METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XV/4, 2014, str. 121 – 136

OCENA POTENCJAŁU PAŃSTW UNII EUROPEJSKIEJ DO GENEROWANIA INNOWACJI Z ZASTOSOWANIEM ANALIZY SKUPIEŃ

Elżbieta Roszko–Wójtowicz
Katedra Statystyki Ekonomicznej i Społecznej
Uniwersytet Łódzki
e-mail: eroszko33@gmail.com

Streszczenie: W ramach wielowymiarowej analizy porównawczej badacze najczęściej stają przed problemem dotyczącym wyboru odpowiedniego sposobu łączenia wielu obiektów w grupy, charakteryzujące się podobnymi właściwościami pod względem wybranych zmiennych diagnostycznych. Niniejsza praca stanowi propozycję zastosowania hierarchicznych metod aglomeracji do poszukiwania związków pomiędzy poszczególnymi państwami członkowskimi UE ze względu na możliwości generowania innowacji. W opisywanym przypadku najkorzystniejszą do poszukiwania związków pomiędzy obiektami, odnośnie zdefiniowanych w pracy zdolności do generowania innowacji, wydaje się być metoda pełnego połączenia i metoda Warda dające podział zestawu obiektów na trzy skupienia.

Słowa kluczowe: analiza skupień, potencjał innowacyjny, metody aglomeracyjne, kryterium Mojeny, kryterium Wisharta, kraje członkowskie UE

WSTĘP

Przedmiotem szczególnej uwagi ekonomistów na przestrzeni ostatniej dekady są czynniki kształtujące rozwój innowacyjny z punktu widzenia ich wpływu i znaczenia dla wzrostu gospodarczego. Co więcej, aktywność innowacyjna kraju nie jest działaniem autonomicznym, a zależy od wielu czynników związanych z obszarem edukacji, wiedzy oraz działalności badawczo-rozwojowej [Fagerberg 1988].

Tempo wzrostu innowacyjności UE spada, a opóźnienie w stosunku do światowych liderów (USA, Japonia, Korea Południowa) utrzymuje się.[1] W związku z tym strategia na rzecz wzrostu gospodarczego i zatrudnienia „Europa 2020" koncentruje się na stymulowaniu działań innowacyjnych i usuwaniu barier, które ograniczają ich wejście na rynek.[2]

Wsparcie innowacyjnego rozwoju regionów UE wymaga szczegółowego badania aktywności innowacyjnej poszczególnych gospodarek. Metodologia OECD [Podręcznik OSLO 2005] podaje, że badając procesy innowacyjne należy przede wszystkim zwrócić uwagę na aktywność innowacyjną przedsiębiorstw. Dla stymulowania innowacyjności jednak nie bez znaczenia pozostają również relacje z otoczeniem, zatem z innymi firmami i instytucjami otoczenia biznesu.

Czynniki, które mają największy wpływ na potencjał innowacyjny kraju związane są w głównej mierze z wykształceniem społeczeństwa, wyposażeniem w infrastrukturę teleinformatyczną, ogólną sytuacją na rynku pracy czy też aktywnością podmiotów w zakresie badań i rozwoju. W związku z tym do analizy państw Unii Europejskiej ze względu na możliwości do generowania innowacji wybrany został zestaw 15 zmiennych sklasyfikowanych w 4 obszary.

Pomiar innowacyjności obejmuje szeroką grupę wskaźników z zakresu nauki, techniki i innowacji (N + T + I) i charakteryzuje się następującymi własnościami [Markowska 2012]:

- Przedmiot pomiaru
  - Ludzie (specjaliści, badacze itp.)
  - Zasoby finansowe (źródła finansowania, rodzaje inwestycji)
  - Wiedza skodyfikowana (publikacje, patenty, kwalifikacje)
  - Wiedza ucieleśniona (urządzenia, komponenty, dobra trwałego użytku)
- Przestrzeń, w której dokonywany jest pomiar (organizacja, przestrzeń w sensie geograficznym)
- Typ działalności (dyscyplina, pole badawcze, dziedzina technologii, branża przemysłu itp.)
- Skala pomiaru (mikro – organizacja, mezo – dyscyplina, branża, gospodarka regionu, makro – gospodarka kraju i większe terytorium)
- Typ pomiaru (parametr zasobu – mierzy rozmiary lub parametr relacji – mierzy zależność lub przepływ)

Mając na uwadze wielowymiarowość zagadnienia celem głównym artykułu jest klasyfikacja państw Unii Europejskiej ze względu na czynniki kształtujące potencjał gospodarki do tworzenia innowacji, ze szczególnym uwzględnieniem pozycji Polski.

[1] http://europa.eu/rapid/press-release_IP-12-102_pl.htm
[2] http://ec.europa.eu/news/science/120208_pl.htm

Cel szczegółowy 1 – wybór, prezentacja i omówienie zmiennych pozwalających na ocenę innowacyjności gospodarek UE.

Cel szczegółowy 2 – omówienie wybranych metod grupowania obiektów oraz przedstawienie praktycznego ich zastosowania.

ANALIZA SKUPIEŃ – ZAŁOŻENIA TEORETYCZNE

Wielowymiarowa analiza porównawcza [Kisielińska, Stańko 2009] zajmuje się badaniem pewnego zbioru obiektów, które zostały opisane za pomocą wielu cech. WAP obejmuje szereg metod statystycznych dzięki którym możliwe jest dokonanie jednoczesnej analizy poziomu przynajmniej dwóch zmiennych opisujących każdą badaną jednostkę. Informacje o obiektach umieszczane są w macierzy obserwacji, której wiersze odpowiadają obiektom, zaś kolumny zebranym cechom.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

gdzie: *n* – liczba obiektów, *p* – liczba cech.

Wśród wielu metod wielowymiarowej analizy danych w badaniach ekonomiczno-społecznych szerokie zastosowanie znajdują:

- Bezwzorcowa klasyfikacja obiektów;
- Wzorcowa klasyfikacja obiektów;
- Liniowe porządkowanie obiektów;
- Regresja wieloraka;
- Analiza czynnikowa;
- Metoda składowych głównych.

