Kinga Kądziołka

Identyfikacja grup obiektów podobnych pod względem struktury zjawisk społeczno – ekonomicznych na przykładzie struktury wieku bezrobotnych

Streszczenie:

Celem pracy jest prezentacja możliwości aplikacyjnych metod grupowania danych do identyfikacji grup obszarów podobnych pod względem struktury analizowanego zjawiska. Zaprezentowana zostanie metoda grupowania hierarchicznego, w której do wyznaczania odległości między skupieniami wykorzystano wskaźnik niepodobieństwa struktur oraz przedstawiona zostanie propozycja metody grupowania niehierarchicznego, stanowiąca pewną analogię do metody k- średnich. Rozważania będą prowadzone na przykładzie oceny podobieństwa struktury wieku zarejestrowanych bezrobotnych w powiatach.

Słowa kluczowe:

podobieństwo struktur, wskaźnik podobieństwa struktur, grupowanie hierarchiczne, metoda k- średnich, struktura wieku bezrobotnych, bezrobocie

Wprowadzenie

Przedmiotem badań społeczno – ekonomicznych często jest struktura wybranych zjawisk. Porównywane są przykładowo wybrane jednostki terytorialne pod względem struktury wydatków, wieku ludności, bezrobocia, itp. W przypadku niewielkiej liczby obiektów, które porównywane są pod względem struktury danego zjawiska, wystarczające wydaje się utworzenie np. macierzy wskaźników podobieństwa struktur, zastosowanie metody eliminacji wektorów[1], celem identyfikacji grup obiektów o podobnej strukturze analizowanego zjawiska czy prezentacja danych statystycznych za pomocą tabel i wykresów. Tego typu metody oceny podobieństwa struktur w przypadku niewielkiej liczby porównywanych obiektów stosowali m. in[2].: E. Sojka, K. Kukuła, L. Luty, B. Bajan i K. Sowa, S. Lisek, K. Wawrzyniak i in[3]. W sytuacji większej ich liczby, przedstawienie wyników za pomocą macierzy wskaźników podobieństwa struktur czy wykresów obrazujących strukturę analizowanego zjawiska z uwzględnieniem wszystkich obiektów byłoby nieczytelne i znacznie utrudniałoby „ręczną" identyfikację grup obiektów podobnych pod względem struktury analizowanego zjawiska.

[1] S. Chomątowski, A. Sokołowski, Taksonomia struktur, Przegląd Statystyczny, 2, 1978, s. 217 – 226.

[2] Metoda eliminacji wektorów może oczywiście zostać zastosowana do dowolnej liczby obiektów porównywanych pod względem struktury.

[3] K. Kukuła, Z problematyki badań nad strukturą agrarną w Polsce w ujęciu przestrzennym, Oeconomia 6(4), 2007, s. 19 – 27; E. Sojka, Analiza porównawcza struktur i procesów ludnościowych w wybranych krajach UE z wykorzystaniem metod taksonomicznych, Folia Oeconomica, 253, 2011, s. 299 – 313; E. Sojka, Zmiany w procesach i strukturach ludnościowych w wybranych krajach UE, Demograficzne uwarunkowania rozwoju gospodarczego (red.) A. Rączaszek, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, 2012, s. 101 – 112; S. Lisek, Struktura wielkościowa przedsiębiorstw w Polsce, Metody Ilościowe w Badaniach Ekonomicznych, XVIII(4), 2017, s. 635-642; L. Luty, Regionalne zróżnicowanie struktury powierzchni użytków rolnych według systemów rolniczych w ujęciu dynamicznym, Metody Ilościowe w Badaniach Ekonomicznych, XVIII/2, 2017, s. 273 – 282; B. Bajan, K. Sowa, Food Consumption Models around the World in the Context of Globalization, Intercathedra, 3(40), 2019, s. 219 – 226; K. Wawrzyniak i in., The Similarity of European Union Countries in Terms of the Structure of the Unemployed, European Research Studies Journal, XXIII(4), 2020, s. 416 - 429.

