2021

Analiza dbałości o środowisko naturalne w państwach członkowskich Unii Europejskiej

KAROL DOLIŃSKI

Spis treści

Wstęp

Postęp cywilizacyjny, definiowany jako pozytywna zmiana w społeczeństwie związana z opanowaniem przyrody, wynalazczością, nowymi technologiami czy poziomem materialno-duchowym społeczeństwa, nigdy nie był tak dynamiczny, jak ten obserwowany obecnie. Wzrost populacji oraz rewolucja technologiczna przyczyniły się do wzrostu zapotrzebowania na zasoby naturalne Ziemi, co przełożyło się na intensyfikację skutków eksploatacji planety. Następstwem jest wzrost zanieczyszczenia środowiska, klęski żywiołowe czy choroby cywilizacyjne. Dlatego też kwestia troski i dbałości o środowisko naturalne nigdy nie była tak żywotną kwestią całej cywilizacji jak obecnie.

Dbałość o środowisko naturalne może być zdefiniowana jako działania, mające minimalizować rozprzestrzenianie się (w długim okresie i znacznym stopniu) szkodliwych dla środowiska czynników (Hulme, 2010, strony 675-677). Czynności te mogą być wykonywane zarówno przez jednostkę jak i organizację. Mają polegać nie tylko na prewencji, ale również na wykazywaniu inicjatywy w poprawianiu obecnego stanu. W przypadku jednostki może być to zmiana nawyków, większa świadomość ekologiczna. Z kolei państwo, czyli organizacja, winno dostosowywać prawo, politykę i modyfikować długofalowe cele przez pryzmat ochrony środowiska naturalnego. Innymi słowy, dbałość o ekosystem to szereg działań mających na celu poprawę stanu rzeczy. Ocena tej dbałości powinna być recenzją stanu obecnego, w jakim stopniu zarówno jednostka jak i państwo przyczyniają się do minimalizacji szkodliwych dla środowiska naturalnego czynników.

Celem niniejszej pracy jest analiza państw członkowskich Unii Europejskiej ze względu na dbałość o środowisko. Wykorzystano analizę skupień, która umożliwia podział obiektów na klasy, dzięki czemu możliwe jest wskazanie państw do siebie podobnych i odwrotnie. Dalsze badanie w pracy, za pomocą porządkowania liniowego, ma pozwolić odpowiedzieć na pytanie, które państwa wypadają najlepiej, a które najgorzej ze względu na dbałość o ekosystem.

W pierwszym rozdziale zawarto przegląd literatury oraz wybór istotnych zmiennych, ze względu na problem dbałości o środowisko. W kolejnym zostały opisane wykorzystane metody analizy skupień i porządkowania liniowego. Finalnie, w rozdziale trzecim, przeprowadzono badanie empiryczne na podstawie wybranego zbioru zmiennych wraz z omówieniem i porównaniem wyników.

1. Problem ochrony środowiska

Rozdział pierwszy został poświęcony przybliżeniu problemu jakim jest dbałość o środowisko naturalne. Omówiono wyniki innych badań dotyczących ochrony ekosystemu. Skupiono się również na wyborze zmiennych, w oparciu o które zostanie przeprowadzona analiza i ocena dbałości o środowisko naturalne dla każdego państwa będącego obecnie członkiem Unii Europejskiej.

1.1. Przegląd literatury

Ochrona środowiska naturalnego jest terminem bardzo obszernym, dlatego też różnorodność literatury jest ogromna. Przeprowadzono szereg badań jak i powołano wiele instytucji, których zadaniem jest monitorowanie stanu środowiska naturalnego w danym państwie lub organizacji. Raporty tych instytucji często są nieocenioną wskazówką, jakie obszary należy poprawić, a także stanowią obiektywną literaturę, na podstawie której można dokonywać oceny stanu środowiska. Europejska Agencja Środowiska jest agencją Unii Europejskiej, a jej celem jest dostarczanie sprawdzonych, obiektywnych i rzetelnych informacji na temat ekosystemu. Ponadto ma monitorować poprawę stanu środowiska. W USA za te działania odpowiada Agencja Ochrony Środowiska EPA. Podobnych instytucji pod względem wykonywanych zadań jest znaczenie więcej, są powołane w prawie każdym państwie.

Ostatnimi laty ochrona środowiska widoczna na agendzie wielu rządów stała się jeszcze bardziej powszechnym tematem wielu artykułów. Większość z nich bada jedynie, acz kompleksowo, jeden obszar ochrony środowiska, jak jakość powietrza czy wód, a nie

skupia się na ocenie całokształtu. Całokształt najczęściej jest oceniany na podstawie wskaźników agencji zajmujących się środowiskiem.

Ze wszystkich wskaźników związanych ze środowiskiem najbardziej rozpoznawalnym jest wskaźnik wydajności środowiskowej EPI (ang. *Environmental Performance Index*) wyznaczany dla prawie wszystkich państw świata z wykorzystaniem ponad trzydziestu mniejszych wskaźników (takich jak na przykład jakość powietrza, trend w intensywności emisji, zalesienie). Pozwala on ocenić stan środowiska naturalnego i witalność ekosystemu. Im wyższa jest wartość EPI, tym stan środowiska w danym państwie jest lepszy. W 2020 roku największą wartość EPI spośród wszystkich krajów świata przypisano Danii. Było to spowodowane wysokiej jakości, długoletnimi programami ochrony środowiska. Dania została doceniona za dekarbonizację. Drugie miejsce zajął Luksemburg, a podium zamknęła Szwajcaria. W pierwszej dziesiątce, spośród państw Unii Europejskiej, znalazły się jeszcze Francja, Austria, Finlandia, Szwecja i Niemcy. Natomiast najmniejsze wartości EPI, wśród państw Unii, przyznano Polsce i Bułgarii. Przesądziły o tym głównie o wiele niższe wartości wskaźników dotyczących jakości powietrza i wody oraz gospodarowania odpadami w stosunku do innych państw Unii Europejskiej (Wendling i in., 2020, strony 20-29).

Kilka lat temu Monika Jaworska (2016) przeprowadziła analizę dystansu Polski do innych krajów Unii Europejskiej ze względu na ochronę środowiska naturalnego. Wykorzystała ona zmienne dotyczące emisji tlenków siarki, azotu, węgla oraz te mówiące o ilości wytwarzanych i składowanych odpadach na jednego mieszkańca. Zestaw zmiennych uzupełniał udział produkowanej energii ze źródeł odnawialnych w energii ogółem. Analiza nie została przeprowadzona z wykorzystaniem analizy skupień, ale zbudowano miarę syntetyczną – wskaźnik Perkala. Każdej zmiennej przypisano taką samą wagę. Badanie, w oparciu o dane z 2000, 2005 i 2013 roku wskazało, iż najlepszy – ze względu na wybrane zmienne – stan środowiska naturalnego cechował Łotwę, Chorwację i Portugalię. Polska uplasowała się na szóstym miejscu. Najgorszy wynik został odnotowany dla Malty, która była bardzo znacząco odróżniająca się od innych państw. Na wartość wskaźnika dla Malty mogły mieć wpływ uwarunkowania geograficzne i demograficzne, a dokładnie to, że jest to najmniejsze i najmniej ludne państwo Unii Europejskiej.

Kharitonova i Ulyankin (2020) przeprowadzili analizę skupień stu siedemdziesięciu trzech państw ze względu na cztery cechy: wskaźniki ekologicznej wydajności, jakość powietrza i wody pitnej oraz śmiertelność przypisywaną jakości środowiska. Podziału dokonano na pięć skupień, ale tylko w dwóch znalazły się państwa będące członkami Unii Europejskiej. W jednym skupieniu znalazły się: Austria, Niemcy, Belgia, Grecja, Dania, Irlandia, Hiszpania, Włochy, Cypr, Malta, Holandia, Portugalia, Finlandia, Francja, Szwecja oraz Estonia. Należy podkreślić, że było to skupienie, dla którego średnia z każdej cechy była najwyższa dla zmiennych typu stymulanty i najniższa dla zmiennych typu destymulanty. Innymi słowy, w tej klasie znalazły się państwa wypadające najlepiej pod względem badanych cech. W drugim najlepszym skupieniu znalazły się pozostałe państwa Unii Europejskiej. Warto podkreślić fakt, iż wszystkie państwa Unii zostały zaklasyfikowane do dwóch najlepszych skupień.