Przedmiotem zainteresowania Autorki w niniejszym opracowaniu będzie zastosowanie analizy skupień do oceny państw członkowskich UE z punktu widzenia ich możliwości do generowania innowacji, co znalazło odzwierciedlenie w sformułowanych celach pracy.

Analiza skupień to jedna z najbardziej znanych metod eksploracji danych. Celem analizy skupień jest dokonanie grupowania obiektów w oparciu o pewien zestaw cech, które te obiekty charakteryzują. Istotą analizy skupień jest taki podział zbioru wyjściowego, aby jednostki jednej grupy były jak najbardziej do siebie podobne ze względu na wyróżniony zestaw zmiennych, a jednocześnie jak najbardziej odmienne od obiektów, które znajdują się w pozostałych skupieniach [Gatnar, Walesiak 2004]. Metoda nie dostarcza narzędzi do optymalnego doboru zmiennych opisujących badane jednostki. Wybór zmiennych i utworzenie

ostatecznego zbioru cech diagnostycznych spoczywa na badaczu i jest ważnym elementem analizy skupień gdyż wywiera istotny wpływ na finalne rozwiązanie.

- W literaturze przedmiotu istnieje ogromna liczba metod grupowania i zarazem mnogość propozycji odnoszących się do ich klasyfikacji. Ogół metod taksonomii numerycznej można podzielić na [Frątczak 2009]:
    - Metody hierarchiczne;
    - Metody optymalizacyjno-iteracyjne;
    - Metody obszarowe.

Przedmiotem zainteresowania Autorki w niniejszym opracowaniu są jedynie metody hierarchiczne, pozwalające na utworzenie pełnej hierarchii skupień, dla której współczynnik podobieństwa wzrasta monotonicznie. Skupienia powstające w kolejnych etapach (wyższego rzędu) zawierają w sobie rozłączne skupienia powstałe na etapach wcześniejszych (niższego rzędu). Struktura skupień tworzona jest w formie wykresu drzewkowego, zwanego dendrogramem. Wyróżnia się dwie grupy metod hierarchicznych [Grabiński, Wydymus, Zeliaś 1989]:

- procedury aglomeracyjne (sekwencyjne, indukcyjne) (ang. *agglomerative*)
- procedury deglomeracyjne (podziałowe) (ang. *divisive*)

Metody aglomeracyjne – początkowo przyjmuje się, że każdy z obiektów stanowi odrębną grupę, w kolejnych etapach łączy się w podzbiory podgrupy najbardziej do siebie podobne, aż do uzyskania finalnego rozwiązania w postaci jednego skupienia, które zawiera wszystkie elementy. W ten sposób powstaje uporządkowane zestawienie podziałów na segmenty. Wśród najczęściej stosowanych metod aglomeracyjnych znajdują się następujące metody: najbliższego sąsiedztwa, najdalszego sąsiedztwa, średnich połączeń oraz metoda Warda. Każdy z wymienionych algorytmów posiada odmienny sposób określania najmniejszej odległości pomiędzy łączonymi skupieniami.

W metodach podziałowych algorytm postępowania jest odwrotny do tego, który występuje w metodach aglomeracyjnych. Zatem punktem wyjścia jest zbiór obiektów, który traktowany jest jako jedna grupa, a w kolejnych krokach dzieli się jedno ze skupień na dwie części, a liczba skupień ulega zwiększeniu o jeden. W konsekwencji po n-1 krokach uzyskuje się n skupień jednoelementowych. Wyniki uzyskane za pomocą metod podziałowych mają analogiczną strukturę do tych uzyskanych za pomocą metod aglomeracyjnych.

Ogólny schemat postępowania w analizie skupień składa się z kilku następujących po sobie kroków. Początek stanowi dobór jednostek do badania, po którym następuje dobór cech diagnostycznych i ocena ich przydatności z punktu widzenia prowadzonej analizy z wykorzystaniem odpowiednich narzędzi statystycznych. Etapy końcowe to przeprowadzenie właściwej analizy skupień

w oparciu o wybrane algorytmy grupowania oraz ustalenie optymalnej liczby klas. [szerzej na ten temat w Milligan, Cooper 1987, Gordon 1999, Mikulec 2012].

WYBÓR ZMIENNYCH DIAGNOSTYCZNYCH

Obiektami poddanymi analizie w niniejszym opracowaniu będę poszczególne kraje członkowskie UE – 28 krajów, nazywane również jednostkami analizy. Cechy diagnostyczne natomiast, to zestaw 15 zmiennych wybranych z bazy Eurostat do opisu obiektów ze względu na ich zdolność do generowania innowacji. Początkowy zestaw składa się z następujących zmiennych diagnostycznych:

X1 – Wydatki rządkowe na działalność badawczo-rozwojową
X 2 – Wskaźnik zatrudnienia (grupa wieku 20-64 lata)
X 3 – Zatrudnienie w sektorze wysokich technologii i wiedzochłonnym
X 4 – Ludność z wykształceniem wyższym
X 5 – Poziom dostępu do Internetu w gospodarstwach domowych
X6 – Studenci w grupie wieku 15-24 lat
X 7 – Wydatki publiczne na edukację
X 8 – Patenty przyznane przez Biuro Patentów i Znaków Towarowych Stanów Zjednoczonych
X 9 – Zgłoszenia patentów do EUP
X 10 – Korzystanie z Internetu przez mieszkańców
X11 – Kompetencje w zakresie obsługi komputera
X 12 – Przedsiębiorstwa zatrudniające specjalistów ICT
X 13 – Wydatki przedsiębiorstw na B+R
X 14 – Wydatki przedsiębiorstw na B+R (% PKB)
X 15 – Kadra sektora B+R