Celem niniejszej pracy jest zaprezentowanie możliwości aplikacyjnych hierarchicznych i niehierarchicznych metod grupowania do identyfikacji grup obiektów podobnych pod względem struktury analizowanego zjawiska. Do wyznaczania odległości między obiektami zastosowano funkcję wykorzystującą wskaźnik niepodobieństwa struktur. Metody te umożliwią przejrzystą wizualizację tworzonych skupień obiektów podobnych pod względem struktury nawet w przypadku dużej liczby obiektów. Prezentowana metoda grupowania hierarchicznego została wcześniej zaprezentowana w pracy K. Kądziołki[4], jednakże analizy dotyczyły struktury przestępczości stwierdzonej na poziomie województw, więc rozpatrywanych obiektów było mało (16 regionów) i możliwa była ocena podobieństwa struktur z wykorzystaniem zarówno prostych wykresów, jak i danych tabelarycznych. W niniejszej pracy rozważania będą prowadzone na przykładzie identyfikacji grup powiatów podobnych pod względem struktury wieku zarejestrowanych bezrobotnych, w związku z czym prezentacja danych za pomocą tabeli wskaźników podobieństwa struktur lub wykresu obrazującego strukturę analizowanego zjawiska z uwzględnieniem wszystkich powiatów (na standardowej kartce formatu A4) byłaby niemożliwa. Przedstawiane wyniki uzyskano z wykorzystaniem darmowego programu R.

Dane i metody

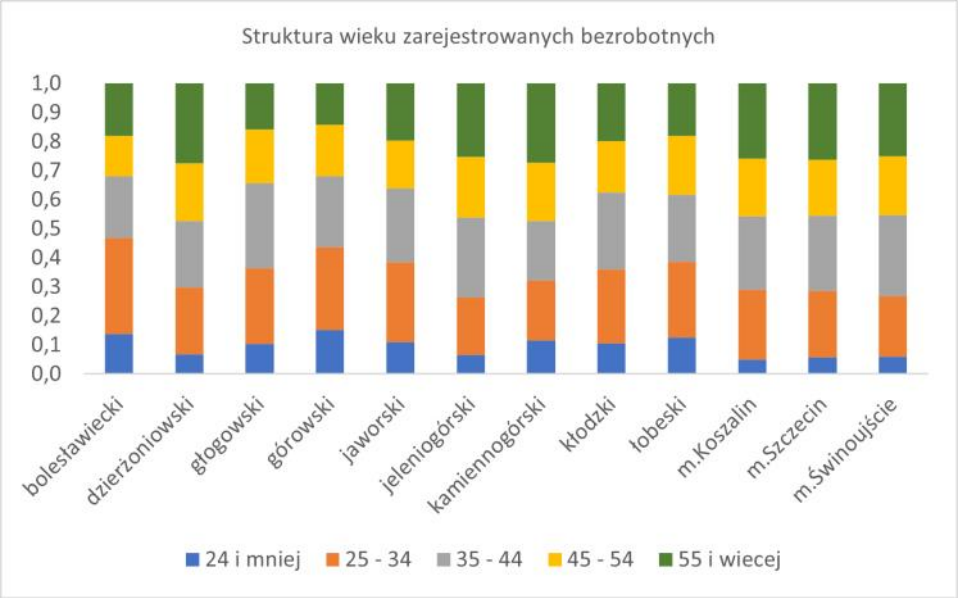
Przedmiotem analizy były ogólnodostępne dane publikowane przez GUS dotyczące wieku zarejestrowanych bezrobotnych w powiatach (wraz z miastami na prawach powiatu) w 2019 roku. Dane dotyczące liczby zarejestrowanych bezrobotnych w poszczególnych kategoriach wiekowych zostały przekonwertowane na dane obrazujące strukturę zarejestrowanych bezrobotnych wg poszczególnych kategorii wiekowych. Tabela 1 przedstawia fragment danych oryginalnych, a rys 1 strukturę wieku zarejestrowanych bezrobotnych dla danych z tabeli 1.

Tab. 1. Dane oryginalne (fragment)

Lp.	Powiat	Ogółem	24 i mniej	25 - 34	35 - 44	45 - 54	55 i więcej
1	bolesławiecki	1 293	178	428	276	177	234
2	dzierżoniowski	1 468	100	339	333	292	404
3	głogowski	2 036	209	533	596	376	322
4	górowski	1 530	231	436	376	269	218
5	jaworski	1 698	185	467	434	279	333
6	jeleniogórski	1 594	106	314	439	332	403
7	kamiennogórski	776	90	160	158	156	212
8	kłodzki	5 330	558	1 366	1 411	937	1 058
			⋮				
377	łobeski	2 032	255	529	471	412	365
378	m.Koszalin	2 202	111	527	558	438	568
379	m.Szczecin	4 274	244	982	1 104	824	1 120
380	m.Świnoujście	495	29	104	138	100	124

