	79,20	28,5
Niemcy	Tak	28 800	3,7	0,0	1,2	0,0	83,0	7,1	83,00	78,00	31,3
Polska	Nie	15 300	3,9	0,0	2,7	0,0	54,8	9,6	80,70	72,10	14,9
Portugalia	Tak	19 600	1,4	0,0	1,4	0,0	93,3	12	82,80	76,70	19,6
Rumunia	Nie	11 400	0,0	1,6	6,1	0,0	30,5	7,3	76,48	69,22	13,7
Słowacja	Tak	17 900	4,2	0,0	0,7	0,0	41,1	14,4	79,30	71,70	15,5
Słowenia	Tak	20 700	1,4	0,0	2,1	0,0	38,8	7,3	83,10	76,40	23,6
Szwecja	Nie	30 300	6,1	0,0	1,9	0,0	39,4	8,4	83,60	79,60	31,9
Węgry	Nie	15 800	1,3	0,0	4,7	0,0	81,4	11,2	78,60	70,70	19,7
W. Brytania	Nie	27 400	2,1	0,0	3,3	0,0	79,6	7,8	82,60	78,60	30,6
Włochy	Tak	24 600	1,8	0,0	1,6	0,0	118,6	8,4	84,15	78,60	21,3

Źródło: opracowanie własne na podstawie danych Eurostatu.

Każdej kolumnie (grupie kolumn) przypisuje się wagę, która stanowi odbicie „ważności" poszczególnych cech w ogólnej ocenie badanych obiektów (w tym przypadku: krajów). Po znormalizowaniu kolumn (grup) każdą wartość mnoży się

przez wagę kolumny (grupy). Dobór wag ma ogromny wpływ na wynik badania, toteż powinien być efektem wnikliwej analizy natury cech oraz ich powiązań.

Drugim etapem normalizacji jest podzielenie każdego wyrazu przez sumę tabeli. W efekcie powstaje nowa tabela o sumie wyrazów równej 1.

Postanowiono, że długość życia kobiet i długość życia mężczyzn zostaną umieszczone w jednej grupie. To podejście zostało podyktowane przekonaniem, że o jakości życia w danym kraju decyduje średnia długość życia mieszkańców bez względu na płeć. Grupie nadano wagę równą 1. Cechy: „PKB na głowę mieszkańca", „dług publiczny", „stopa bezrobocia" i „szerokopasmowy Internet" umieszczono w samodzielnych grupach o wagach równych 1. Oznacza to, że każda z tych cech ma takie samo znaczenie oraz że każda z nich jest równie ważna, jak średnia długość życia kobiet i mężczyzn łącznie.

Cechy „stopa wzrostu PKB", „stopa spadku PKB", „stopa inflacji" i „stopa deflacji" stanowią pewien problem.

Po pierwsze: należy zdecydować, jak je pogrupować. Intuicja podpowiada, by połączyć w jednej grupie stopę wzrostu i stopę spadku PKB, a w drugiej – stopę inflacji i stopę deflacji. Jednak byłoby to posunięcie błędne. Rozważmy to na przykładzie stopy wzrostu i stopy spadku PKB. Po połączeniu ich w jedną grupę zostałyby znormalizowane do sumy obu kolumn, co nie ma sensu. Stopa spadku to ujemna stopa wzrostu; dane należałoby normalizować do różnicy (a nie sumy) pierwszej i drugiej kolumny. Podobna sytuacja ma miejsce w przypadku stopy inflacji i stopy deflacji. Tak więc każda z wymienionych cech musi się znaleźć w samodzielnej grupie.

Po drugie: trzeba ustalić wagi dla dla tych cech. Stopa wzrostu i stopa spadku PKB, które są cechami „siostrzanymi", powinny otrzymać identyczne wagi. Podobnie – stopa inflacji i stopa deflacji. Przyjmiemy, że stopa zmiany PKB ma znaczenie dwukrotnie mniejsze niż inne cechy – przypiszemy więc zarówno stopie wzrostu PKB, jak i stopie spadku wagę równą 0,5. Podobnie uczynimy ze stopą inflacji i stopą deflacji.

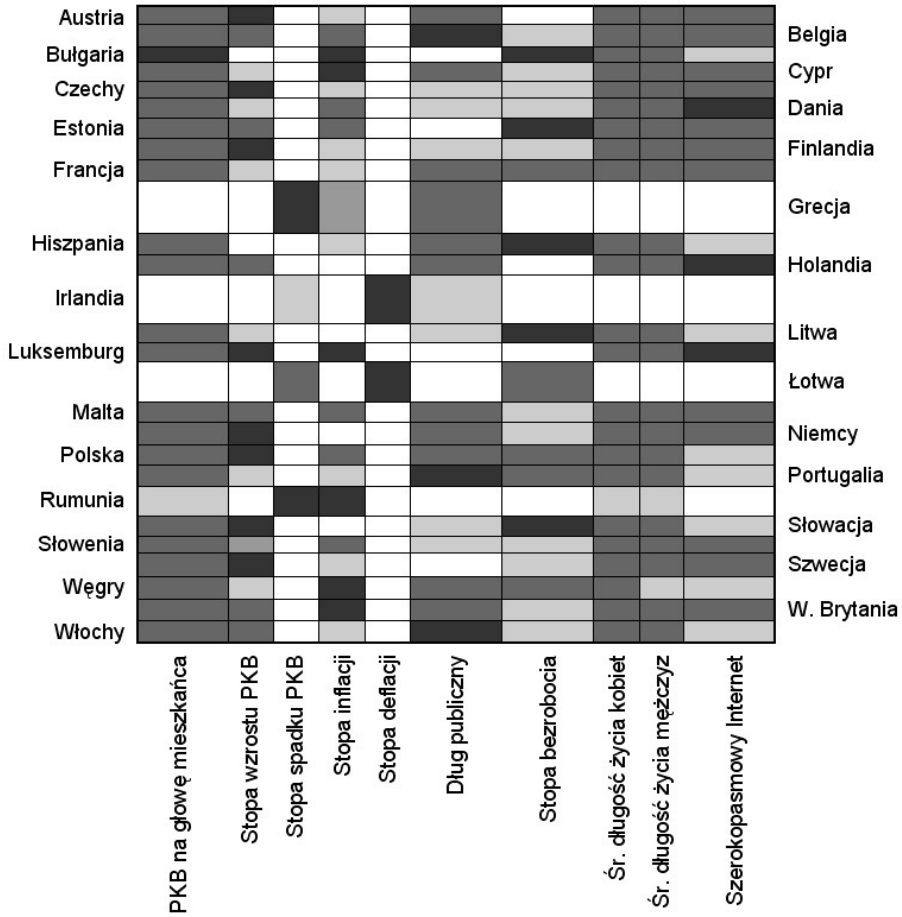
Cechę „Strefa euro" wyłączono z analizy, nie jest ona bowiem cechą liczbową. Zostanie w dalszej części wykorzystana do oznaczenia na wykresie krajów, których walutą jest euro.