Kijewska i Bluszcz (2016) przeprowadziły analizę skupień ze względu na emisję gazów cieplarnianych przez państwa Unii Europejskiej. Wykorzystano cztery cechy: ilość emitowanego dwutlenku węgla, tlenku węgla, metanu i podtlenku azotu. Wszystkie zmienne zostały wykorzystane w przeliczeniu na jednego mieszkańca. Osobne dwa skupienia stworzyła Dania i Irlandia. W trzecim znalazły się: Rumunia, Portugalia, Hiszpania, Słowacja, Szwecja, Włochy, Litwa, Węgry, Chorwacja i Francja. Ostatnie, czwarte skupienie składało się z pozostałych państw Unii. We wnioskach podkreślono, iż państwem emitującym najwięcej gazów cieplarnianych na mieszkańca była Dania.

Frodyma (2017) przeprowadziła badanie, w którym skupiono się na ocenie wpływu udziału energii produkowanej ze źródeł odnawialnych na stan środowiska naturalnego w Unii Europejskiej. Autorka wskazała na wysoką korelację ujemną pomiędzy emisją zanieczyszczeń a produkcją zielonej energii. Wzrost wykorzystania odnawialnych źródeł energii ma prowadzić do zmniejszania emitowanych zanieczyszczeń. W artykule została również opisana zrealizowana analiza skupień ze względu na emisję zanieczyszczeń siarkowych, azotowych, amoniaku, gazów cieplarnianych i niemetanowych lotnych związków organicznych. Państwa Unii podzielono tylko na dwie grupy. Do pierwszej, mniej licznej, zaklasyfikowano Belgię, Niemcy, Luksemburg, Maltę i Holandię. Podkreślono, iż te kraje cechowały najniższe udziały energii produkowanej ze źródeł odnawialnych w stosunku do całościowej produkcji. Polska znalazła się w grupie drugiej,

której średnie wartości przyjętych zmiennych były niższe, a więc bardziej pożądane. Dane wykorzystane w badaniu pochodziły z 2013 roku.

Rok temu zostało przeprowadzone badanie zrównoważonego rozwoju energii odnawialnej w państwach Unii Europejskiej (Wang i Yang, 2020). Analizę oparto na szeregu zmiennych. Wykorzystano zmienne natury demograficznej jak dochód, rozmiar populacji, współczynnik urbanizacji oraz te związane z zanieczyszczeniem powietrza, emisją gazów cieplarniach. Analizę uzupełniło dobranie zmiennych takich jak podatki od energii czy średnie ceny energii elektrycznej. Badanie pozwoliło wyróżnić trzy grupy państw. Pierwsza, czyli Dania i Szwecja, zostały uznane za państwa najbardziej zrównoważone pod względem energii odnawialnej. Druga grupa, gdzie znalazły się państwa południowo-wschodniej części Europy oraz Finlandia, została określona jako ta gdzie postęp jest widoczny, natomiast te państwa nie mogą jeszcze równać się z liderami, czyli Danią i Szwecją. Grupa trzecia, sklasyfikowana jako słaba pod względem zrównoważonego rozwoju zielonej energii, to państwa wschodniej Europy, w tym Polska.

Analizie poddano również udział odpadów komunalnych poddanych recyklingowi, odzyskowi, przetwarzaniu na energię w krajach Unii Europejskiej (Ríos i Picazo-Tadeo, 2021). Została wykorzystana metoda Warda. W rezultacie podzielono zbiór na pięć klas. Osobne klasy stanowiły zarówno Niemcy jak i Szwecja. Do trzeciej zaklasyfikowano Belgię, Finlandię oraz Danię. W czwartej znalazła się Austria i Holandia, a w piątej wszystkie pozostałe państwa Unii Europejskiej. W podsumowaniu artykułu podkreślono najlepszy wynik Szwecji spośród wszystkich państw. Najbliżej osiągnieć Szwecji są państwa centralnej i północnej części Europy. Osobne skupienia tworzone przez państwa śródziemnomorskie i wschodniej Europy zostało ocenione jako najgorsze ze względu na wykorzystany zestaw cech.

Analiza wybranych artykułów oraz wskaźnika wydajności środowiskowej pozwala na wyciagnięcie wstępnych wniosków odnośnie stanu środowiska w poszczególnych krajach Unii Europejskiej lub jej regionach. Należy zaznaczyć, iż kraje skandynawskie wraz z Finlandią często tworzą jedno skupienie, które ponadto charakteryzuje się wysokimi wartościami zmiennych typu stymulanty i niskimi typu destymulanty. Innymi słowy jest to najlepsza klasa. Krajem, który często odstaje od innych, jest Malta. Powtarza

się wniosek, iż jest to państwo, które najtrudniej dopasować do któregoś ze skupień. Najwięcej jest państw przeciętnych, są to głównie kraje południowej i zachodniej części Europy. Natomiast w badaniach jako najgorsze państwa są klasyfikowane te ze wschodniej części Europy, również Polska.

1.2. Istotne zmienne

Wszystkie dane wykorzystane w pracy zostały pobrane ze strony Europejskiego Urzędu Statystycznego (Eurostat, 2019) i Europejskiej Agencji Ochrony Środowiska (EEA, 2019) w dniu 7 marca 2021 roku. Każda zmienna składa się z 27 obserwacji (za okres 2019 roku), które reprezentują 27 państw Unii Europejskiej (bez Wielkiej Brytanii).

Problem dbałości o środowisko jest niezwykle szeroki, istnieją setki czynności i miar, które można by klasyfikować jako te dotyczące dbałości o ekosystem. W niniejszej pracy do oceny dbałości państwa, a także jednostki, o stan ekosystemu wykorzystano siedem zmiennych (zob. Tabela 1), które zostały wybrane jako te najważniejsze, na podstawie wiedzy autora i szeregu publikacji oraz raportów.

Tabela 1. Wykorzystane zmienne

Zmienna	Nazwa zmiennej	Opis zmiennej	Jednostka
X_1	Ekologiczne użytki rolne	Udział całkowitej powierzchni użytków rolnych zajmowanych przez rolnictwo ekologiczne	%
X_2	Recykling odpadów komunalnych	Udział odpadów komunalnych poddanych recyklingowi	%
X_3	Oczyszczane ścieki bytowe	Udział ścieków bytowych poddanych oczyszczaniu	%
X_4	Energia odnawialna	Udział energii odnawialnej w produkcji energii ogółem	%
X_5	Wydatki na środowisko	Wydatki na ochronę środowiska w stosunku do PKB	%
X_6	Intensywność emisji	Intensywność emisji dwutlenku węgla	g CO_2 / kWh
X_7	Emisja gazów cieplarnianych	Emisja gazów cieplarnianych	t / osoba

Źródło: opracowanie własne

Udział całkowitej powierzchni użytków rolnych zajmowanych przez rolnictwo ekologiczne

Pod terminem rolnictwa ekologicznego zawierają się istniejące obszary uprawiane ekologicznie i obszary w trakcie konwersji. Rolnictwo ekologiczne, jako metoda produkcji, kładzie największy nacisk na ochronę środowiska, a w przypadku produkcji zwierzęcej – na ich dobrostan. Niezwykle ważne jest również ograniczenie użycia syntetycznych środków chemicznych, takich jak nawozy, pestycydy, dodatki i produkty medyczne (European Commission, 2013, strony 6-9). Rolnictwo ekologiczne przyczynia się do poprawy jakości wody, w związku z ograniczeniem nawożenia szkodliwymi substancjami. Wpływa na wzrost długoterminowej żyzności gleby i ogranicza emisję gazów cieplarnianych. Finalnie oddziałuje na konsumenta i jego zdrowie dostarczając mu produkty lepszej jakości (HDRA, 1998, str. 3).