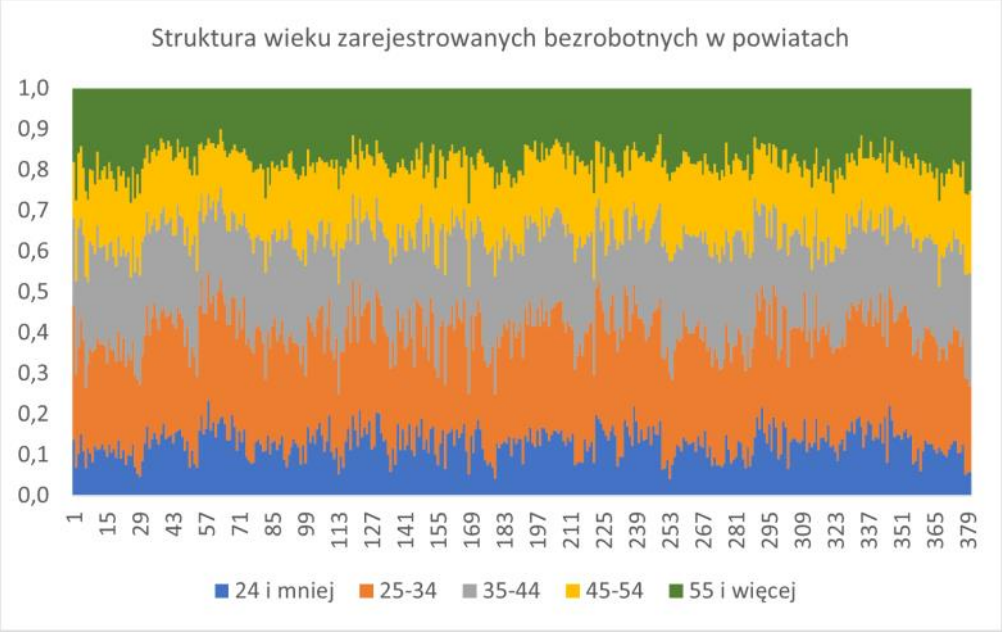
Źródło: GUS (Bank Danych Lokalnych).

[4] K. Kądziołka, Przestrzenne zróżnicowanie, struktura i dynamika przestępczości stwierdzonej w Polsce, Przestrzeń, Ekonomia, Społeczeństwo, 8/II, 2015, s. 223-235.



Rys. 1. Struktura wieku zarejestrowanych bezrobotnych dla danych z tabeli 1.
Źródło: opracowanie własne .

W przypadku analizy wszystkich powiatów identyfikacja obszarów podobnych pod względem struktury wieku zarejestrowanych bezrobotnych na podstawie wykresu analogicznego do rys. 1 byłaby niemożliwa. Dane te zaprezentowano na rys. 2. Na osi poziomej znajdują się numery porządkowe powiatów zamiast ich nazw z uwagi na przejrzystość wykresu. Tabela 2 przedstawia wartość minimalną i maksymalną odsetka bezrobotnych w ramach poszczególnych kategorii wiekowych dla wszystkich powiatów.



Rys. 2. Struktura wieku zarejestrowanych bezrobotnych w powiatach
Źródło: opracowanie własne.

Tab. 2. Wartość minimalna i maksymalna odsetka bezrobotnych w poszczególnych kategoriach

	24 i mniej	25-34	35-44	45-54	55 i więcej
min	0,0409	0,1759	0,1695	0,1243	0,0999
max	0,2321	0,3399	0,3194	0,2363	0,2820

Źródło: opracowanie własne.

Podobieństwo par struktur można wyznaczyć za pomocą wskaźnika podobieństwa struktur danego wzorem:

$$P_{ij}^{*} = \sum_{k=1}^{r} min(p_{ik}, p_{jk}) \Big/ \sum_{k=1}^{r} max(p_{ik}, p_{jk})$$

gdzie: *i, j* – numery obiektów, *k* – numer składnika struktury, pik – udział k- tego składnika w strukturze obiektu *i*, p_{jk} – udział k- tego składnika w strukturze obiektu *j*.

Przyjmuje się następującą wartość wskaźnika podobieństwa struktur[5]:

⇒ 0 – 0,2 – podobieństwo bardzo niskie

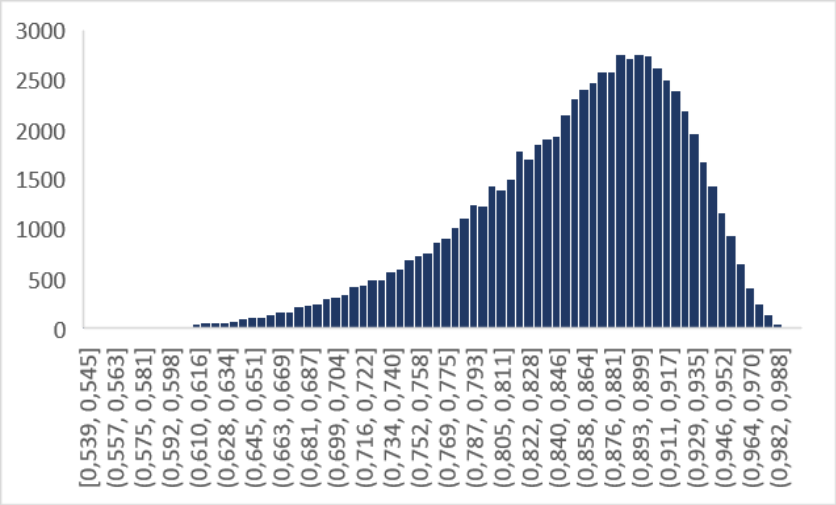
⇒ 0,2 – 0,4 – podobieństwo niskie

⇒ 0,4 – 0,6 – podobieństwo umiarkowane

⇒ 0,6 – 0,8 – podobieństwo duże

⇒ 0,8 – 1 – podobieństwo bardzo duże

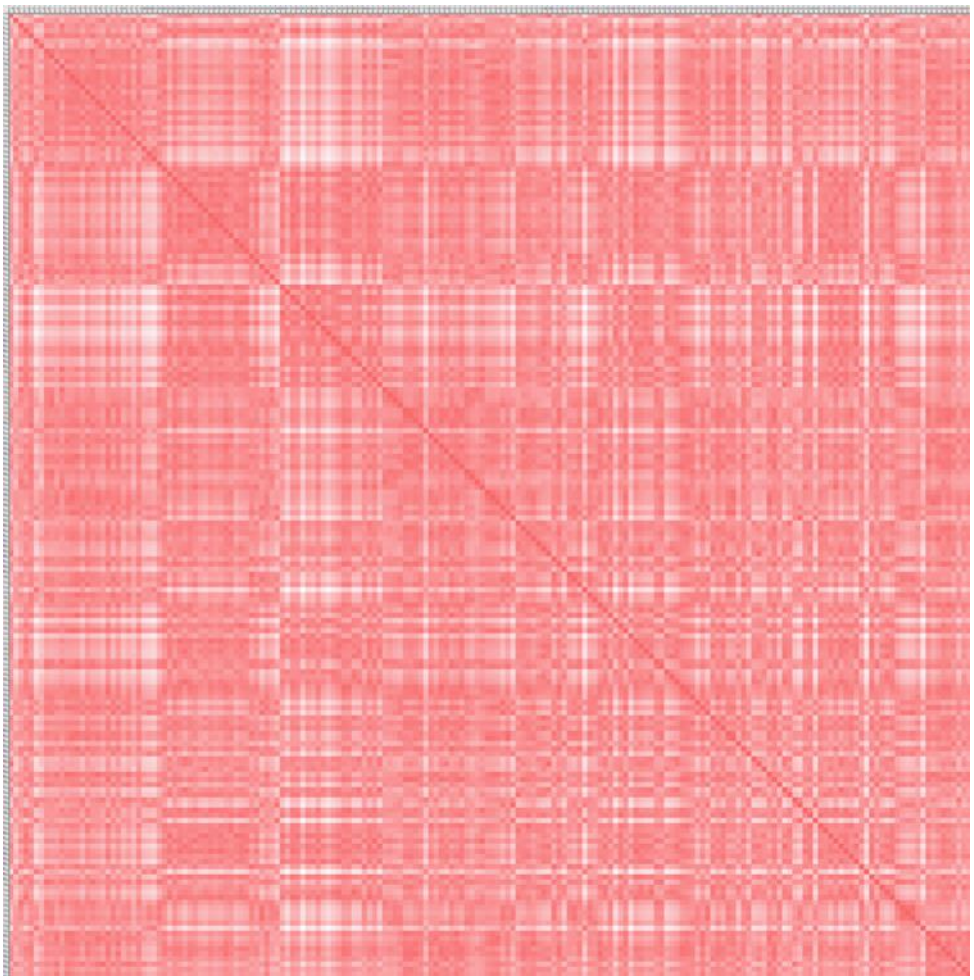
W przypadku gdy wartość wskaźnika podobieństwa struktur wynosi 1, struktury są identyczne. Rys. 3 przedstawia histogram wartości wskaźników podobieństwa struktur dla rozważanego zbioru danych.



Rys. 3. Histogram wartości wskaźników podobieństwa struktur dla analizowanych danych
Źródło: opracowanie własne.