Na Rys. 13 pokazano mapę nadreprezentacji dla surowych danych. Uwagę zwracają kolumny odpowiadające średniej długości życia kobiet i mężczyzn, stopom wzrostu i spadku PKB oraz inflacji i deflacji: są dwukrotnie węższe od pozostałych. W przypadku średniej długości życia jest to efektem umieszczenia obu kolumn w jednej grupie, w przypadku pozostałych cech – przypisania im mniejszych wag.

Nietrudno także spostrzec, że wiersze odpowiadające poszczególnym krajom mają różne szerokości. Co ciekawe, najszersze są wiersze odpowiadające krajom małym (Grecji, Irlandii i Łotwie), a krajom dużym (np. Francji, Hiszpanii, Niemcom) odpowiadają wiersze wąskie. Szerokość wierszy mapy nadreprezentacji jest proporcjonalna do sumy wierszy znormalizowanej tabeli cech. Wiersze odpowiada-

jące Grecji, Irlandii i Łotwie mają największe sumy (zob. Tabela 11), stąd też najszersze wiersze na mapie. Największe sumy wierszy znormalizowanej tabeli, a tym samym: najszersze wiersze macierzy nadreprezentacji oznaczają, że dane kraje „najwięcej ważą" w analizie – mają szczególnie wysokie wartości niektórych cech.

Rysunek 13. Mapa nadreprezentacji dla 27 krajów Unii Europejskiej.



Źródło: opracowanie własne na podstawie danych Eurostatu.

Tabela 11. Znormalizowana tabela danych.

Kraj	PKB na głowę mieszkańca	Stopa wzrostu PKB	Stopa spadku PKB	Stopa inflacji	Stopa deflacji	Dług publiczny	Stopa bezrobocia	Śr. długość życia kobiet	Śr. długość życia mężczyzn	Szerokopasmowy Internet	Suma wiersza
Austria	0,0068	0,0032	0,0000	0,0021	0,0000	0,0062	0,0023	0,0028	0,0026	0,0051	0,0310
Belgia	0,0064	0,0031	0,0000	0,0028	0,0000	0,0083	0,0044	0,0028	0,0026	0,0065	0,0368
Bułgaria	0,0024	0,0006	0,0000	0,0037	0,0000	0,0014	0,0054	0,0026	0,0024	0,0030	0,0213
Cypr	0,0053	0,0015	0,0000	0,0032	0,0000	0,0053	0,0033	0,0028	0,0026	0,0050	0,0290
Czechy	0,0043	0,0037	0,0000	0,0015	0,0000	0,0033	0,0038	0,0027	0,0025	0,0044	0,0262
Dania	0,0068	0,0018	0,0000	0,0027	0,0000	0,0037	0,0039	0,0027	0,0026	0,0083	0,0325
Estonia	0,0035	0,0032	0,0000	0,0033	0,0000	0,0006	0,0089	0,0027	0,0024	0,0056	0,0301
Finlandia	0,0062	0,0051	0,0000	0,0021	0,0000	0,0042	0,0044	0,0028	0,0026	0,0063	0,0337
Francja	0,0058	0,0021	0,0000	0,0021	0,0000	0,0071	0,0052	0,0029	0,0026	0,0068	0,0345
Grecja	**0,0048**	**0,0000**	**0,0417**	**0,0057**	**0,0000**	**0,0125**	**0,0066**	**0,0028**	**0,0026**	**0,0040**	**0,0808**
Hiszpania	0,0054	0,0000	0,0012	0,0024	0,0000	0,0053	0,0106	0,0029	0,0027	0,0049	0,0353
Holandia	0,0071	0,0024	0,0000	0,0011	0,0000	0,0054	0,0024	0,0028	0,0027	0,0083	0,0321
Irlandia	**0,0068**	**0,0000**	**0,0048**	**0,0000**	**0,0408**	**0,0080**	**0,0072**	**0,0028**	**0,0026**	**0,0050**	**0,0780**
Litwa	0,0031	0,0019	0,0000	0,0015	0,0000	0,0033	0,0094	0,0028	0,0026	0,0042	0,0288
Luksemburg	0,0146	0,0037	0,0000	0,0034	0,0000	0,0016	0,0024	0,0027	0,0023	0,0072	0,0379
Łotwa	**0,0027**	**0,0000**	**0,0048**	**0,0000**	**0,0306**	**0,0039**	**0,0098**	**0,0026**	**0,0023**	**0,0041**	**0,0608**
Malta	0,0044	0,0032	0,0000	0,0024	0,0000	0,0060	0,0036	0,0028	0,0027	0,0062	0,0313
Niemcy	0,0063	0,0051	0,0000	0,0015	0,0000	0,0072	0,0037	0,0028	0,0026	0,0068	0,0360
Polska	0,0034	0,0054	0,0000	0,0033	0,0000	0,0047	0,0051	0,0027	0,0024	0,0032	0,0302
Portugalia	0,0043	0,0019	0,0000	0,0017	0,0000	0,0080	0,0063	0,0028	0,0026	0,0042	0,0319
Rumunia	0,0025	0,0000	0,0190	0,0075	0,0000	0,0026	0,0038	0,0026	0,0023	0,0030	0,0433
Słowacja	0,0039	0,0058	0,0000	0,0009	0,0000	0,0035	0,0076	0,0027	0,0024	0,0034	0,0302
Słowenia	0,0046	0,0019	0,0000	0,0026	0,0000	0,0033	0,0038	0,0028	0,0026	0,0051	0,0267
Szwecja	0,0067	0,0085	0,0000	0,0023	0,0000	0,0034	0,0044	0,0028	0,0027	0,0069	0,0376
Węgry	0,0035	0,0018	0,0000	0,0057	0,0000	0,0070	0,0059	0,0026	0,0024	0,0043	0,0332
W. Brytania	0,0060	0,0029	0,0000	0,0040	0,0000	0,0069	0,0041	0,0028	0,0026	0,0066	0,0360
Włochy	0,0054	0,0025	0,0000	0,0020	0,0000	0,0102	0,0044	0,0028	0,0026	0,0046	0,0346
Suma kolumny	0,1429	0,0714	0,0714	0,0714	0,0714	0,1429	0,1429	0,0743	0,0686	0,1429	1,0000