Udział odpadów komunalnych poddanych recyklingowi

Recykling odpadów komunalnych obejmuje nie tylko recykling materiałowy, ale także kompostowanie i biodegradacje w celu pozyskania biogazu do produkcji energii. Zarządzanie odpadami komunalnymi generuje koszty rzędu kilkudziesięciu miliardów euro rocznie w całej wspólnocie. Unia Europejska od wielu lat dąży do możliwie najefektywniejszego zarządzania odpadami, ponieważ są one ogromną stratą zasobów i energii (European Commission, 2007, strony 3-4). W roku 2016 odnotowano wzrost współczynnika recyklingu odpadów komunalnych o 11% w stosunku do roku 2007 (European Commission, 2019, str. 98). Wysokość odsetka odpadów komunalnych poddanych recyklingowi jest miarą dbałości o środowisko nie tylko ze strony państwa, ale również obywatela i jego decyzji dotyczących recyklingu.

Udział ścieków bytowych poddanych oczyszczaniu

Ściekami bytowymi można nazwać te powstałe w gospodarstwach domowych czy instytucjach publicznych. Nieoczyszczone ścieki bytowe zawierają patogeny, substancje organiczne, co powoduje choroby i niszczenie ekosystemów. Zanieczyszczenie wody ogranicza możliwość ponownego jej wykorzystania, co jest szczególnie ważne w regionach bez wielkich zasobów słodkiej wody (World Health Organization i UN-Habitat, 2018, strony 10-15). Na ilość oczyszczanych ścieków ma wpływ nie tylko polityka państwa, ale również postawa obywateli, którzy świadomi istotności problemu, odprowadzają ścieki tak, by powołane do tego podmioty mogły je oczyścić we właściwy

sposób. Dlatego można stwierdzić, iż udział ścieków bytowych poddanych oczyszczaniu jest miarą dbałości o środowisko.

Udział energii odnawialnej w produkcji energii ogółem

Europejska Agencja Środowiska definiuje energię odnawialną jako taką, która została wyprodukowana z wykorzystaniem biomasy, geotermii, wiatru, wody lub słońca (ECOTEC Research & Consulting Limited, 2008, strony 1-2). Energia pochodząca ze źródeł odnawialnych może być doskonałą odpowiedzią na zmiany klimatu i globalne ocieplenie (Bilgen i in., 2008, strony 373-374). Wiodącym celem Unii Europejskiej w walce o czyste środowisko jest zwiększanie udziału energii odnawialnej w produkcji energii ogółem. Według prognozy udział energii odnawialnej w 2030 roku przekroczy 24% (European Commission, 2018, str. 39). Warto podkreślić, iż odsetek ten jest bardzo zróżnicowany w państwach Unii i często znajduje odzwierciedlenie w jakości powietrza.

Wydatki na ochronę środowiska w stosunku do PKB

Do wydatków na ochronę środowiska można zaliczyć nakłady inwestycyjne i koszty bieżące, jeśli są związane z redukcją zanieczyszczeń, ochroną środowiska lub naprawą szkód środowiskowych. Do takich wydatków nie zalicza się tych, które korzystanie wpływają na stan środowiska, ale których głównym celem nie jest poprawa jakości ekosystemu. Nakładem inwestycyjnym są też wydatki gospodarstw domowych na ochronę środowiska (Główny Urząd Statystyczny, 2019, str. 10). Wydatki na ochronę środowiska w stosunku do PKB stanowią wartość niemianowaną, która odzwierciedla jak bardzo państwo i obywatele są zaangażowani w poprawę i ochronę środowiska. Nakłady inwestycyjne dotyczą między innymi termomodernizacji, wymiany pieców grzewczych, zakładania paneli fotowoltaicznych, a także maja wpływ na wysokość i ilość dotacji oferowanych przez państwo.

Intensywność emisji dwutlenku węgla

Intensywność emisji dwutlenku węgla oblicza się jako stosunek emisji dwutlenku węgla (w gramach) z publicznej produkcji energii elektrycznej do produkcji energii elektrycznej ogółem (wyrażonej w kilowatogodzinach). Ponad 40% światowej emisji tlenku węgla (IV) pochodzi z produkcji energii elektrycznej (International Energy Agency, 2019, str. 11). Jest to spowodowane tym, że znaczna część energii na świecie (prawie 70%) jest produkowana z wykorzystaniem paliw kopalnych. Warto również podkreślić

światowy trend w produkcji energii, który nieprzerwanie rośnie (Bin i Ang, 2016, str. 56). Intensywność emisji dwutlenku węgla może być miarą dbałości państwa o środowisko ze względu na dwa powody. Pierwszy z nich jest bezpośrednio skorelowany z ilością energii produkowanej z odnawialnych źródeł energii, gdzie produktem ubocznym nie jest dwutlenek węgla. Zaś drugim jest jakość i modernizacja elektrowni wykorzystujących paliwa kopalne (Dołęga, 2016, strony 2-5).

Emisja gazów cieplarnianych

Zgodnie z protokołem z Kioto, gazy cieplarniane to takie składniki atmosfery, które składają się z dwutlenku węgla, metanu, podtlenku azotu lub gazów fluorowanych (Kijewska i Bluszcz, 2016, str. 133). Powstają na skutek spalania paliw kopalnych, wylesiania. Ogromne ilości są wytwarzane na skutek hodowli zwierząt gospodarskich czy stosowania nawozów azotowych. Emisja gazów cieplarniach jest przyczyną globalnego ocieplenia klimatu, na skutek którego można zaobserwować topnienie pokrywy lodowej na biegunach, częstsze powodzie w rejonach nadmorskich, wyginięcie wielu gatunków roślin i zwierząt (Shahzad, 2015, strony 2-3). Reakcją na to zjawisko jest między innymi Europejski Zielony Ład, czyli program Unii Europejskiej, którego założeniem jest osiągnięcie neutralności klimatycznej przez państwa członkowskie do 2050 roku. Redukcja emisji gazów cieplarnianych jest więc przede wszystkim miarą dbałości państwa o stan środowiska, ale również jednostki, która codziennymi decyzjami może przybliżać państwo do realizacji celu, jakim jest neutralność klimatyczna do 2050 roku.

2. Wykorzystane metody

Rozdział drugi został poświęcony omówieniu wykorzystanych metod. Podstawowym narzędziem, które zostało zastosowane do analizy problemu była analiza skupień, która pozwala na rozwiązywanie problemów związanych z zagadnieniem klasyfikacji obiektów. Uzupełnieniem analizy skupień jest porządkowanie liniowe, które umożliwia stworzenie rankingu ze względu na wybrane cechy.

2.1. Analiza skupień

Uczenie nienadzorowane jest rodzajem uczenia maszynowego, które wykorzystuje się w sytuacji, kiedy celem analizy nie jest prognoza zmiennej zależnej, ale znalezienie zależności pomiędzy cechami (James i in., 2013, str. 322).

Metoda analizy skupień (zwanej też grupowaniem lub klasteryzacją) umożliwia pogrupowanie *n* obiektów na *u* grup (które są niepuste, rozłączne i zupełne) na podstawie wektora *p* cech. Sklasyfikowane obiekty powinny być jak najbardziej podobne do innych ze swojej grupy, będąc tym samym najbardziej różne od obiektów z innych grup (Krzyśko i in., 2008, str. 345). Milligan (1996, strony 342-343) wyodrębnił następujące etapy analizy skupień:

1. Wybór obiektów i charakteryzujących ich zmiennych oraz formuły normalizacji.
2. Wybór miary odległości pomiędzy cechami.
3. Wybór metody klasyfikacji.
4. Wyznaczenie liczby klas.
5. Klasyfikacja obiektów, ocena wyników wraz z interpretacją.