W ocenie krajów członkowskich UE do generowania innowacji każdą z zaprezentowanych zmiennych diagnostycznych będziemy traktować jako stymulantę, oznacza to, że wzrost jej wartości ma pozytywny wpływ na badane zjawisko. W przypadku wystąpienia zmiennych określonych jako destymulanty, tzn. takich, których wzrost wartości ma niekorzystny wpływ na badane zjawisko, należy dokonać ich przekształcenia stosując odpowiednią formułę. Wybranych 15 zmiennych diagnostycznych podzielono na 4 obszary analizy. (Tabela 1)

Tabela 1. Obszary analizy pokryte wybranymi cechami diagnostycznymi – początkowy i ostateczny zestaw zmiennych diagnostycznych

Obszar	Początkowy zestaw zmiennych diagnostycznych	Ostateczny zestaw zmiennych diagnostycznych
1 – Finansowanie	X1 – Wydatki rządkowe na działalność badawczo-rozwojową X 7 – Wydatki publiczne na edukację X 14 – Wydatki przedsiębiorstw na B+R (%PKB) X 13 – Wydatki przedsiębiorstw na B+R	X1 – Wydatki rządkowe na działalność badawczo-rozwojową X 7 – Wydatki publiczne na edukację X 14 – Wydatki przedsiębiorstw na B+R (% PKB) ~~X 13 – Wydatki przedsiębiorstw na B+R~~
2 – Rynek pracy	X 2 – Wskaźnik zatrudnienia (grupa wieku 20-64 lata) X 3 – Zatrudnienie w sektorze wysokich technologii i wiedzochłonnym X 12 – Przedsiębiorstwa zatrudniające specjalistów ICT X 15 – Kadra sektora B+R	X 2 – Wskaźnik zatrudnienia (grupa wieku 20-64 lata) X 3 – Zatrudnienie w sektorze wysokich technologii i wiedzochłonnym X 12 – Przedsiębiorstwa zatrudniające specjalistów ICT ~~X 15 – Kadra sektora B+R~~
3 – Edukacja	X 4 – Ludność z wykształceniem wyższym X 5 – Poziom dostępu do Internetu w gospodarstwach domowych X6 – Studenci w grupie wieku 15-24 lata X 10 – Korzystanie z Internetu przez mieszkańców X11 – Kompetencje w zakresie obsługi komputera	X 4 – Ludność z wykształceniem wyższym ~~X 5 – Poziom dostępu do Internetu w gospodarstwach domowych~~ X6 – Studenci w grupie wieku 15-24 lata ~~X 10 – Korzystanie z Internetu przez mieszkańców~~ X11 – Kompetencje w zakresie obsługi komputera
4 – Patenty	X 8 – Patenty przyznane przez Biuro Patentów i Znaków Towarowych Stanów Zjednoczonych X 9 – Zgłoszenia patentów do EUP	~~X 8 – Patenty przyznane przez Biuro Patentów i Znaków Towarowych Stanów Zjednoczonych~~ X 9 – Zgłoszenia patentów do EUP

Źródło: opracowanie własne

Ostateczny wybór cech ze wstępnie zaprezentowanej listy zmiennych (Tabela 1) opisujących gospodarki unijne pod względem zdobności do tworzenia innowacji dokonany został na podstawie następujących kryteriów [por. Ostasiewicz 1998, Malina, Zieliaś 1996, Kunasz 2006]:

1. Kryterium uniwersalności – cechy diagnostyczne, opisujące badane zjawisko muszą być mierzalne, oraz powinny być źródłem istotnych informacji w obszarze poddawanym analizie.
2. Kryterium zmienności – analizowane cechy powinny dostatecznie różnicować badane obiekty, w tym celu dla analizowanych cech wyznacza się wartość współczynnika zmienności.
3. Kryterium stopnia skorelowania – oceny stopnia skorelowania zmiennych dokonuje się najczęściej za pomocą współczynnika korelacji liniowej Pearsona

lub współczynnika rang Spearmana. Wysoka wartość współczynnika korelacji świadczy o silnej zależności korelacyjnej między dwoma cechami diagnostycznymi i oznacza, że są one nośnikiem podobnych informacji.

4. Kryterium ważności – przyjmuje się, że cechy są ważne, jeśli trudno osiągają wysokie wartości. Celem sprawdzenia ważności cech i tym samym eliminacji cech nieważnych oblicza się wartości współczynnika asymetrii.

Weryfikacji kryterium zmienności dokonujemy na podstawie wyliczonych wartości współczynników zmienności dla poszczególnych cech diagnostycznych. Uzyskane wartości porównujemy z arbitralnie zadaną wartością krytyczną wynoszącą w tym przypadku V_S = 10%. Z początkowego zestawu zmiennych eliminacji podlegają te, dla których wyznaczone wartości współczynnika zmienności są mniejsze od przyjętej wartości progowej. W niniejszym przypadku współczynnik zmienności dla 14 spośród 15 analizowanych zmiennych jest wyższy niż 10%, co oznacza, iż wybrane cechy diagnostyczne mają dużą zdolność do różnicowania krajów członkowskich UE ze względu na potencjał do generowania innowacji. Wyjątek stanowi wartość wskaźnika zatrudnienia w grupie wieku 20-64 lata, gdzie współczynnik zmienności osiąga poziom 9,5%. Niemniej jednak, ze względu na przydatność merytoryczną tej zmiennej z punktu widzenia omawianej tematyki nie została ta cecha wyeliminowana ze zbioru cech diagnostycznych. Rozkład żadnej z wybranych zmiennych nie charakteryzuje się silną asymetrią lewostronną (wartość współczynnika asymetrii nie jest niższa od -0,7). Zatem, można przyjąć, iż każda z wybranych cech diagnostycznych dostatecznie różnicuje obiekty, nie osiągając tym samym wysokiego stopnia nasycenia.