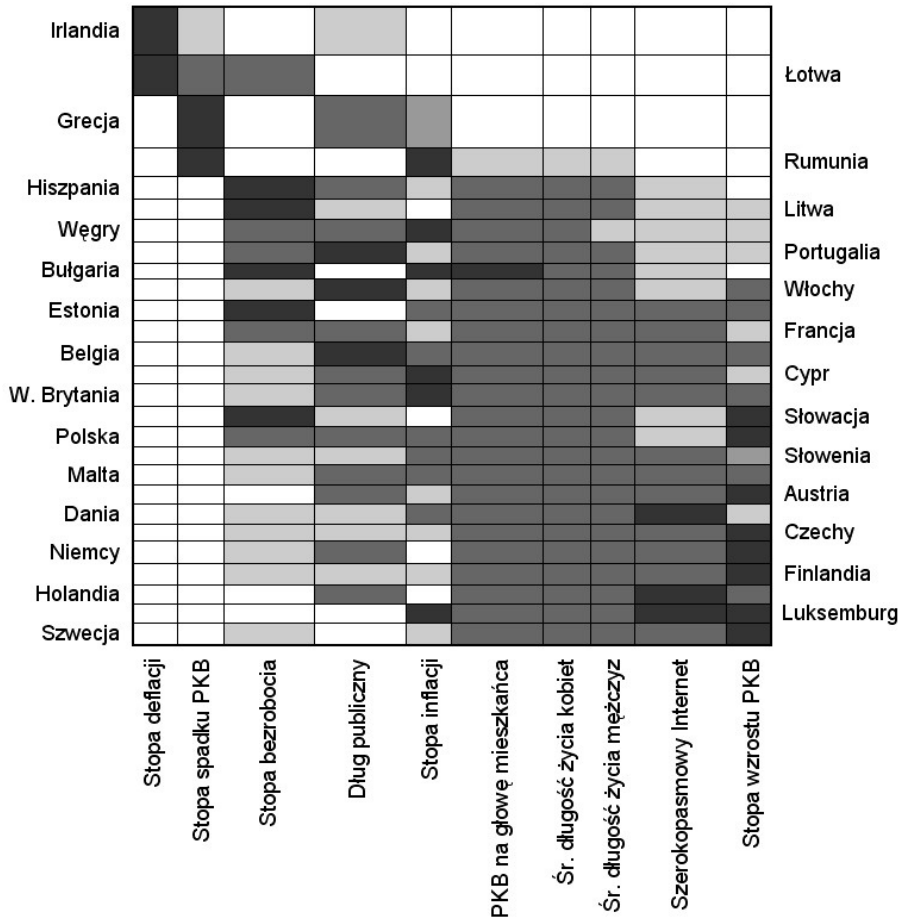
Źródło: opracowanie własne.

4.3. GCA

Po pogrupowaniu cech i przypisaniu im wag dane poddano przetworzeniu algorytmem GCA, wykorzystując program GradeStat. Na Rys. 14 pokazano mapę nadreprezentacji dla przetworzonych danych.

Jak widać, Irlandia, Łotwa i Grecja znalazły się blisko siebie. Oznacza to, że mają sporo cech wspólnych (na razie nie precyzujemy: dobrych czy złych). Daje się też zauważyć duży pionowy obszar o ciemnoszarym kolorze obejmujący aż 23 kraje. Jest to obszar odpowiadający produktowi krajowemu brutto na głowę mieszkańca oraz średniej długości życia (kobiet i mężczyzn). Można zaryzykować stwierdzenie, że badane kraje, choć w wielu obszarach znacznie różnią się między sobą, stanowią obszar, którego mieszkańcy dość długo żyją i cieszą się w miarę dobrym statusem materialnym.