2.1.1. Wybór zmiennych, formuły normalizacji, miar odległości

Wybór m zmiennych (gdzie m to liczba zmiennych), które zostaną wykorzystane do klasyfikacji n obiektów (gdzie n to liczba obiektów, obserwacji), bardzo często jest dokonywany na podstawie wiedzy eksperckiej i literatury poświęconej danemu zagadnieniu. Mowa wtedy o wyborze merytorycznym. Natomiast wybór formalny to zbadanie zmienności oraz zależności pomiędzy zmiennymi (Walesiak, 2004, str. 54). Współczynnik zmienności dla każdej zmiennej wyznacza się ze wzoru:

$$V = \frac{s}{\bar{x}} \; ; \; \bar{x} \neq 0, \tag{1}$$

gdzie:

s – odchylenie standardowe z próby,

$\bar{x}$ – średnia arytmetyczna z próby.

Natomiast zależność pomiędzy zmiennymi można zbadać z wykorzystaniem współczynnika korelacji Pearsona (Jóźwiak i Podgórski, 2012, str. 118), wykorzystując wzór:

$$\rho = \frac{cov(X,Y)}{D(X)D(Y)}, \tag{2}$$

gdzie:

$cov(X,Y)$ – kowariancja zmiennych,

$D(X), D(Y)$ – odchylenie standardowe.

Przyjmuje się, iż wartość współczynnika zmienności (1) powinna być większa niż 0,1, natomiast współczynnika korelacji Pearsona (2) – na moduł mniejsza niż 0,9.

Normalizację zmiennych stosuje się w celu ujednolicenia ich wielkości (Walesiak, 2004, str. 55). Najczęściej wykorzystywaną metodą jest standaryzacja klasyczna opisana wzorem:

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j}, \tag{3}$$

gdzie:

x_{ij} – obserwacja *j*-tej zmiennej *i*-tego obiektu,

$\overline{x_j}$ – średnia arytmetyczna wszystkich obserwacji *j*-tej zmiennej,

s_j – odchylenie standardowe wszystkich obserwacji *j*-tej zmiennej.

Standaryzacja (3) jest wykorzystywana dla zmiennych na skali ilorazowej lub przedziałowej.

Istnieje szereg różnych miar odległości, a ich wybór jest uzależniony od skali pomiaru zmiennych. Dwiema podstawowymi miarami dla zmiennych na skali przedziałowej lub ilorazowej są odległość miejska (4) i euklidesowa (5) wyrażone wzorami (Walesiak, 2011, strony 23-24):

$$\sum_{j=1}^{m} \left| z_{ij} - z_{kj} \right|, \tag{4}$$

$$\sqrt{\sum_{j=1}^{m} \left(z_{ij} - z_{kj} \right)^2}, \tag{5}$$

gdzie:

z_{ij}, z_{kj} – znormalizowana wartość *j*-tej zmiennej dla *i*-tego (*k*-tego) obiektu,

j – numer zmiennej,

m – liczba zmiennych.

2.1.2. Metody klasyfikacji

Można wyodrębnić kilka podstawowych metod analizy skupień, do których zaliczają się metody hierarchiczne, a także metody optymalizujące, w których to jednak niezbędna jest wcześniejsza znajomość liczby klas, na które zostanie podzielony zbiór obiektów (Gatnar i Walesiak, 2009, str. 413). Jako podstawową metodę optymalizującą można sklasyfikować metodę *k*-średnich.

Metody hierarchiczne

Metody hierarchiczne dzieli się na aglomeracyjne i deglomeracyjne, natomiast znacznie częściej wykorzystuje się te pierwsze. Istota metod hierarchicznych sprowadza się do sukcesywnego dzielenia lub łączenia obserwacji. Rezultatem takiego działania jest dendrogram, czyli drzewo-podobna struktura (Kłopotek i Wierzchoń, 2015, str. 34).

Cechą charakterystyczną hierarchicznej klasyfikacji aglomeracyjnej jest sytuacja, w której każdy obiekt A_i poddawany analizie ($i = 1, \dots, n$; gdzie n jest liczbą obiektów) tworzy na początku oddzielną klasę (Gatnar i Walesiak, 2009, str. 413). Następnie algorytm można przedstawić w trzech krokach:

1. Należy znaleźć dwie najbardziej podobne do siebie klasy (na podstawie macierzy odległości) i połączyć je w jedno skupienie. Liczba klas po połączeniu zmniejszy się o jeden.
2. Następnie właściwe jest zaktualizowanie macierzy odległości pomiędzy nowo utworzoną klasą a pozostałymi klasami.
3. Kroki 1 i 2 należy powtarzać do momentu, aż wszystkie obiekty znajdą się w jednej klasie.

Gatnar i Walesiak (2009, strony 413-415) podkreślają, iż różnice w kontekście macierzy odległości wykorzystywanej w etapie drugim wynikają z istnienia różnych metod wyznaczania odległości miedzyklasowych.

Odległość międzyklasową można obliczyć na podstawie wzoru (Gatnar i Walesiak, 2009, str. 413):

$$d(P_i \cup P_k, P_l) = \alpha_i d(P_i, P_l) + \alpha_k d(P_k, P_l) + \beta d(P_i, P_k) + \gamma |d(P_i, P_l - d(P_k, P_l)|, \quad (6)$$

gdzie:

P – dana klasa,

$d(P_i \cup P_k, P_l)$ – odległość pomiędzy połączonymi klasami $P_i \cup P_k$ i inną klasą P_l,

$\theta = (\alpha_i, \alpha_k, \beta, \gamma)$ – wektor parametrów (uzależniony od metody klasyfikacji).

Metoda pojedynczego połączenia zwana też metodą najbliższego sąsiada (ang. *single linkage*) sprowadza się do liczenia odległości między klasami jako odległości pomiędzy dwoma najbliższymi obiektami (które należą do różnych klas). Wyraża się ją wzorem (6) z wykorzystaniem wektora parametrów (Gatnar i Walesiak, 2009, str. 414):

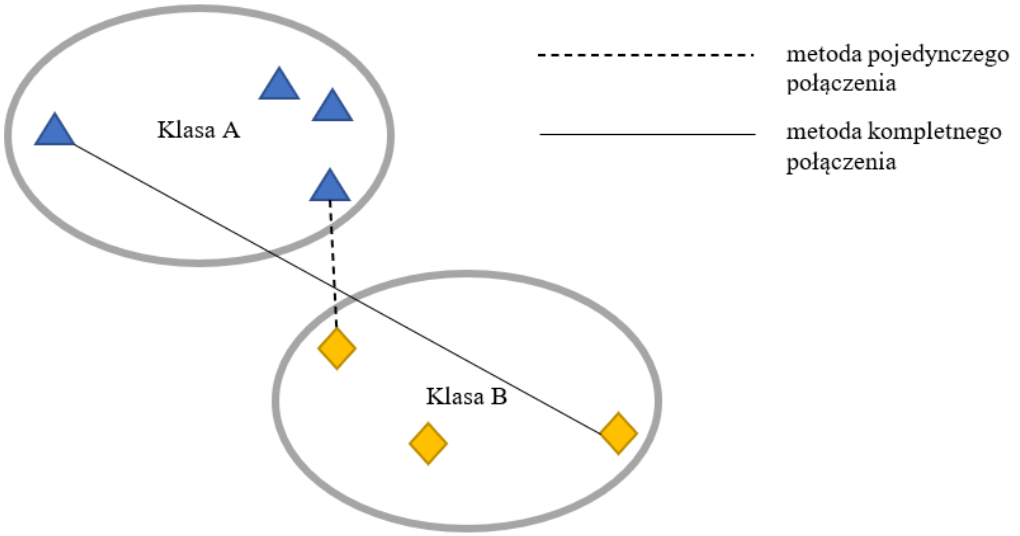
$$\theta = (\alpha_i, \alpha_k, \beta, \gamma) = \left(\frac{1}{2}, \frac{1}{2}, 0, -\frac{1}{2}\right). \quad (7)$$

Cechą charakterystyczną tej metody jest tworzenie łańcuchów (Kłopotek i Wierzchoń, 2015, str. 35). Gatnar i Walesiak (2009, str. 414) podkreślają, iż konsekwencją tworzenia się łańcucha może być powstawanie klas mało się siebie podobnych. Jest to spowodowane tym, że w tej metodzie o wiele częściej do istniejących klas dołączane są nowe obiekty, niż tworzone są nowe klasy. Efekt łańcucha występuje również w innych metodach, natomiast najsilniej jest widoczny w metodzie pojedynczego połączenia.