[5] M. Sobczyk, Statystyka opisowa, Wydawnictwo C.H. Beck, Warszawa, 2010, s. 81.

Rys. 4 przedstawia dla analizowanych danych fragment macierzy wskaźników podobieństwa struktur z zaznaczonymi kolorem poziomami podobieństwa- im kolor ciemniejszy tym wskaźnik podobieństwa struktur wyższy. Jak zostało wspomniane, w przypadku dużej liczby obiektów, „ręczna" identyfikacja grup obiektów podobnych pod względem struktury analizowanego zjawiska na podstawie macierzy wskaźników podobieństwa struktur, byłaby zadaniem trudnym, o ile w ogóle możliwym do wykonania. W związku z czym wykorzystanie w tym celu innych metod, wydaje się zasadne.



Rys. 3. Graficzna prezentacja poziomów wartości wskaźników podobieństwa struktur (fragment)
Źródło: opracowanie własne.

Celem identyfikacji grup obiektów podobnych pod względem struktury analizowanego zjawiska wykorzystana zostanie w pierwszej kolejności metoda grupowania hierarchicznego[6]. Rezultatem działania metody grupowania hierarchicznego jest tzw. dendrogram, czyli drzewo hierarchicznie ułożonych skupień. Procedura grupowania hierarchicznego z wykorzystaniem tzw. metod aglomeracyjnych obejmuje następujące kroki[7]:

[6] Do identyfikacji grup obiektów podobnych pod względem struktury badanych zjawisk wykorzystywana była m. in. metoda Warda z odległością euklidesową, por. A. Majdzińska, Regionalizacja demograficzna. Wybrane metody i ich aplikacje, Wydawnictwo Uniwersytetu Łódzkiego, 2016, s. 106. Jednakże do oceny podobieństwa struktur powszechnie wykorzystywany bywa wskaźnik podobieństwa struktur, stąd zastosowanie miary niepodobieństwa struktur obiektów, wydaje się być zasadne.

[7] E. Gatnar, Statystyczna analiza danych z wykorzystaniem programu R, Wydawnictwo PWN, Warszawa, 2009, s. 413.

1) W macierzy odległości znajdź parę klas (skupień) najbardziej podobnych (najmniej odległych w sensie przyjętej miary odległości). Załóżmy, że są to klasy P_i i P_k.

2) Zredukuj liczbę skupień o jeden, łącząc skupienia P_i i P_k.

3) Przekształć odległości (zgodnie z przyjętą metodą wiązania skupień) między połączonymi skupieniami a pozostałymi skupieniami.

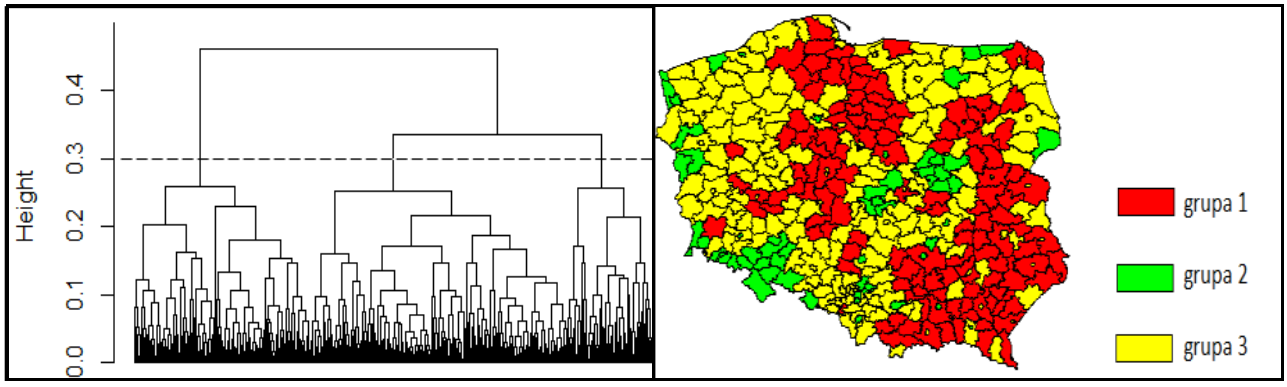
Powtarzaj kroki 1 – 3 aż wszystkie obiekty znajdą się w jednej klasie.

W niniejszej pracy do wiązania skupień wykorzystana zostanie metoda pełnego wiązania, w której odległość między skupieniami jest równa największej odległości między dwoma dowolnymi obiektami należącymi do różnych skupień[8].