Rysunek 14. Mapa nadreprezentacji dla 27 krajów Unii Europejskiej po przeprowadzeniu GCA.



Źródło: opracowanie własne na podstawie danych Eurostatu.

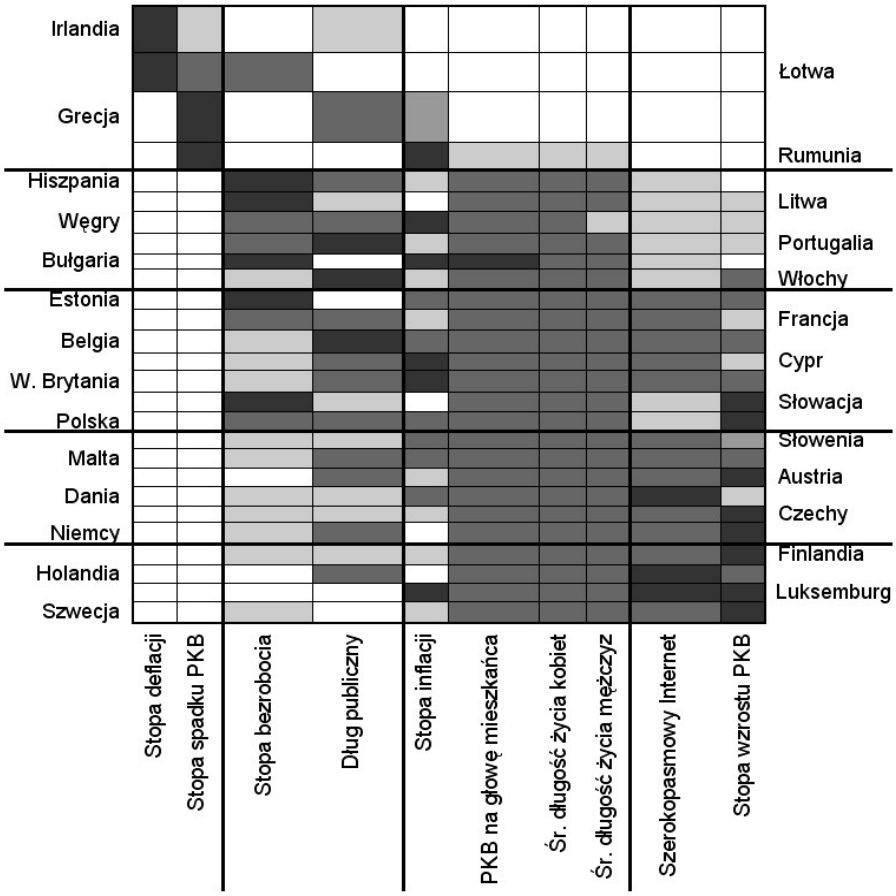
4.4. Analiza skupień

Dla danych przetworzonych GCA przeprowadzono analizę skupień. Przyjęto 5 skupień dla krajów i 4 skupienia dla cech. Wynik, przy tych założeniach, pokazano na Rys. 15.

Okazało się, że najważniejszą cechą różnicującą kraje jest zmiana PKB na głowę mieszkańca. Stopa wzrostu PKB oraz stopa spadku PKB znalazły się na przeciwległych krańcach tabeli. Stopa wzrostu utworzyła skupienie z dostępem do szerokopasmowego Internetu, stopa spadku – ze stopą deflacji. Pierwsze skupienie możemy nazwać „skupieniem rozwoju", drugie – „skupieniem stagnacji". Kolejne

skupienie cech objęło stopę bezrobocia i dług publiczny. Można byłoby je nazwać „skupieniem wad", bowiem należą do niego te cechy gospodarki, z którymi każdy kraj usiłuje walczyć (a przynajmniej je ograniczyć). Ostatnie skupienie cech tworzą: stopa inflacji, PKB na głowę mieszkańca oraz średnia długość życia kobiet i mężczyzn. To skupienie można byłoby nazwać „skupieniem jakości życia".

Rysunek 15. Analiza skupień dla 27 krajów Unii Europejskiej.



Źródło: opracowanie własne na podstawie danych Eurostatu.

Na uwagę zasługuje kwestia, do którego skupienia dołączyła stopa inflacji. Nie do tego, w którym znalazły się stopa bezrobocia i dług publiczny („skupienie wad"), tylko do zawierającego PKB na głowę mieszkańca oraz średnią długość życia („skupienie jakości życia"), choć intuicja podpowiada, że bardziej pasuje do tego pierwszego.

Kraje Unii ułożyły się w następujących skupieniach:

– Irlandia, Łotwa, Grecja i Rumunia. Kraje, w których panuje stagnacja. Łączy je spadek PKB. Ponadto Irlandia i Łotwa odczuwają skutki deflacji (zmniejszona

opłacalność produkcji, spadek konsumpcji i zamówień przemysłowych w oczekiwaniu niższe ceny). Warto zauważyć, że spadek PKB w dużym stopniu jest spowodowany deflacją – spadek konsumpcji powoduje zmniejszenie produkcji, a ono z kolei – spadek produktu krajowego.

Kraje tego skupienia charakteryzuje również silna niedoreprezentacja w obszarach rozwoju (wszystkie kraje) i jakości życia (poza Rumunią, która ma nieznaczną nadreprezentację). Rumunię i Grecję dodatkowo dotyka problem wysokiej inflacji.