Metoda pełnego wiązania lub najdalszego sąsiedztwa (ang. *complete linkage*) sprowadza się natomiast do liczenia odległości między klasami jako odległości pomiędzy dwoma najdalszymi obiektami (które należą do różnych klas). Wyraża się ją wzorem (6) z wykorzystaniem wektora parametrów (Gatnar i Walesiak, 2009, str. 414):

$$\theta = (\alpha_i, \alpha_k, \beta, \gamma) = \left(\frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2}\right). \tag{8}$$

W tej metodzie nie występuje efekt łańcucha. Stosuje się ją, kiedy obiekty tworzą zwarte klasy, które są dobrze separowalne (Kłopotek i Wierzchoń, 2015, str. 35). Ilustracją działania metod opisanych wzorami (7) i (8) jest Rysunek 1. Pokazuje on sposób wyznaczania odległości międzyklasowych w przestrzeni dwuwymiarowej.



Rysunek 1. Wyznaczanie odległości międzyklasowej (w przestrzeni dwuwymiarowej) dla metody pojedynczego i kompletnego połączenia dla przykładowych klas A i B
Źródło: opracowanie własne

Niezwykle popularną metodą aglomeracyjną jest metoda Warda. Jej cechą charakterystyczną jest tworzenie skupień o podobnej liczebności. Wyraża się ją za pomocą wzoru (6) z wektorem parametrów (Gatnar i Walesiak, 2009, str. 414):

$$\theta = (\alpha_i, \alpha_k, \beta, \gamma) = \left(\frac{n_i + n_l}{n_i + n_k + n_l}, \frac{n_k + n_l}{n_i + n_k + n_l}, \frac{-n_l}{n_i + n_k + n_l}, 0\right), \tag{9}$$

gdzie:

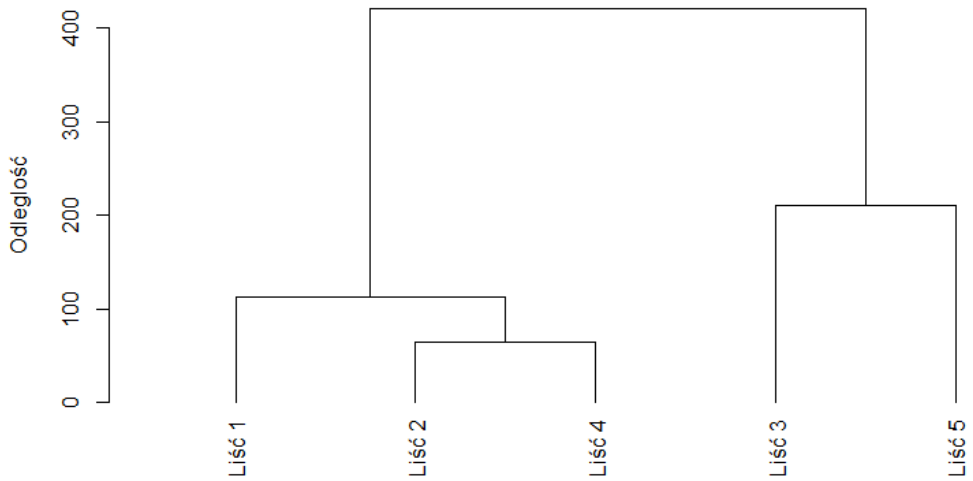
n_i, n_k, n_l – liczba obiektów w klasie (odpowiednio) P_i, P_k, P_l.

Według wielu analiz, które przeprowadził Milligan (1996, str. 358), metoda Warda jest jedną z dwóch najlepszych wśród hierarchicznych metod aglomeracyjnych. Natomiast

metoda pojedynczego połączenia wypadła najgorzej spośród wszystkich kilku badanych metod. Podkreślono jej niskie wyniki w każdym analizowanym przypadku. Z kolei metoda kompletnego połączenia uplasowała się o wiele lepiej, w wielu przypadkach będąc tylko nieznacznie gorsza od metody Warda.

Hierarchiczne metody deglomeracyjne nazywane są również klasyfikacją przez podział. Algorytm jest niejako odwrotny do tego dla metod aglomeracyjnych, czyli podział rozpoczyna się od sytuacji, w której wszystkie obiekty $A_1, \ldots, A_n$ (gdzie n jest liczbą obiektów) są przyporządkowane do jednego skupienia. Algorytm polega na zwiększaniu liczby klas o jedną w każdej iteracji, poprzez rozdzielnie jednej z istniejących już klas. Algorytm należy zakończyć w momencie, kiedy każda klasa zawiera tylko jeden obiekt (Gatnar i Walesiak, 2009, str. 415). Zasadniczą wadą metod deglomeracyjnych jest o wiele większa złożoność obliczeniowa niż w przypadku metod aglomeracyjnych. Koronacki i Ćwik (2005, str. 271) wskazują, iż trudno jest podać jakiekolwiek korzyści ze stosowania klasyfikacji przez podział, dlatego też w niniejszej pracy nie wykorzystano żadnych metod opierających się na tym właśnie sposobie klasyfikacji.

Graficznym odzwierciedleniem metod hierarchicznych jest dendrogram (zob. Rysunek 2), czyli binarne drzewo. Skupienia są reprezentowane przez węzły dendrogramu, a obiekty przez liście.



Rysunek 2. Dendrogram
Źródło: opracowanie własne

Metody optymalizujące

Jednym z problemów analizy skupień jest liczba klasyfikowanych obiektów, a co za tym idzie liczba możliwych podziałów na klasy. Do wyznaczenia liczby wszystkich podziałów zbioru n obiektów na u klas wykorzystuje się wzór (Gordon, 1999, str. 40):

$$L(n,u) = \frac{1}{u!}\sum_{r=1}^{u}(-1)^{u-r}\binom{u}{r}r^n, \tag{10}$$

gdzie:

$r = 1, \dots, u$ – numer klasy.

Aby wyznaczyć liczbę wszystkich możliwych podziałów zbioru n obiektów stosuje się wzór (Gatnar i Walesiak, 2009, str. 407):

$$L(n) = L(n,1) + L(n,2) + L(n,3) + \cdots + L(n,n). \tag{11}$$

Liczba możliwych podziałów uzyskana za pomocą wzorów (10) i (11) jest ogromna. Przy większej liczbie obiektów wręcz niemożliwe (lub bardzo złożone obliczeniowo) jest znalezienie optimum globalnego. Metody optymalizujące pozwalają wyznaczyć optimum lokalne znacznie ograniczając złożoność obliczeniową (Gatnar i Walesiak, 2009, str. 416).

Niech $H = [h_{ij}]$ będzie macierzą, której elementy świadczą o przynależności *i*-tego obiektu do *j*-tej klasy. Jeżeli *i*-ty obiekt należy do *j*-tej klasy to $h_{ij} = 1$ (w przeciwnym wypadku $h_{ij} = 0$). Środek ciężkości zbioru n obiektów można wyrazić wzorem (Kłopotek i Wierzchoń, 2015, str. 41):

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{12}$$

gdzie:

n – liczba obiektów,

x_i – wektor cech *i*-tego obiektu.