Miara odległości (niepodobieństwa) między obiektami wyznaczana będzie następująco[9]: $d_{ij} = 1 - P_{ij}^*$, gdzie P_{ij}^* - wskaźnik podobieństwa struktur obiektów i oraz j. Miara ta nazywana bywa również wskaźnikiem niepodobieństwa struktur[10].

Wyniki

Rys. 5 przedstawia uzyskany dendrogram i podział powiatów na grupy obszarów podobnych pod względem struktury wieku zarejestrowanych bezrobotnych. Przyjęto podział powiatów na trzy grupy. Miejsce podziału dendrogramu zaznaczone zostało przerywaną linią. Rys. 6 przedstawia przeciętne wartości poszczególnych składników struktury dla uzyskanych grup.

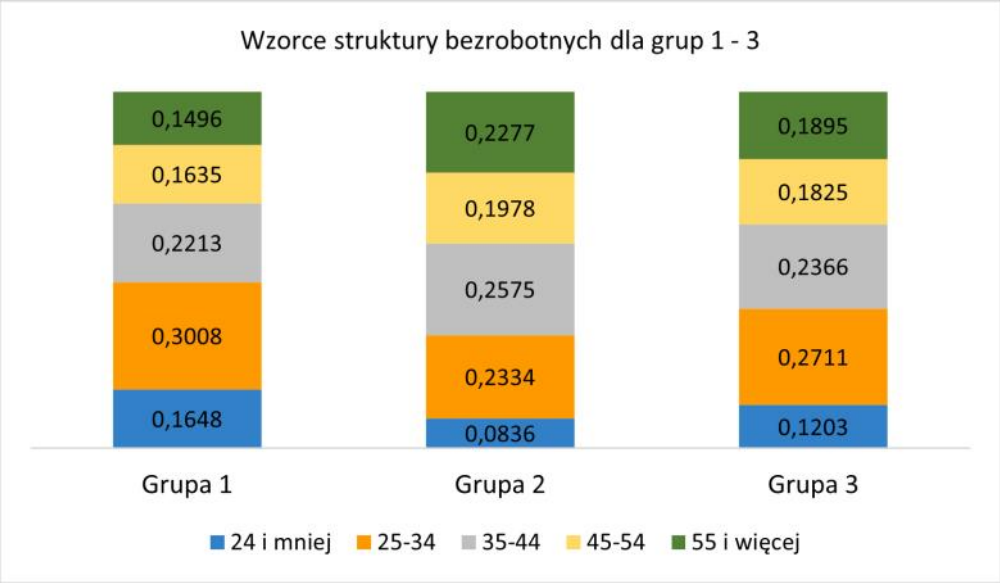


Rys. 5. Podział powiatów wg struktury wieku bezrobotnych
Źródło: opracowanie własne.

[8] A. Stanisz, Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 3. Analizy wielowymiarowe, Kraków, 2007, s. 120.

[9] Por. K. Kądziołka, Przestrzenne zróżnicowanie..., s. 231. Podobna koncepcja zastosowania miernika niepodobieństwa struktur w algorytmie grupowania hierarchicznego została wykorzystana w pracy M. Markowska i in., Podobieństwo struktur zatrudnienia w krajach Unii Europejskiej w latach 2008 – 2017 – ocena dynamiki, Prace Komisji Geografii Przemysłu Polskiego Towarzystwa Geograficznego, 33(4), 2019, s. 285, przy czym zamiast względnego wskaźnika podobieństwa struktur, wykorzystany został tzw. bezwzględny wskaźnik podobieństwa struktur, por. J. Zaród, Jakość produktów żywnościowych i stan sanitarny zakładów produkujących żywność, Stowarzyszenie Ekonomistów Rolnictwa i Agrobiznesu, XVI(4), 2014, s. 352.

[10] M. Walesiak, Pojęcie, klasyfikacja i wskaźniki podobieństwa struktur gospodarczych, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, 285, 1984, s. 68.



Rys. 6. Wzorce struktury dla poszczególnych grup
Źródło: opracowanie własne.

Powiaty grupy 1 charakteryzowały się przeciętnie najwyższym udziałem osób młodych w strukturze bezrobotnych i przeciętnie najniższym odsetkiem bezrobotnych w wieku 55 i więcej lat. Z kolei powiaty grupy 2 charakteryzowały się przeciętnie najniższym udziałem osób młodych w strukturze bezrobotnych i przeciętnie najwyższym odsetkiem osób w wieku 55 lat i więcej w strukturze zarejestrowanych bezrobotnych.