Wydarzenia ostatnich miesięcy potwierdzają, że Grecja jest krajem wymagającym działań naprawczych. Dziwić jednak może obecność w tym skupieniu Irlandii uważanej przez wielu mieszkańców „nowych" krajów Unii za obszar dobrobytu. Przyszłość z pewnością zweryfikuje słuszność tej opinii – oraz wyników niniejszego badania.

To skupienie wyraźnie odróżnia się od pozostałych, które – w porównaniu z nim – są relatywnie jednorodne. Można zaryzykować stwierdzenie, że należące do niego kraje stanowią dla Unii „grupę specjalnej troski". Ma to, niewątpliwie, istotne znaczenie dla polityki europejskiej.

- Hiszpania, Litwa, Węgry, Portugalia, Bułgaria i Włochy. Kraje mające problemy z finansami publicznymi oraz bezrobociem. Problemy te są jednak „maskowane" przez dość wysoki standard życia ludności. Mieszkańcy tych krajów mają względnie dobry (na tle średniej unijnej) dostęp do szerokopasmowego Internetu. W większości krajów (poza Hiszpanią i Bułgarią) PKB na głowę mieszkańca rośnie, można więc mieć nadzieję, że problemy z bezrobociem i długiem publicznym uda się im przezwyciężyć, choć ani nie będzie to łatwe, ani nie nastąpi szybko.

  Na tle doniesień na temat pomocy unijnej, jakiej udzielono Hiszpanii, a której wkrótce mogą potrzebować również Portugalia i Włochy, umieszczenie tych krajów w jednym skupieniu raczej nie dziwi.

- Estonia, Francja, Belgia, Cypr, Wielka Brytania, Słowacja i Polska. Kraje o przyzwoitej, choć niekoniecznie rewelacyjnej kondycji. Ich mieszkańcy cieszą się dość dobrymi warunkami życiowymi i korzystają w nowych technologii (dość powszechny dostęp do szerokopasmowego Internetu). Kraje te odnotowują również wzrost PKB, a jednym z liderów tej grupy jest Polska. Mają jednak dość wysoką stopę bezrobocia i spory dług publiczny. Kłopoty z wypłacalnością, z jakimi boryka się poprzednie skupienie, mogą dosięgnąć również niektóre kraje tej grupy – choć raczej nie w najbliższej przyszłości.

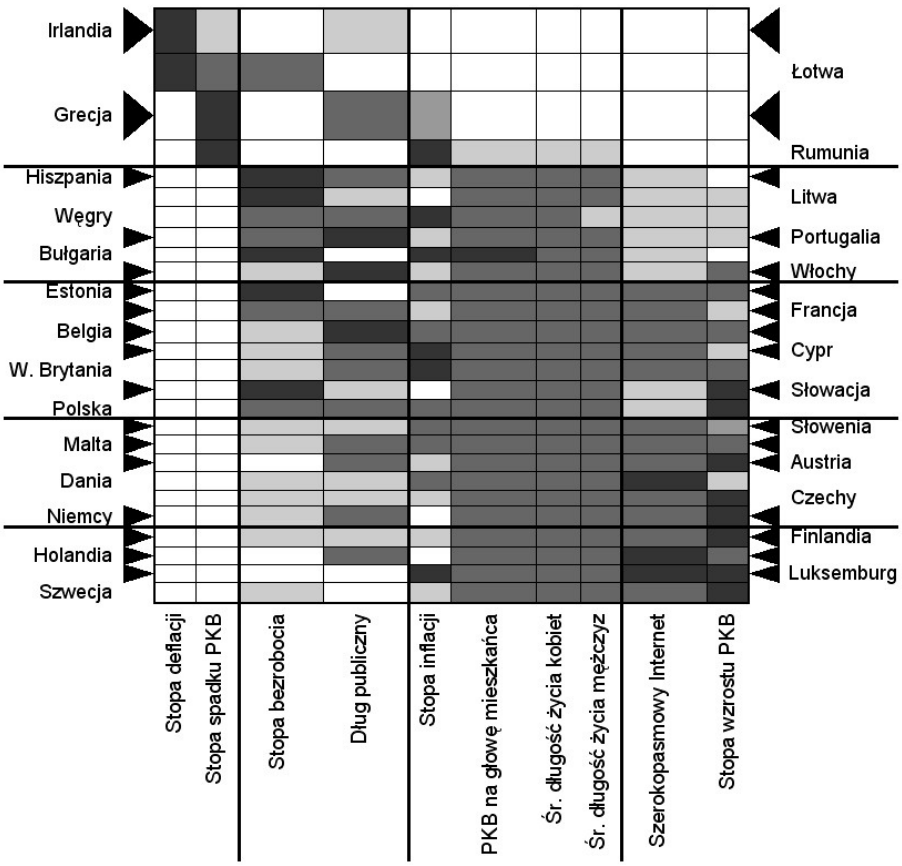
- Słowenia, Malta, Austria, Dania, Czechy i Niemcy. Kraje zapewniające swoim mieszkańcom dobre warunki życiowe i dostęp do nowoczesnych technologii. Mają stosunkowo niewielkie bezrobocie. Kłopoty z długiem publicznym, jakie ich nie omijają, będą mogły pokonać dzięki wysokiej (lub bardzo wysokiej) stopie wzrostu PKB.

- Finlandia, Holandia, Luksemburg, Szwecja. Kraje o bardzo wysokim tempie wzrostu PKB i powszechnym dostępie do nowoczesnych technologii. Zapewniają swoim mieszkańcom wysoki standard życia przy znikomym (z wyjątkiem Ho-

landii) zadłużeniu publicznym oraz niewielkim bezrobociu. Zdecydowani liderzy Unii.