Celem dokonania oceny stopnia skorelowania wybranego zestawu zmiennych najczęściej oblicza się współczynnik korelacji liniowej Pearsona lub współczynnik rang Spearmana. W niniejszym opracowaniu do oceny stopnia skorelowania zmiennych wybrano współczynnik korelacji liniowej Pearsona, przyjmując, że cechy charakteryzują się relatywnie dużą zmiennością, jeżeli współczynnik korelacji przyjmuje wartości wyższe niż 0,7. W niniejszym opracowaniu przyjęto arbitralnie zadaną wartość współczynnika korelacji i eliminacji cech dokonano na podstawie przesłanek merytorycznych. Niemniej jednak, oprócz arbitralnie przyjętej wartości współczynnika korelacji, w literaturze przedmiotu opisywane są również inne sposoby wskazania wartości progowej współczynnika korelacji, powyżej której dokonuje się eliminacji cech będących nośnikiem podobnych informacji, i tak w podejściu formalnym stosuje się najczęściej metodę minimaksową lub procedurę weryfikacji istotności współczynnika korelacji z populacji generalnej na podstawie próby [por. Grabiński, Wydymus, Zeliaś 1989, Panek 2009].

W analizie skupień wysoki stopień skorelowania zmiennych może powodować dominację tych wymiarów w tworzeniu klastrów i tym samym nieistotność pozostałych zmiennych. Dlatego też odrzucono zmienne, których korelacje z pozostałymi przekraczały wartość 0,7 [Nowak 1990]. Wyjątek tutaj

stanowi zależność korelacyjna między zmiennymi X9 i X14, r_{yx} = 0,86. Ze względu na duże znaczenie merytoryczne i przydatność społeczno-ekonomiczną tych zmiennych dla omawianej problematyki oraz fakt, iż zmienne należą do odrębnych obszarów analizy (X9 – obszar patenty, X14 – obszar finansowanie) cechy zostały uwzględnione w ostatecznym zbiorze. Na podstawie współczynników korelacji między zmiennymi i sum ich wartości bezwzględnych, wykluczone zostały następujące zmienne: X5, X8, X10, X13, X15. (zob. tab. 1)

Na podstawie przeprowadzonej analizy w zakresie doboru cech diagnostycznych można stwierdzić, iż w ostatecznym zestawie znalazły się zmienne, charakteryzujące się wysoką zdolnością do dyskryminacji jednostek w analizowanym obszarze. Oznacza to ich dużą zmienność w przestrzeni oraz niski stopień skorelowania między sobą. Tym samym usunięte zostały zmienne będące nośnikiem podobnych informacji.

ZAŁOŻENIA I SCHEMAT POSTĘPOWANIA EMPIRYCZNEGO

W literaturze przedmiotu opisanych jest wiele miar „podobieństwa" i „niepodobieństwa" obiektów oraz szeroka gama metod tworzenia skupień, o czym będzie mowa w tej części opracowania.

Empiryczna analiza danych została przeprowadzona w następującym układzie:

1. Liczba obiektów poddanych badaniu – 28 krajów członkowskich UE;
2. Liczba zmiennych diagnostycznych użytych do opisu potencjału innowacyjnego gospodarek UE – 10 wskaźników;
3. Normalizacja zmiennych;
4. Wybór miar odległości;
5. Wybór metod aglomeracyjnych.

Przebieg etapu 1 oraz 2, które zamieszczono powyżej, zostały już scharakteryzowane w poprzednim rozdziale. Kolejno, dla wybranego zestawu zmiennych diagnostycznych została przeprowadzona procedura normalizacji w formie unitaryzacji zerowej oraz standaryzacji. Ostatecznie jednak zdecydowano o przeprowadzeniu analizy skupień dla wybranych obiektów (N = 28 państw członkowskich) bez poddawania zmiennych procedurze normalizacji. Poddanie zmiennych procedurze normalizacji wprowadziło istotne zakłócenia w formułowaniu ostatecznych wniosków dotyczących podziału zbioru na klastry. Wartości indeksu sylwetkowego dla skupień uzyskanych za pomocą zmiennych poddanych procedurze normalizacji są znacznie niższe aniżeli w przypadku wyodrębnienia skupień na zmiennych rzeczywistych. Ze względu na mierzalny charakter zmiennych diagnostycznych analiza skupień została przeprowadzona dla dwóch miar odległości, tj. dla odległości euklidesowej oraz dla kwadratu odległości euklidesowej [StatSoft – Electronic Statistics Textbook].

W literaturze przedmiotu prezentowany jest szereg algorytmów analizy skupień dokonywanej przy pomocy metod aglomeracyjnych, wśród których w zależności od techniki grupowania rozróżnia się procedury kombinatoryczne i niekombinatoryczne. Poszczególne algorytmy różnią się sposobem wyznaczania odległości pomiędzy skupieniami. W niniejszym opracowaniu Autorka przeprowadziła grupowanie państw UE ze względu na ich zdolność do generowana innowacji stosując dziewięć algorytmów analizy skupień hierarchicznych metod aglomeracyjnych [por. Wishart 2006, Frątczak 2009]:

- Pojedynczego połączenia (najbliższego sąsiedztwa);
- Całkowitego połączenia (najdalszego sąsiedztwa);
- Średniego połączenia (średniej odległości między skupieniami);
- Ważona średniego połączenia (ważonej średniej odległości między skupieniami);
- Środka ciężkości;
- Minimalnej wariancji Warda;
- Sumy kwadratów;
- Średniego sąsiedztwa.

W artykule zaprezentowane zostały trzy spośród licznie dostępnych procedur wyboru liczby skupień, dwa kryteria Mojeny oraz kryterium Wisharta. Są to jedne z nielicznych procedur dedykowanych metodom klasyfikacji hierarchicznej stąd znalazły one zastosowanie empiryczne w niniejszym opracowaniu [Mikulec 2013]. (zob. Tabela 2)

Tabela 2. Kryteria wyboru wraz z opisem

KRYTERIUM	Formuła, przedział zmienności, kryterium wyboru
Mojeny I – Górnego obszaru odrzucenia (eng. *upper tail rule*)	$d_{i+1} > \bar{d} + k * S(d)$ KRYTERIUM WYBORU Klasyfikacja P_x, aby odpowiadający jej krok i: i = 2, 3, …, n – 1 pierwszy spełniał nierówność; d_i = długość i-tego wiązania (i-tej gałęzi drzewa) d_{i+1} = odległość połączenia grup w kroku i+1 $\bar{d}, S(d)$ – średni poziom (średnia arytmetyczna) i odchylenie standardowe długości wiązań (odległości połączenia grup) $k-$ stała, której wartość według R. Mojeny powinna zawierać się w przedziale [2,5; 3,5].