Propozycja niehierarchicznej metody identyfikacji grup obiektów podobnych pod względem struktury zjawisk

W niniejszej pracy przedstawiona zostanie również propozycja niehierarchicznej metody grupowania, celem identyfikacji obiektów podobnych pod względem struktury zjawisk. Metoda wzorowana będzie na algorytmie k-średnich, przy czym odległość między obiektami oparta jest na wskaźniku niepodobieństwa struktur. Proponowany algorytm grupowania obejmuje następujące kroki:

1) Ustal liczbę (ozn. k) grup, na które dzielone będą obiekty i wygeneruj k – środków (reprezentantów) grup.

2) Przypisz każdy z obiektów do najbliższego środka. Odległość między obiektami wyznaczana jest zgodnie ze wzorem: $d_{ij} = 1 - P_{ij}^{*}$, gdzie P_{ij}^{*} - wskaźnik podobieństwa struktur obiektów i oraz j.

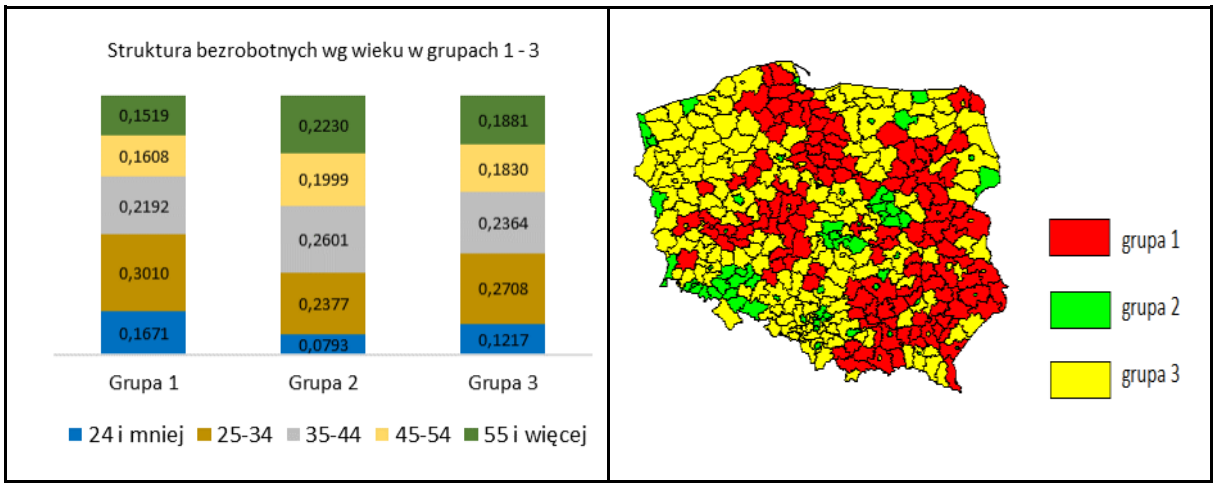
3) Wyznacz nowe środki grup, jako średnie poszczególnych składników struktury obiektów przypisanych do poszczególnych grup i przypisz obiekty do najbliższych mu środków.

[6] Do identyfikacji grup obiektów podobnych pod względem struktury badanych zjawisk wykorzystywana była m. in. metoda Warda z odległością euklidesową, por. A. Majdzińska, Regionalizacja demograficzna. Wybrane metody i ich aplikacje, Wydawnictwo Uniwersytetu Łódzkiego, 2016, s. 106. Jednakże do oceny podobieństwa struktur powszechnie wykorzystywany bywa wskaźnik podobieństwa struktur, stąd zastosowanie miary niepodobieństwa struktur obiektów, wydaje się być zasadne.

[7] E. Gatnar, Statystyczna analiza danych z wykorzystaniem programu R, Wydawnictwo PWN, Warszawa, 2009, s. 413.

4) Jeśli brak jest zmian w przypisaniu obiektów do grup, zakończ algorytm.

Opisaną powyżej procedurę zastosowano do analizowanego zbioru danych. Początkowe środki grup (utożsamiane tutaj z reprezentantami poszczególnych grup) zostały wyznaczone na podstawie przedstawionej wcześniej metody grupowania hierarchicznego. Rys. 7 przedstawia podział powiatów na grupy wg struktury wieku zarejestrowanych bezrobotnych i środki (reprezentantów) poszczególnych grup, po ich „dostrojeniu" opisaną powyżej metodą grupowania niehierarchicznego[11].



Rys. 7. Podział powiatów na grupy z wykorzystaniem połączenia obu metod
Źródło: opracowanie własne.