Na podstawie cechy „Strefa euro" wprowadzono do tabeli znaczniki i sporządzono nową mapę nadreprezentacji (Rys. 16). Jak widać, w każdym skupieniu znalazły się zarówno kraje należące do „eurolandu", jak i spoza niego. Jednak z uwagi na fakt, że euro funkcjonuje stosunkowo krótko (o wiele krócej niż sama Unia Europejska), nie należałoby wyciągać z tego zbyt daleko idących wniosków.

Rysunek 16. Analiza skupień dla 27 krajów Unii Europejskiej.
Zaznaczono kraje strefy euro.



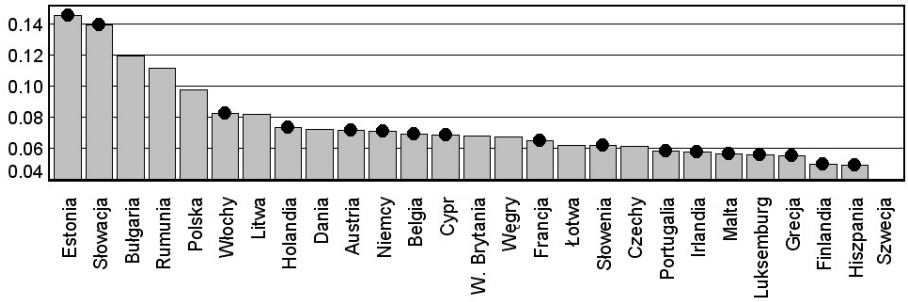
Źródło: opracowanie własne na podstawie danych Eurostatu.

4.5. Elementy odstające

Na Rys. 17 pokazano wyniki analizy elementów odstających wśród rozpatrywanych krajów. Okazuje się, że są to przede wszystkim Estonia i Słowacja, w

drugiej zaś kolejności: Bułgaria i Rumunia oraz – w mniejszym stopniu – Polska. Trudno byłoby wskazać te kraje na podstawie intuicji.

Rysunek 17. Elementy (kraje) odstające. Znacznikami oznaczono kraje strefy euro.



Źródło: opracowanie własne na podstawie danych Eurostatu.

Porównanie tych wyników z analizą skupień pozwala na nieco głębszą ocenę. Estonia i Słowacja, najbardziej „odstające od reszty", należą do trzeciego skupienia – tego, które na razie nie ma poważnych kłopotów. Mają jednak znacznie wyższe bezrobocie niż pozostałe kraje skupienia, stąd „nie przystają do reszty". Jeśli bezrobocie w najbliższym czasie nie spadnie, w badaniu przeprowadzonym na podstawie danych za kolejny rok mogą się przesunąć do skupienia drugiego.

Bułgarię zaliczono do skupienia drugiego – tego „z problemami". Jednak na tle pozostałych krajów skupienia wyróżnia się ona wysoką niedoreprezentacją w obszarze długu publicznego i stopy wzrostu PKB. Nie rozwija się w zawrotnym tempie, ale nie finansuje rozwoju „pożyczonymi pieniędzmi" (kraj niezbyt bogaty, ale rozsądny?). Dlatego jest w stanie poprawić swoją kondycję – i być może wkrótce przesunąć się do skupienia trzeciego.

Rumunia weszła w skład skupienia pierwszego, tego o najgorszej kondycji. Od pozostałych krajów różni ją jednak lepszy standard życiowy mieszkańców i – co bardzo ważne – małe zadłużenie i bezrobocie. Ma więc szansę przesunąć się do skupienia drugiego.

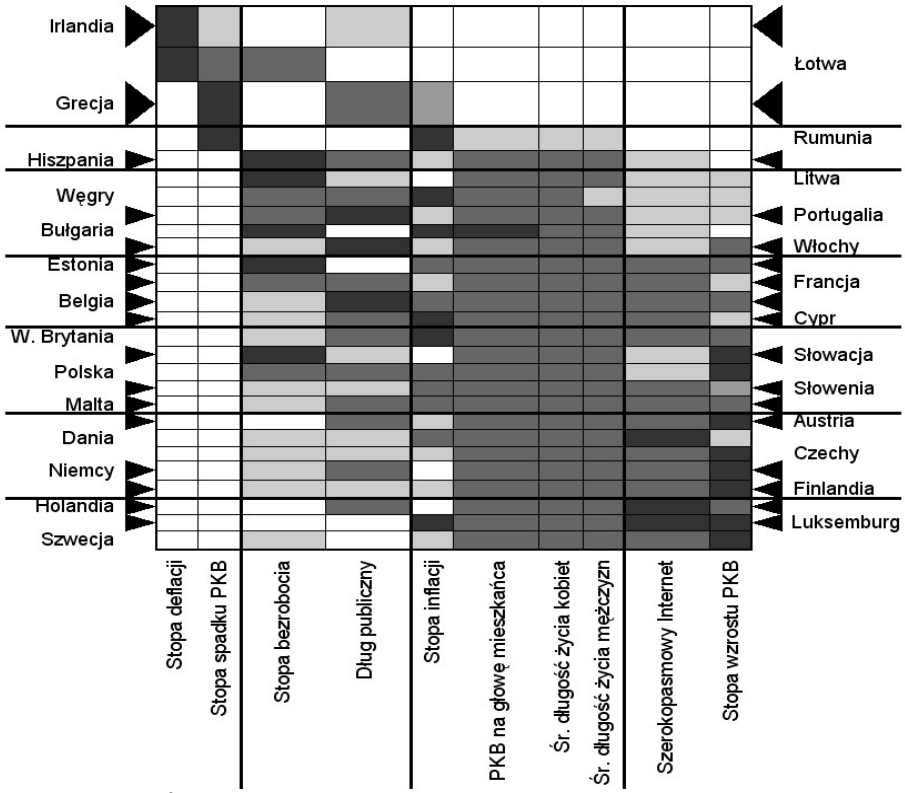
Aby zweryfikować powyższe rozważania, przeprowadzono podział krajów na siedem skupień. Wyniki przedstawia Rys. 18.