Środek ciężkości (centroid) *j*-tej klasy można zdefiniować za pomocą wzoru (Ćwik i Koronacki, 2005, str. 265):

$$\mu_j = \frac{1}{\sum_{i=1}^{n} h_{ij}}\sum_{i=1}^{n} h_{ij}x_i. \tag{13}$$

Macierz kowariancji wewnątrzgrupowych definiuje się wzorem (Kłopotek i Wierzchoń, 2015, str. 41):

$$W = \sum_{i=1}^{n} \sum_{j=1}^{u} h_{ij}(x_i - \mu_j)(x_i - \mu_j)^T, \tag{14}$$

natomiast macierz kowariancji międzygrupowych można przedstawić za pomocą wzoru:

$$B = \sum_{j=1}^{u} \left(\sum_{i=1}^{n} h_{ij} \right) (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T. \tag{15}$$

Najczęściej wykorzystywaną metodą optymalizującą jest metoda *k*-średnich. W tej metodzie reprezentacją każdej klasy jest centroid (czyli środek ciężkości). Celem metody *k*-średnich jest znalezienie minimum wartości miary (Gatnar i Walesiak, 2009, str. 416):

$$tr(W), \tag{16}$$

gdzie:

tr – ślad macierzy,

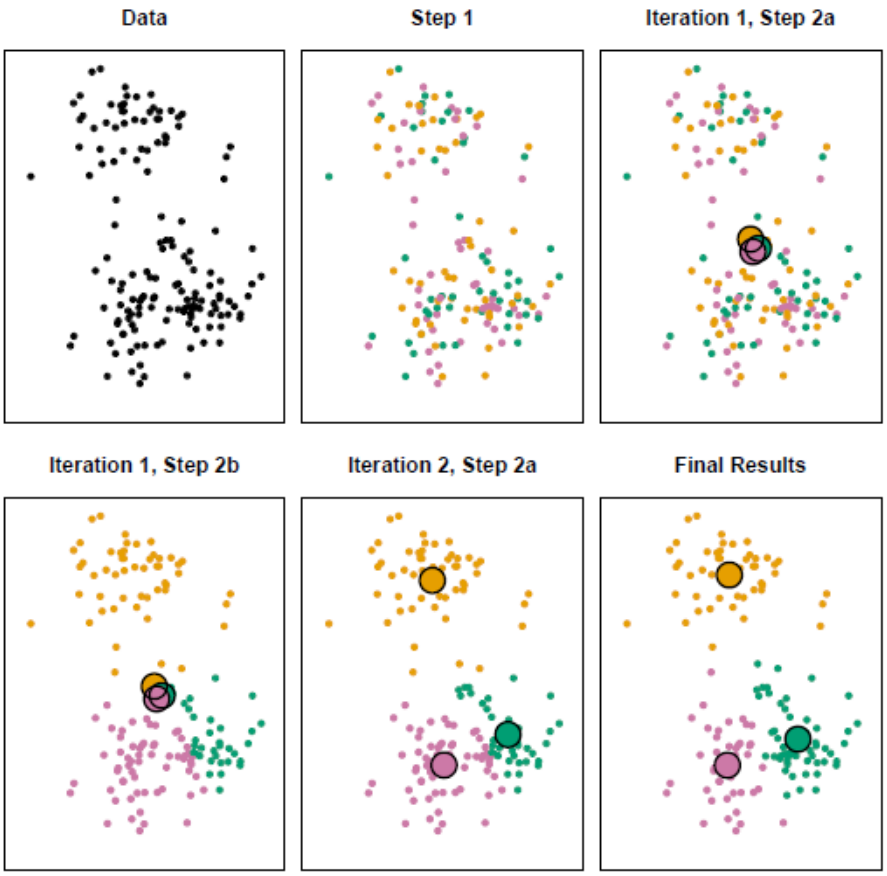
W – macierz kowariancji wewnątrzklasowej.

Schemat metody *k*-średnich zawiera się w następujących krokach (James i in., 2013, str. 392):

1. Najpierw deklaruje się liczbę klas, a następnie przydziela się każdy obiekt do tylko jednej z klas (losowo).
2. Kroki A i B powtarza się aż do momentu, kiedy kolejne ich powtarzanie nie sprawia, że obiekty zostają zaklasyfikowane do innych klas.
    A. Oblicza się środki ciężkości oraz odległości każdego obiektu od każdego z tych środków ciężkości.
    B. Obiekty przyporządkowuje się na nowo, tak aby każdy obiekt należał do klasy o najbliższym centroidzie (czyli środku ciężkości).

Należy podkreślić, że metoda *k*-średnich w większości przypadków znajduje optimum lokalne, a nie globalne. Dlatego otrzymany wynik zależy od początkowego, losowego przyporządkowania obiektów. Z tego powodu ważne jest, aby algorytm powtórzyć wiele razy, za każdym razem rozpoczynając od innego, losowego przyporządkowania obiektów (James i in., 2013, str. 393).

Rysunek 3 ilustruje sposób działania metody *k*-średnich. Początkiem jest wyjściowy zbiór danych. W kroku 1 losowo przyporządkowano obiekty do trzech różnych klas. W pierwszej iteracji, w kroku 2a wyliczono centroidy, a w kroku 2b na nowo przyporządkowano obiekty. Kroki powtarzano, aż do otrzymania finalnych wyników.



Rysunek 3. Proces działania metody *k*-średnich
Źródło: (James i in., 2013, str. 394)

2.1.3. Liczba klas

Wybór liczby klas, na które zostaną podzielone obiekty ma istotne znaczenie. Należy zachować równowagę pomiędzy zwiększaniem różnorodności klas, a minimalizacją ilości przypadków, kiedy to tylko jeden obiekt stanowi jedną klasę (Austin i in., 2013, str. 247).

Określenie liczby klas oraz obiektów do nich wchodzących (dla metod hierarchicznych) jest możliwe poprzez odcięcie dendrogramu na ustalonej wysokości. Zdarza się, że odcięcie jest wynikiem wiedzy merytorycznej autora na temat badanego problemu. Natomiast istnieją również podejścia bardziej formalne. Jednym z nich jest reguła Mojena

(1977, strony 359-360), dzięki której możliwe jest wyznaczenie liczby skupień w badaniu. Wykorzystuje się do tego nierówność:

$$\alpha_{j+1} > \bar{\alpha} + ks_{\alpha}, \tag{17}$$

gdzie:

$\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_{u-1}$ – poziomy (odległości) połączeń odpowiadające podziałowi na u, u – 1, …, 1 klas,

$\bar{\alpha}$ – średnia arytmetyczna długości wiązań,

s_{α} – odchylenie standardowe długości wiązań,

k – przyjęty parametr.

Przyjmuje się taką liczbę klas u która odpowiada pierwszemu poziomowi połączenia α_j na dendrogramie spełniającemu nierówność (17). Mojena (1977, strony 359-360) proponuje przyjąć wartość parametru $k \in$ (2,75; 3,50). Natomiast Milligan i Cooper (1985, str. 164) twierdzą, że najlepszą wielkością parametru jest k = 1,25.

W ramach metod optymalizujących, gama indeksów do wyznaczania najlepszej liczby klas jest o wiele większa. Nie jest możliwe korzystanie z dendrogramu, jako że przy metodach optymalizujących takowy nie powstaje, gdyż określenie liczby klas musi być znane a priori (Gatnar i Walesiak, 2009, str. 416). Zdaniem Walesiaka (2004, str. 64) trzema najlepszymi globalnymi kryteriami są indeks Calińskiego i Harabasza (18), Bakera i Huberta (19), Huberta i Levine'a (20). Natomiast Kaufman i Rousseeuw (1990, str. 85) zaproponowali indeks Silhouette (21) oparty na odległościach między obiektami w skupieniu i między skupieniami. W pracy zdecydowano się wykorzystać właśnie te cztery wskaźniki opisane wzorami (Gatnar i Walesiak, 2009, str. 418):

$$G1(u) = \frac{tr(B_u)(n-u)}{tr(W_u)(u-1)}, \tag{18}$$

$$G2(u) = \frac{s(+) - s(-)}{s(+) + s(-)}, \tag{19}$$

$$G3(u) = \frac{D(u) - l_w D_{min}}{l_w D_{max} - l_w D_{min}}, \tag{20}$$

$$S(u) = \frac{1}{n}\sum_{i=1}^{n} \frac{b(i) - a(i)}{\max\{a(i); b(i)\}}, \tag{21}$$

gdzie:

tr – ślad macierzy,

B_u – macierz kowariancji międzyklasowej,

W_u – macierz kowariancji wewnątrzklasowej,

n – liczba obiektów,

u – liczba klas,

c, r – numer klasy,

$s(+)$ – liczba par odległości zgodnych (czyli odległości wewnątrzklasowych, które są mniejsze od odległości międzyklasowych),

$s(-)$ – liczba par odległości niezgodnych (czyli odległości wewnątrzklasowych, które są większe od odległości międzyklasowych),

$D(u)$ – suma wszystkich odległości wewnątrzklasowych,

l_w – liczba odległości wewnątrzklasowych,

$D_{min}(D_{max})$ – najmniejsza (największa) odległość wewnątrzklasowa,

$a_i = \sum_{k\in\{P_r\setminus i\}} \frac{d_{ik}}{n_r-1}$ – średnia odległość obiektu i od pozostałych obiektów należących do klasy P_r,

$b_i = \min_{c\neq r}\{d_{iP_c}\}$,

$d_{iP_c} = \sum_{k\in P_c} \frac{d_{ik}}{n_c}$ – średnia odległość obiektu i od obiektów należących do klasy P_c.