KRYTERIUM	Formuła, przedział zmienności, kryterium wyboru
Mojeny II – Średniej ruchomej (eng. *moving average quality control rule*)	$\alpha_{i+1} > \overline{\alpha}_i + L_i + b_i + k \cdot s_i$ KRYTERIUM WYBORU Klasyfikacja P_i, aby odpowiadający jej krok $i : i = y, y+1,..., n-2$ pierwszy spełniał nierówność $y -$ liczba wartości poziomu (odległości) połączenia klas α w danym kroku wykorzystana do wyznaczenia średniej ruchomej; $\overline{\alpha}_i$ – średnia ruchoma wartości parametru α obliczona w kroku i; $L_i -$ korekta dla opóźnionego „trendu" poziomu (odległości) połączenia klas obliczona w kroku i; $b_i -$ „ruchome" średniokwadratowe nachylenie linii trendu poziomu połączenia klas w kroku i; $s_i -$ „ruchome" odchylenie standardowe wartości parametru α (odległości).
Wisharta – Losowości podziału obiektów na wykresie drzewa (eng. *tree validation*)	Porównywanie wyników ciągu klasyfikacji uzyskanych metodami aglomeracyjnymi z rodziną drzew generowanych na podstawie losowej permutacji zbioru danych. KRYTERIUM WYBORU H_0 mówiąca o tym, że struktura grupowania obiektów w postaci danego drzewa jest losowa (brak struktury), $H_1 :\sim H_0$

Źródło: na podstawie [Mikulec 2013]

ANALIZA SKUPIEŃ – WYNIKI

Stosując 9 algorytmów aglomeracyjnych przy trzech wybranych kryteriach podziału zbioru na skupienia uzyskano 54 rozwiązania opisane poniżej. (zob. Tabela 3) Rozwiązania A wygenerowano dla kwadratu odległości euklidesowej a Rozwiązania B dla odległości euklidesowej. Każde rozwiązanie zostało poparte stosowną ilustracją graficzną w postaci dendrogramu. Mając na uwadze przesłanki natury merytorycznej i statystycznej Autorka zdecydowała o ostatecznym podziale zbioru na 3 skupienia. Po pierwsze, wyłonienie większej aniżeli 3 liczby skupień, nie dawało lepszych rezultatów, tzn. kolejne skupienia były wyłonione ze skupień najmniej licznych, a nie ze skupienia 15-elementowego, czyli najbardziej licznego w przypadku podziału na 3 grupy (sprawdzony szczegółowo został podział na 4 i 5 skupień). Po drugie, tworzenie większej liczby skupień nie poprawiało jakości grupowania mierzonej wartością indeksu sylwetkowego[3].

[3] Ocenie jakości grupowania wyrażonej za pomocą indeksu sylwetkowego oraz skorygowanego indeksu Randa będzie poświęcone odrębne opracowania.

Tabela 3. Rozwiązanie uzyskane dla wybranych algorytmów i kryteriów wyboru liczby skupień

Lp.	R algorytmu	Kryterium	Skupienia	Lp.	R algorytmu	Kryterium	Skupienia
Rozw_A_1	Ward	UT	3	Rozw_B_1	Ward	UT	3
Rozw_A_2	Ward	MA	6	Rozw_B_2	Ward	MA	5
Rozw_A_3	Ward	TREE	6	Rozw_B_3	Ward	TREE	3
Rozw_A_4	Pojedynczego połączenia	UT	4	Rozw_B_4	Pojedynczego połączenia	UT	7
Rozw_A_5	Pojedynczego połączenia	MA	4	Rozw_B_5	Pojedynczego połączenia	MA	2
Rozw_A_6	Pojedynczego połączenia	TREE	1	Rozw_B_6	Pojedynczego połączenia	TREE	2
Rozw_A_7	Pełnego połączenia	UT	3	Rozw_B_7	Pełnego połączenia	UT	5
Rozw_A_8	Pełnego połączenia	MA	5	Rozw_B_8	Pełnego połączenia	MA	3
Rozw_A_9	Pełnego połączenia	TREE	3	Rozw_B_9	Pełnego połączenia	TREE	4
Rozw_A_10	Średniego połączenia	UT	3	Rozw_B_10	Średniego połączenia	UT	5
Rozw_A_11	Średniego połączenia	MA	5	Rozw_B_11	Średniego połączenia	MA	5
Rozw_A_12	Średniego połączenia	TREE	3	Rozw_B_12	Średniego połączenia	TREE	3
Rozw_A_13	Średniego ważonego połączenia	UT	3	Rozw_B_13	Średniego ważonego połączenia	UT	5
Rozw_A_14	Średniego ważonego połączenia	MA	5	Rozw_B_14	Średniego ważonego połączenia	MA	2
Rozw_A_15	Średniego ważonego połączenia	TREE	2	Rozw_B_15	Średniego ważonego połączenia	TREE	2
Rozw_A_16	Środków ciężkości	UT	3	Rozw_B_16	Środków ciężkości	UT	4
Rozw_A_17	Środków ciężkości	MA	5	Rozw_B_17	Środków ciężkości	MA	4
Rozw_A_18	Środków ciężkości	TREE	5	Rozw_B_18	Środków ciężkości	TREE	5
Rozw_A_19	Mediany	UT	4	Rozw_B_19	Mediany	UT	4
Rozw_A_20	Mediany	MA	5	Rozw_B_20	Mediany	MA	4
Rozw_A_21	Mediany	TREE	4	Rozw_B_21	Mediany	TREE	4
Rozw_A_22	Sumy kwadratów	UT	3	Rozw_B_22	Sumy kwadratów	UT	3
Rozw_A_23	Sumy kwadratów	MA	5	Rozw_B_23	Sumy kwadratów	MA	3
Rozw_A_24	Sumy kwadratów	TREE	5	Rozw_B_24	Sumy kwadratów	TREE	5
Rozw_A_25	Średniego sąsiedztwa	UT	4	Rozw_B_25	Średniego sąsiedztwa	UT	6
Rozw_A_26	Średniego sąsiedztwa	MA	5	Rozw_B_26	Średniego sąsiedztwa	MA	2
Rozw_A_27	Średniego sąsiedztwa	TREE	3	Rozw_B_27	Średniego sąsiedztwa	TREE	3