Podsumowanie

W pracy zaprezentowano możliwości aplikacyjne metod grupowania hierarchicznego i niehierarchicznego w zagadnieniu identyfikacji grup obiektów podobnych pod względem struktury analizowanego zjawiska. Do wyznaczania odległości między obiektami wykorzystany został wskaźnik niepodobieństwa struktur. W przypadku dużej liczby analizowanych obiektów, proste metody graficznej i tabelarycznej prezentacji danych są niewystarczające do oceny podobieństwa struktur. Wykorzystanie zaawansowanych metod analizy danych może być pomocne w takiej sytuacji, celem identyfikacji zależności w analizowanym zbiorze danych.

[11] Połączenie metody grupowania hierarchicznego z algorytmem grupowania niehierarchicznego może poprawić jakość grupowania, por. K. Kądziołka, Propozycja metody wspomagającej wybór miernika taksonomicznego na przykładzie oceny atrakcyjności giełd kryptowalut, Firma i Rynek, 1(59), 2021, s. 65-76.

Bibliografia

- Bajan B., Sowa K., Food Consumption Models around the World in the Context of Globalization, Intercathedra, 3(40), 2019.
- Chomątowski S., Sokołowski A., Taksonomia struktur, Przegląd Statystyczny, 2, 1978.
- Gatnar E., Statystyczna analiza danych z wykorzystaniem programu R, Wydawnictwo PWN, Warszawa, 2009.
- Kądziołka K., Przestrzenne zróżnicowanie, struktura i dynamika przestępczości stwierdzonej w Polsce, Przestrzeń, Ekonomia, Społeczeństwo, 8/II, 2015.
- Kądziołka K., Propozycja metody wspomagającej wybór miernika taksonomicznego na przykładzie oceny atrakcyjności giełd kryptowalut, Firma i Rynek, 1(59), 2021.
- Kukuła K., Z problematyki badań nad strukturą agrarną w Polsce w ujęciu przestrzennym, Oeconomia 6(4), 2007.
- Lisek S., Struktura wielkościowa przedsiębiorstw w Polsce, Metody Ilościowe w Badaniach Ekonomicznych, XVIII(4), 2017.
- Luty L., Regionalne zróżnicowanie struktury powierzchni użytków rolnych według systemów rolniczych w ujęciu dynamicznym, Metody Ilościowe w Badaniach Ekonomicznych, XVIII/2, 2017.
- Majdzińska A., Regionalizacja demograficzna. Wybrane metody i ich aplikacje, Wydawnictwo Uniwersytetu Łódzkiego, 2016.
- Markowska M., Strahl D., Sobczak E., Hlavacek P., Podobieństwo struktur zatrudnienia w krajach Unii Europejskiej w latach 2008 – 2017 – ocena dynamiki, Prace Komisji Geografii Przemysłu Polskiego Towarzystwa Geograficznego, 33(4), 2019.
- Sobczyk M., Statystyka opisowa, Wydawnictwo C.H. Beck, Warszawa, 2010.
- Sojka E., Analiza porównawcza struktur i procesów ludnościowych w wybranych krajach UE z wykorzystaniem metod taksonomicznych, Folia Oeconomica, 253, 2011.
- Sojka E., Zmiany w procesach i strukturach ludnościowych w wybranych krajach UE, Demograficzne uwarunkowania rozwoju gospodarczego (red.) A. Rączaszek, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, 2012.
- Stanisz A., Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 3. Analizy wielowymiarowe, Kraków, 2007.
- Walesiak M., Pojęcie, klasyfikacja i wskaźniki podobieństwa struktur gospodarczych, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, 285, 1984.
- Wawrzyniak K., Bąk I., Cheba K., Oesterreich M., The Similarity of European Union Countries in Terms of the Structure of the Unemployed, European Research Studies Journal, XXIII(4), 2020.
- Zaród J., Jakość produktów żywnościowych i stan sanitarny zakładów produkujących żywność, Stowarzyszenie Ekonomistów Rolnictwa i Agrobiznesu, XVI(4), 2014.
- GUS, Bank Danych Lokalnych: www.stat.gov.pl

Identification of groups of similar objects in terms of the structure of socio-economic phenomena on-the example of age structure of the unemployed

Summary:

The aim of the work is to present the application possibilities of clustering methods to identify groups of objects similar in terms of the structure of the analyzed phenomenon. The hierarchical clustering method was proposed, in which the structure dissimilarity indicator was used to determine the distance between the clusters. Then a proposal of the non-hierarchical clustering method was presented. Considerations were conducted on the example of the similarity of the age structure of the registered unemployed in poviats.

Keywords:

structure similarity, structure similarity index, hierarchical clustering, k-means method, age structure of the unemployed, unemployment