Optymalna liczba klas u to taka, dla której indeks (18), (19), (21) osiąga maksimum, natomiast indeks (20) minimum (Gatnar i Walesiak, 2009, str. 418). Możliwa jest sytuacja, w której każdy indeks będzie wskazywał inną najlepszą liczbę klas. Wtedy częstą praktyką jest kierowanie się wiedzą merytoryczną odnośnie optymalnej liczby klas lub wybranie tej liczy klas, która została wskazana przez największą liczbę indeksów.

2.2. Porządkowanie liniowe

Analizując niejednorodny zbiór obiektów i wykorzystując metody analizy skupień możliwe jest podzielenie tego zbioru na względnie zwarte grupy. Nie jest natomiast możliwe stwierdzenie, które obiekty są lepsze lub gorsze od innych (na podstawie przyjętego punktu widzenia). Aby można było to zrobić, obiekty należałoby uporządkować według nasilenia (lub braku nasilenia) wybranej cechy lub zestawu cech (Balicki, 2009, str. 317).

Porządkowanie liniowe jest porządkowaniem ze względu na tylko jedną zmienną i umożliwia uszeregowanie wszystkich obiektów. Z racji tego, że najczęściej potrzeba

dokonać porządkowania zestawu cech, to porządkowanie odbywa się na podstawie syntetycznego wskaźnika (zmiennej syntetycznej), który jest wyliczany na podstawie wszystkich zmiennych tworzących dany obiekt. Dzięki temu możliwe jest porządkowanie liniowe zestawu zmiennych (Balicki, 2009, str. 318). Metody służące do szacowania wartości zmiennej syntetycznej można podzielić na wzorcowe i bezwzorcowe (Bąk, 2018, str. 10).

Pierwszym etapem porządkowania liniowego jest określenie charakteru zmiennych. Wyróżnia się trzy typy (Balicki, 2009, strony 319-320):

1. stymulanty (zmienne, których wysokie wartości są pożądane),
2. destymulanty (zmienne, których niskie wartości są pożądane),
3. nominanty (zmienne, których pożądana jest konkretna, optymalna wartość).

Wybór typu najczęściej odbywa się na podstawie wiedzy merytorycznej.

Metoda Hellwiga jest popularną metodą wzorcową (Bąk, 2016, str. 26). Schemat metody jest następujący (Hellwig, 1981, str. 62; Bąk, 2016, strony 26-27):

1. Zmienne poddawane są standaryzacji według wzoru (3).
2. Wyznaczane są współrzędne wzorca według wzoru:

$$z_{0j} = \begin{cases} \max\limits_{i}\{z_{ij}\} & \text{dla stymulant} \\ \min\limits_{i}\{z_{ij}\} & \text{dla destymulant} \end{cases} \tag{22}$$

3. Obliczana jest odległość obiektów od wzorca według wzoru:

$$d_{i0} = \sqrt{\sum_{j=1}^{m}\left(z_{ij} - z_{0j}\right)^2}. \tag{23}$$

4. Wyznaczana jest wartość zmiennej agregatowej:

$$q_i = 1 - \frac{d_{i0}}{d_0}, \tag{24}$$

gdzie:

$d_0 = \overline{d_0} + 2s_d$,

$\overline{d_0} = \frac{1}{n}\sum_{i=1}^{n} d_{i0}$,

$$s_d = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(d_{i0} - \overline{d_0}\right)^2}.$$

5. Obiekty są szeregowane według zmiennej agregatowej q_i. Najlepszym obiektem i jest obiekt, dla którego spełniony jest warunek $q_i = \max_i q_i$. Analogicznie w przypadku najgorszego.

W większości przypadków wartości zmiennej agregowanej $q_i \in [0; 1]$, natomiast mogą wystąpić wartości spoza tego przedziału. Będzie to świadczyć o tym, że w zbiorze występują obiekty silnie odstające od innych.

Metoda TOPSIS jest oparta na wyznaczeniu nie tylko wzorca, ale też i antywzorca. Schemat metody został ustalony następująco: (Hwang i Yoon, 1981, strony 128-132; Bąk, 2016, str. 26):

1. Zmienne poddawane są normalizacji (przekształceniu ilorazowemu) według wzoru:

$$z_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}^2}, \tag{25}$$

gdzie:

x_{ij} – obserwacja *j*-tej zmiennej *i*-tego obiektu.

2. Wyznaczane są współrzędne wzorca według wzoru:

$$z_{0j}^{+} = \begin{cases} \max_i\{z_{ij}\} & \text{dla stymulant} \\ \min_i\{z_{ij}\} & \text{dla destymulant} \end{cases}. \tag{26}$$

3. Wyznaczane są współrzędne antywzorca według wzoru:

$$z_{0j}^{-} = \begin{cases} \max_i\{z_{ij}\} & \text{dla destymulant} \\ \min_i\{z_{ij}\} & \text{dla stymulant} \end{cases}. \tag{27}$$

4. Obliczana jest odległość obiektów od wzorca według wzoru:

$$d_{i0}^{+} = \sqrt{\sum_{j=1}^{m}\left(z_{ij} - z_{0j}^{+}\right)^2}. \tag{28}$$

5. Obliczana jest odległość obiektów od antywzorca według wzoru:

$$d_{i0}^{-} = \sqrt{\sum_{j=1}^{m} \left(z_{ij} - z_{0j}^{-}\right)^2}. \tag{29}$$

6. Wyznaczana jest wartość zmiennej agregatowej:

$$q_i = \frac{d_{i0}^{-}}{d_{i0}^{+} + d_{i0}^{-}}. \tag{30}$$

7. Obiekty są szeregowane według zmiennej agregatowej q_i. Najlepszym obiektem i jest obiekt, dla którego spełniony jest warunek $q_i = \max_i q_i$. Analogicznie w przypadku najgorszego.

Metoda Hellwiga była pierwszą propozycją porządkowania liniowego wykorzystującą wzorzec. Natomiast metoda TOPSIS była pierwszą z wykorzystaniem wzorca i antywzorca (Bąk, 2018, str. 10).

Ocena zależności pomiędzy wynikami dwóch metod możliwa jest dzięki współczynnikowi τ-Kendalla wyrażanego wzorem (Geller, 2016, str. 3):

$$\tau = \frac{\sum_{1 \le i < j \le n}^{n} Q\left((a_i, b_i), (a_j, b_j)\right)}{n(n-1)/2}, \tag{31}$$

gdzie:

$Q\left((a_i, b_i), (a_j, b_j)\right) = -1$ jeśli $(a_i, b_i)(a_j, b_j) < 0$,

$Q\left((a_i, b_i), (a_j, b_j)\right) = 1$ jeśli $(a_i, b_i)(a_j, b_j) > 0$,

(a_i, b_i) – para obserwacji i (gdzie obserwacja to miejsce w rankingu, po uszeregowaniu względem zmiennej agregatowej).

Im wartość współczynnika (31) jest bliższa 1, tym porównywane dwa rankingi dały bliższe sobie wyniki.

3. Analiza empiryczna

Rozdział trzeci został poświęcony przeprowadzeniu analizy skupień dbałości o środowisko państw Unii Europejskiej w 2019 roku. Zdecydowano się również uszeregować te państwa w rankingu z wykorzystaniem porządkowania liniowego.