Źródło: opracowanie własne na podstawie danych statystycznych z bazy Eurostat, obliczenia wykonano w Clastan Graphics

Bazując na wartości indeksu sylwetkowego (0,791) Autorka zdecydowała o ostatecznym podziale zbioru na 3 grupy. Identycznych rozwiązań dostarczyły

dwie metody tj., metoda Warda przy I kryterium Mojeny oraz metoda pełnego połączenia stosując I kryterium Mojeny oraz kryterium Wisharta.

Tabela 4. Przyporządkowanie krajów do poszczególnych grup

Numer klastra	Państwa
Klaster 1	Rumunia Bułgaria Grecja Chorwacja Słowacja Portugalia Republika Czeska Węgry Malta Litwa Łotwa Polska Cypr Hiszpania Estonia
Klaster 2	Włochy Słowenia Irlandia Wielka Brytania Luksemburg Belgia Francja
Klaster 3	Austria Niderlandy Dania Finlandia Szwecja Niemcy

Źródło: opracowanie własne na podstawie obliczeń wykonanych w ClustanGraphics

W klastrze 3, dla każdej ze zmiennych diagnostycznych, wartości średnich wewnątrzgrupowych są znacznie wyższe aniżeli średnie ogólne, jest to jednocześnie grupa najlepsza z punktu widzenia potencjału do generowania innowacji. (zob. tab. 5) W grupie tej znajdują się wszystkie kraje skandynawskie oraz Austria, Niemcy i Niderlandy. Skupienie trzecie jest również najbardziej jednorodne, współczynniki zmienności wewnątrzgrupowe stojące przy każdej ze zmiennych diagnostycznych są zdecydowanie niższe aniżeli wartości współczynników zmienności wyznaczone dla całej badanej zbiorowości. Największa różnica w poziomie zmienności występuje dla cechy patenty zgłoszone do Europejskiego Urzędu Patentowego. Tutaj współczynnik zmienności ogólny jest ponad siedmiokrotnie wyższy aniżeli wewnątrzgrupowy. W skupieniu trzecim znalazły się jednostki, dla których odnotowano najwyższą wartość, ze wszystkich możliwych wartości zarejestrowanych dla całej populacji, w odniesieniu do następujących cech: wskaźnik zatrudnienia, wydatki publiczne na edukację, zgłoszenia patentów do EUP, kompetencje w zakresie obsługi komputera oraz wydatki przedsiębiorstw na B+R. Analizując wartości minimalne dla poszczególnych zmiennych, należy stwierdzić, iż w niniejszym klastrze nie wystąpiły państwa, dla których zaobserwowano najniższą z możliwych wartości.

W przypadku grupy 1, do której należy również Polska, wraz ze wszystkimi państwami które dołączyły do UE nie wcześniej jak w 2004, średnie wewnątrzgrupowe dla każdej z 10 analizowanych cech są niższe aniżeli średnie ogólne, jest to jednocześnie najsłabsza grupa ze względu na potencjał do generowania innowacji. (zob. tab. 5) Grupa ta, na tle dwóch pozostałych, jest najmniej jednorodna. Potwierdzają to wewnątrzgrupowe wartości współczynników zmienności, które aż dla 7 spośród 10 zmiennych diagnostycznych są wyższe aniżeli wartości ogólne tej miary. Ponadto, poddając analizie wartości maksymalne i minimalne stojące przy wybranych zmiennych diagnostycznych, widzimy, że w skład skupienia pierwszego weszły kraje, w których dla wybranych cech odnotowano wartość najwyższą bądź najniższą spośród występujących w całej populacji. Taka sytuacja ma miejsce w odniesieniu do następujących cech diagnostycznych: wydatki rządowe na działalność badawczo-rozwojową, studenci w grupie wieku 15-24 lata, przedsiębiorstwa zatrudniające specjalistów ICT.

Polska w zakresie potencjału innowacyjnego ma poziom zbliżony do następujących krajów: Rumunia, Bułgaria, Grecja, Chorwacja, Słowacja, Portugalia, Republika Czeska, Węgry, Malta, Litwa, Łotwa, Cypr, Hiszpania, Estonia.

Obecnie (dane z roku 2012) wydaje się, że jedynie Słowenia ma szansę dorównać do poziomu państw byłej „piętnastki". Grupa 2 to grupa złożona w przeważającej mierze z krajów dawnej „15", które w znaczący sposób odczuły skutki kryzysu gospodarczego. (zob. tab. 5) Słowenia jest w tej grupie jedynym reprezentantem „nowych" krajów członkowskich. Porównując tę grupę z grupą pierwszą należy stwierdzić, iż jest ona bardziej jednorodna. Dla większości spośród badanych cech współczynniki zmienności wewnątrzgrupowe są niższe aniżeli współczynniki ogólne obliczone dla całej populacji. W niniejszym klastrze znalazły się państwa, dla których zaobserwowano najniższą wartość, ze wszystkich możliwych w populacji, jedynie w odniesieniu do dwóch zmiennych: ludność z wykształceniem wyższym oraz wydatki publiczne na edukację.