Estonia, „najbardziej odstająca", utworzyła nowe skupienie wraz z Francją, Belgią i Cyprem. Podobnie Słowacja, drugi w kolejności kraj na liście „odstających"; w ten sposób powstało skupienie „Wielka Brytania – Słowacja – Polska – Słowenia – Malta". Poprzednio Estonia i Słowacja znalazły się w jednym skupieniu. Ich obecne rozdzielenie oznacza, że nie tylko znacznie różnią się od reszty badanej zbiorowości, ale również – między sobą.

Trzecia na liście Bułgaria pozostała w swoim skupieniu, z którego jednak usunięto Hiszpanię. Do Hiszpanii dołączyła Rumunia, wychodząc z pierwszego (najgorszego) skupienia. Tym samym Hiszpania opuściła Litwę, Węgry, Portugalię, Bułgarię i Włochy, z którymi poprzednio ją połączono w skupieniu drugim. Świadczy to o tym, że jej obecna kondycja bardziej przypomina tę, jaką ma Rumunia (po-

wszechnie uważana za kraj biedny) niż tę, jaką mają zasobne (ale też mające problemy) Włochy.

Rysunek 18. Podział krajów na 7 skupień.



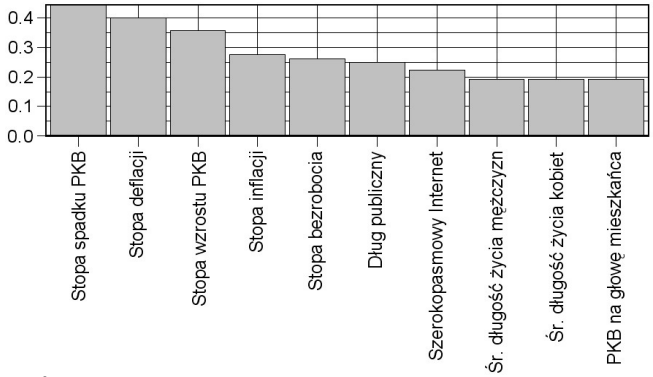
Źródło: opracowanie własne na podstawie danych Eurostatu.

Wśród cech (Rys. 19) brak wyraźnych elementów odstających. Należy jednak pamiętać, że na wynik analizy elementów odstających wśród cech znaczący wpływ mają przypisane im wagi.

4.6. Uwagi końcowe

Przeprowadzone obliczenia miały na celu przede wszystkim zilustrowanie gradacyjnej analizy odpowiedniości. Dla pełności badania należałoby dokonać bardziej szczegółowego doboru danych: uwzględnić większą liczbę cech zmierzonych w dłuższym okresie czasu. To, co zaprezentowano powyżej, stanowi „fotografię" Unii Europejskiej w roku 2010, bowiem obliczenia wykonano na podstawie wielkości pochodzących tylko z tego roku. Nie sposób więc ocenić tendencję (poprawa czy pogorszenie) w gospodarkach poszczególnych krajów.

Rysunek 19. Cechy odstające.



Źródło: opracowanie własne na podstawie danych Eurostatu.

Spore znaczenie dla uzyskanych wyników ma odpowiedni dobór wag cech. Tutaj był on bardzo subiektywny, podyktowany osobistym przekonaniem autora. Dlatego do niniejszej analizy należy podchodzić z dystansem. Nawet tak niewielka modyfikacja jak rozdzielenie średniej długości życia kobiet i mężczyzn (cech połączonych w grupę) i przypisanie każdej z nich wagi równej 1 zmienia uporządkowanie krajów na mapie nadreprezentacji po GCA, a w konsekwencji – wpływa na wyniki całego badania.

Wypada również nadmienić, że krytyczny wpływ na wynik badania ma wiarygodność danych źródłowych. Ufamy, że tutaj jest ona stuprocentowa.

5. Podsumowanie

Gradacyjna analiza danych wciąż jest metodą mało znaną, choć ma już stosunkowo długą historię. W niniejszym tekście opisaliśmy w zarysie jej podstawy teoretyczne i pokazaliśmy proste przykłady jej wykorzystania.

Podejście gradacyjne ma szereg zalet. Pozwala na jednolite traktowanie danych ciągłych i dyskretnych, jak też danych mierzonych na różnych skalach. Bada siłę powiązań pomiędzy danymi, umożliwiając wykrycie głównych trendów, często umykających intuicji, a czasami – sprzecznych z intuicją. Wyraża wyniki badań w liczbach i prezentuje je w przejrzystej postaci graficznej. Nadaje nowy sens pojęciu eksploracji danych.

Podstawowa metoda gradacyjnej analizy danych: gradacyjna analiza odpowiedniości opiera się na solidnym aparacie matematycznym, którego dogłębna znajomość nie jest jednak konieczna, by móc ją wykorzystywać. Uniwersalność, wszechstronność analizy i bogactwo oferowanych wyników czynią GCA przydatnym narzędziem dla badaczy o bardzo różnorodnych specjalnościach. Warto, by stała się bardziej popularna i znalazła liczniejsze zastosowania.