Tabela 5. Wybrane miary statystyczne dla poszczególnych klastrów

Klaster					Klaster				
	Ogółem	1	2	3		Ogółem	1	2	3
Wielkość	28	15	7	6	Wielkość	28	15	7	6
Wydatki rządkowe na działalność badawczo-rozwojową					Wydatki publiczne na edukację (% PKB)				
$\overline{X}$	1,25	1,08	1,22	1,74	$\overline{X}$	5,49	5,06	5,49	6,60
S_x	0,47	0,52	0,14	0,19	S_x	1,30	1,14	1,24	1,28
Min	0,4	0,4	1,09	1,54	Min	3,15	3,53	3,15	5,08
Max	2,12	2,12	1,49	2,02	Max	8,8	7,92	6,57	8,8
V_S	37,5%	48,2%	11,7%	10,8%	V_S	23,7%	22,5%	22,7%	19,5%
Wskaźnik zatrudnienia (grupa wieku 20-64 lata)					Zgłoszenia patentów do EUP				
$\overline{X}$	67,94	64,58	67,89	76,38	$\overline{X}$	79,04	13,57	91,53	228,14
S_x	6,42	5,19	4,47	1,85	S_x	88,48	12,70	26,77	34,74
Min	55,3	55,3	61	74	Min	1,53	1,53	63,75	193,95
Max	79,4	72,1	74,2	79,4	Max	272,25	44,25	132,55	272,25
V_S	9,4%	8,0%	6,6%	2,4%	V_S	111,9%	93,6%	29,2%	15,2%
Zatrudnienie w sektorze technologii i wiedzochłonnym					Kompetencje w zakresie obsługi komputera				
$\overline{X}$	3,95	3,32	4,70	4,63	$\overline{X}$	27,79	24,73	29,29	33,67
S_x	1,30	1,16	1,32	0,87	S_x	7,79	7,44	5,77	7,74
Min	2,2	2,2	3,3	3,5	Min	8	8	23	21
Max	7,5	6,1	7,5	5,7	Max	42	35	40	42
V_S	33,0%	34,8%	28,1%	18,8%	V_S	28,0%	30,1%	19,7%	23,0%
Ludnośc z wykształceniem wyższym (grupa wieku 30-34 lata)					Przedsiębiorstwa zatrudniające specjalistów ICT				
$\overline{X}$	36,41	32,41	42,31	39,53	$\overline{X}$	23,14	21,93	24,00	25,17
S_x	10,02	9,32	9,93	8,49	S_x	7,17	7,71	8,17	4,54
Min	21,7	21,8	21,7	26,3	Min	5	5	14	20
Max	51,1	49,9	51,1	47,9	Max	36	36	33	31
V_S	27,5%	28,8%	23,5%	21,5%	V_S	31,0%	35,2%	34,0%	18,0%

Tabela 5 cd.

Klaster					Klaster				
	Ogółem	1	2	3		Ogółem	1	2	3
Studenci w grupie wieku 15-24 lata					Wydatki przedsiębiorstw na B&R (% PKB)				
$\overline{X}$	61,66	60,38	60,36	66,37	$\overline{X}$	1,01	0,48	1,30	1,97
S_x	8,63	9,10	8,95	6,34	S_x	0,73	0,35	0,47	0,42
Min	41,3	41,3	48,5	55,5	Min	0,06	0,06	0,69	1,22
Max	74,8	74,8	72,4	72,7	Max	2,44	1,25	2,16	2,44
V_S	14,0%	15,1%	14,8%	9,6%	V_S	72,3%	72,6%	36,1%	21,5%

Źródło: opracowanie własne na podstawie obliczeń wykonanych w ClustanGraphics

PODSUMOWANIE I WNIOSKI

Wybrane wskaźniki do oceny potencjału innowacyjnego europejskich gospodarek wykazały, iż poziom zróżnicowania poszczególnych krajów w wybranych obszarach – finansowanie, B+R i nauka, edukacja, rynek pracy, patenty – jest bardzo duży. Wyróżnienie państw, które znacznie odbiegają od czołówki państw członkowskich pozwoli na efektywniejsze wdrażanie polityki unijnej poprzez tworzenia rozwiązań adekwatnych do poziomu zaawansowania innowacyjnego w każdym z analizowanych krajów. Wyróżnienie grup państw o podobnym potencjale jest również korzystne z punktu widzenia prowadzenia analiz porównawczych i tworzenia oraz implementowania wspólnych rozwiązań dla wybranych grup krajów.

Zaproponowany w analizie zestaw cech diagnostycznych zawierał 15 zmiennych, z którego po wstępnej weryfikacji merytorycznej i statystycznej zostało wyeliminowanych 5 zmiennych, ostatecznie zestaw 10 zmiennych diagnostycznych stanowił podstawę klasyfikacji państw. W zależności od algorytmu grupowania wyróżniony podział państw zawierał od 1 do 7 skupień.

W wyniku przeprowadzonej analizy skupień sprawdzono 54 rozwiązania stosując 9 algorytmów grupowania oraz 3 kryteria wyboru optymalnej liczby klas. Ostatecznie obliczony dla wszystkich 54 wariantów indeks sylwetkowy wskazał na podział zbioru na 3 klastry.

Najkorzystniejszą do poszukiwania związków pomiędzy obiektami, odnośnie zdefiniowanych w pracy zdolności do generowania innowacji, wydają się być metoda pełnego połączenia i metoda Warda dające podział zestawu obiektów na trzy skupienia. Skład grup uzyskany tymi metodami jest jednakowy, co wskazuje na możliwość wykorzystania tych metod w badaniach podobnego typu.

Polska na tle pozostałych krajów członkowskich, które dołączyły do struktur unijnych nie wcześniej niż w 2004 roku wypada nie najgorzej, niemniej jednak jej przynależność do najsłabszej z wyróżnionych grup potwierdza, iż nakłady jakie muszą jeszcze zostać poniesione na rozwój nauki, edukacji w tym sferę badawczo-rozwojową są znaczące. Dodatkowo analiza skupień pokazała, że Rumunia

najbardziej odstaje od pozostałych państw członkowskich UE i jednocześnie różni się najbardziej od Finlandii i Szwecji. Jednocześnie Finlandia, Dania i Szwecja wyróżniają się istotnie na tle państw UE-28. Przy czym to Dania i Finlandia są do siebie najbardziej podobne ze względu na wybrany zbiór danych.

Warto podkreślić, iż w analizie skupień badacz przed zaproponowaniem ostatecznego rozwiązania powinien przetestować wiele różnych możliwych grupowań, ponieważ zastosowanie różnych algorytmów grupowania oraz odmiennych kryteriów wyboru optymalnej liczby klastrów prowadzi do różnych rezultatów. Niemniej jednak w przypadku metody pełnego połączenia zarówno dla I kryterium Mojeny jak i kryterium Wisharta otrzymujemy te same wyniki.

BIBLIOGRAFIA

Fagerberg J. (1998) International Competitiveness, Economic Journal, vol. 98, No. 2, s. 102-104.

Frątczak E. (red.) (2009) Wielowymiarowa analiza statystyczna. Teoria – przykłady zastosowań z systemem SAS, Szkoła Główna Handlowa w Warszawie, Warszawa, s. 120-122.

Gatnar E., Walesiak M. (2004) Metody statystycznej analizy wielowymiarowej w badaniach marketingowych, Wydawnictwo Akademii Ekonomicznej, Wrocław, s. 310-317.

Gordon A. D. (1999) Classification, (2nd edition), Chapman & Hall/CRC, Boca Raton, s. 6-10.

Grabiński T., Wydymus S., Zeliaś A. (1989) Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych, PWN, Warszawa, s. 50.

Kisielińska J., Stańko S. (2009) Istota analizy wielowymiarowej, Roczniki Nauk Rolniczych, Seria G, T. 96, z. 2 , Warszawa, s. 64.

Kunasz M. (2006) Przykład zastosowania metod WAP do analizy procesów gospodarowania zasobami ludzkimi w przedsiębiorstwie, [w:] Kapitał ludzki w gospodarce opartej na wiedzy, D. Kopycińska (red.), Katedra Mikroekonomii Uniwersytetu Szczecińskiego, Szczecin, s. 131-139 http://mikroekonomia.net/system/publication_files/904/original/11.pdf?1315223800.

Malina A., Zieliaś A. (1996) Taksonomiczna analiza przestrzennego zróżnicowania jakości życia ludności w Polsce w 1994 r., [w:] Ekonometryczne modelowanie danych finansowo-księgowych, Nowak E., Urbaniak M. (red.), UMCS, Lublin, s. 85-89.

Markowska M. (2012) Dynamiczna taksonomia innowacyjności regionów, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław, s. 73.

Mikulec A. (2013) Kryterium Mojeny i Wisharta w analizie skupień – przypadek skupień o różnych macierzach kowariancji [w:] Taksonomia 20. Klasyfikacja i analiza danych – teoria i zastosowania, Jajuga K., Walesiak M. (red.), Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Nr 278, Wrocław, ISSN 1899-3192, s. 206-208.

Mikulec A. (2012)Analiza skupień z wykorzystaniem programu komputerowego Clustangraphics [w:] Rola informatyki w naukach ekonomicznych i społecznych. Innowacje i implikacje interdyscyplinarne, Zieliński Z. E. (red.), Kielce, ISBN 978-83-89274-75-5, Tom. II, s. 214-216.

Milligan G.W., Cooper M.C. (1987) Methodology review: clustering methods, Applied Psychological Measurement, Vol. 11, No 4, 1987, s. 329-331.

Nowak E. (1990) Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych, PWN, Warszawa, [za:] Wojnar J., Zastosowanie metod taksonomicznych do klasyfikacji państw Unii Europejskiej, Zeszyty Naukowe, Świętokrzyskie Centrum Edukacji na Odległość, s. 396, http://www.sceno.edu.pl/cms_tmp/2438_Jolanta%20Wojnar%20-%20III%20SCENO.pdf.

Ostasiewicz W. (red.) (1998) Statystyczne metody analizy danych, Akademia Ekonomiczna we Wrocławiu, Wrocław, s. 118.

Panek T. (2009) Statystyczne metody wielowymiarowej analizy porównawczej, Szkoła Główna Handlowa w Warszawie, Warszawa, s. 20-23.

Podręcznik OSLO (2005) Zasady gromadzenia i interpretacji danych dotyczących innowacji, OECD, Eurostat, Paryż, s. 20-23, http://www.uwm.edu.pl/ciitt/wp-content/uploads/2013/10/Podrecznik-OSLO-MANUAL1.pdf.

http://europa.eu/rapid/press-release_IP-12-102_pl.htm

http://ec.europa.eu/news/science/120208_pl.htm

COMPARATIVE STUDIES ON THE INNOVATIVE POTENTIAL OF THE EUROPEAN UNION MEMBER COUNTRIES BASED ON THE CLUSTER ANALYSIS

Abstract: Under multidimensional data analysis researchers face the problem of grouping methods selection. The aim of the paper is to find the inner homogenic groups created by 28 EU member countries in relation to the selected characteristics describing the potential of economies to create innovations. The issue of clustering objects – countries was conducted with the use of ClustanGraphics software. Ten variables were selected from the Eurostat database in order to describe the potential of the EU member countries to create innovations. These were tested using nine hierarchical clustering methods. The *Increase in Sum of Squares* method and the *Complete Linkage* method seem to be the most appropriate for further discussion regarding cross-group analysis. The rest of tested methods do not seem suitable for deep analysis between the groups. This is mainly due to the tendency to form rather unclear clusters with the structure of long chains.

Keywords: cluster analysis, agglomerative methods, Mojena criteria, Wishart criteria, innovative potential, EU member countries