Literatura

Aldenderfer M.S., Blashfield R.K. (1984) *Cluster Analysis*. Series: Quantitative Applications in the Social Sciences, N^{o} 44. Sage University Papers, Newbury Park.

Biernacki M. (2006) Porządki generowane krzywą Lorenza. *Mathematical Economics*, 3 (10), 125-134, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Ciok A. (2004) *On the number of clusters – a grade approach*. Instytut Podstaw Informatyki PAN, Warszawa.

Ciok A., Kowalczyk T.,Pleszczyńska E., Szczesny W. (1995) Algorithms of grade correspondence – cluster analysis. *The Collected Papers of Theoretical and Applied Computer Science*, **7**, 1-4, 5-22.

Grabowska G., Wiech M. (2009) Grade analysis of data from the European Economic Survey 2005 on Economic Climate in Polish Servicing Sector. *Control and Cybernetics*, **28**, 3.

Jarochowska E., Grzegorek M., Hirny J., Maryja O., Wiech M. (2005) *Analiza danych medycznych i demograficznych przy użyciu programu GradeStat*. Instytut Podstaw Informatyki PAN oraz Instytut „Pomnik – Centrum Zdrowia Dziecka", Warszawa.

Johann M. (2005) *Polska – UE. Porównanie poziomu życia ludności.* Difin, Warszawa.

Kowalczyk T., Pleszczyńska E., Rulad F. (2004) Grade Models and Methods for Data Analysis with Applications for the Analysis of Data Populations. *Studies in Fuzziness and Soft Computing*, **151**; Springer Verlag, Berlin – Heidelberg – New York.

Krajewska-Siuda E., Szromek A.R. (2009) Wykorzystanie cech auksologicznych w klasyfikacji diagnostycznej niskorosłych pacjentów – próba wypracowania nowej metody klasyfikacji. *Zeszyty Naukowe Politechniki Śląskiej*, seria: *Organizacja i Zarządzanie*, 49, 175-185, Wydawnictwo Politechniki Śląskiej, Gliwice.

Majchrzyk Z. (2001) *Nieletni, młodociani i dorośli sprawcy zabójstw. Analiza procesów motywacyjnych i dyspozycji osobowościowych*. Instytut Psychiatrii i Neurologii, Warszawa.

Myatt G.J., Johnson W.P. (2009) *Making Sense of Data II. A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications.* John Wiley & Sons, Hoboken.

Noworol Cz. (1989) *Analiza skupień w badaniach empirycznych. Rozmyte modele hierarchiczne.* PWN, Warszawa.

Pleszczyńska E., Jarochowska E., Szczesny W. (2006) Wielowymiarowa analiza danych oparta na modelach gradacyjnych z implementacją w programie GradeStat. *Inżynieria wiedzy i systemy ekspertowe,* A. Grzech, red., tom 2, 73-83; Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

Romesburg H.Ch. (2004) *Cluster Analysis for Researchers*. Lulu Press, North Carolina.

Szromek A.R., Krajewska-Siuda E. (2008) *Koncepcja klasyfikacji diagnostycznej dzieci z niskorosłością i jej ekonomiczne implikacje.* Wydawnictwo Politechniki Śląskiej, Gliwice.

Siedlecka G. (2006) *Analiza porównawcza klimatów biznesowych państw europejskich przy wykorzystaniu metod gradacyjnych*. Plik dostępny w Internecie: http://gradestat.ipipan.waw.pl/ppt/Mgr_G.Siedlecka.ppt (05/05/2012). IPI PAN, Warszawa.

Szczesny W., Jarochowska E. (2006) *Gradacyjna analiza danych.* Plik dostępny w Internecie: http://www.ibspan.waw.pl/strony/podstrony/seminaria/Analiza_gradacyjna.pdf (05/05/2012), Instytut Podstaw Informatyki PAN, Warszawa.

Wiech M. (2007) *Metody gradacyjne w analizie danych wielowymiarowych: infrastruktura i implementacja.* Plik dostępny w Internecie: http://gradestat.ipipan.waw.pl/ppt/ MetodyGradacyjne.pdf (05/05/2012), Instytut Podstaw Informatyki PAN, Warszawa.

Wyka J. (2009) *Stan odżywienia ludzi po 60. roku życia w aspekcie uwarunkowań żywieniowych, zdrowotnych, środowiskowych i socjodemograficznych*. Wydawnictwo Uniwersytetu Przyrodniczego we Wrocławiu, Wrocław.

Wyrozumski T. (2002) Eksploracja danych – dlaczego nie w przemyśle? VIII Konferencja Polish Oracle User Group, Kościelisko.

GRADE DATA ANALYSIS – THE CONCEPT AND AN INSTANCE OF APPLICATION

Abstract: The article presents Grade Data Analysis – an original method of data exploration which has been elaborated and is still being developed by the Institute of Computer Science, Polish Academy of Sciences – as well as its basic tool: Grade Correspondence Analysis. Though based on sophisticated theory, the grade method is so universal that it may be used by researchers dealing with various areas of knowledge, even those very far from statistics and computer science.

The text is composed of three parts. The first is a description (rather general, because of its limited volume) of the grade analysis idea. In the second part we present the GCA algorithm. The third one is an example of GCA application: on the basis of statistical data available from the Eurostat website we divide the 27 EU countries into several groups. The results obtained may be surprising.

Keywords: data exploration, cluster analysis, Grade Correspondence Analysis, concentration curve, overrepresentation, outlier.