3.1. Charakterystyka zbioru danych

Zbiór wykorzystanych danych składa się z siedmiu zmiennych (zob. Tabela 1). Analizę rozpoczęto od wyznaczenia podstawowych statystyk opisowych (zob. Tabela 2), które już na samym początku umożliwiają wysunięcie wstępnych wniosków na temat badanego problemu. Określono również czy dana zmienna jest stymulantą czy destymulantą (zob. Tabela 3).

Tabela 2. Wybrane statystyki opisowe analizowanych zmiennych

Zmienna	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Średnia	9,39	39,49	80,47	22,38	0,73	327,8	9,29
Odchylenie standardowe	6,32	14,81	18,26	12,09	0,34	233,38	3,35
Mediana	8,14	38,5	85,51	18,36	0,70	260,4	8,50
Skośność	0,84	-0,30	-0,76	0,96	0,71	0,55	1,36
Kurtoza	-0,01	-0,67	-0,70	0,32	-0,45	-0,89	2,06
Wartość minimalna	0,47	8,90	36,82	7,05	0,20	13,3	5,40
Wartość maksymalna	25,33	66,7	99,96	56,39	1,40	818,9	20,3

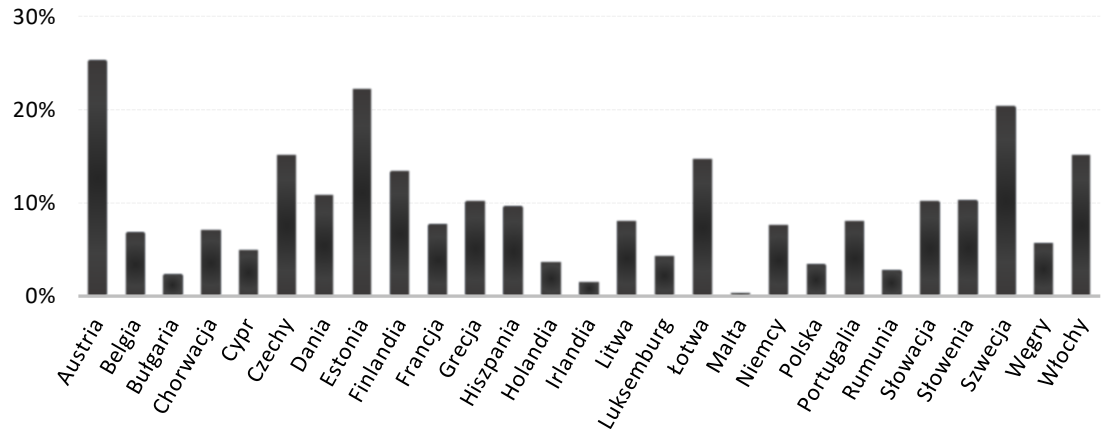
Źródło: opracowanie własne

Tabela 3. Charakter zmiennych

Zmienna	Nazwa zmiennej	Charakter
X_1	Ekologiczne użytki rolne	stymulanta
X_2	Recykling odpadów komunalnych	stymulanta
X_3	Oczyszczane ścieki bytowe	stymulanta
X_4	Energia odnawialna	stymulanta
X_5	Wydatki na środowisko	stymulanta
X_6	Intensywność emisji	destymulanta
X_7	Emisja gazów cieplarnianych	destymulanta

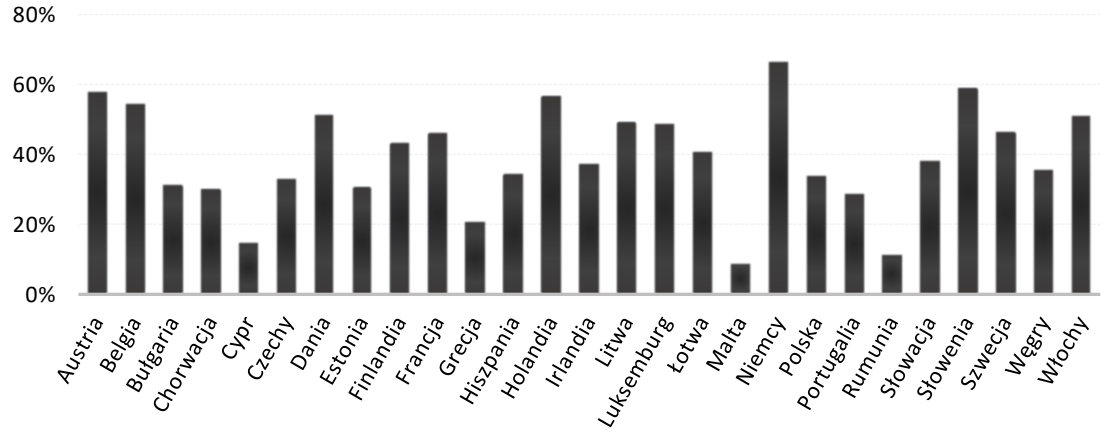
Źródło: opracowanie własne

Zmienna X_1, czyli udział całkowitej powierzchni użytków rolnych zajmowanych przez rolnictwo ekologiczne, charakteryzuje się prawostronną skośnością, co wskazuje, że wartości zmiennej dla większości obserwacji są niższe od średniej (zob. Rysunek 4). Należy podkreślić stosunkowo wysokie odchylenie standardowe świadczące o różnorodności próby. Niektóre państwa (jak Austria czy Estonia) przewyższają inne (między innymi Maltę czy Irlandię) nawet kilkudziesięciokrotnie ze względu na wartości tej cechy. W Austrii ponad 25% użytków rolnych stanowią te uprawiane w sposób ekologiczny, kiedy w Polsce jest to zaledwie 3,49%. Wartości zmiennej często są powiązane z ilością i wysokościami dopłat, ze względu na większe koszty jakie niesie ze sobą ekologiczne rolnictwo w porównaniu do nieekologicznego.



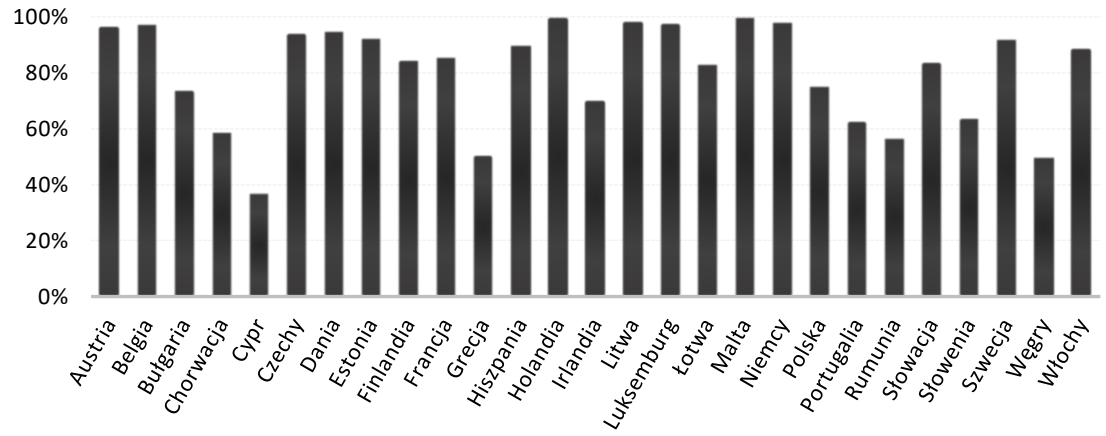
Rysunek 4. Wartości zmiennej X_1(udział całkowitej powierzchni użytków rolnych zajmowanych przez rolnictwo ekologiczne)
Źródło: opracowanie własne

Zmienna X_2 reprezentująca udział odpadów komunalnych poddanych recyklingowi jest o wiele mniej zróżnicowana niż X_1. Liderem są Niemcy, zaś Malta uplasowała się najniżej. Ujemna kurtoza wskazuje na słabą koncentrację cechy wokół średniej. Można zauważyć (zob. Rysunek 5), że obserwacji bliskich średnich jest o wiele mniej niż tych znacznie od niej odbiegających.



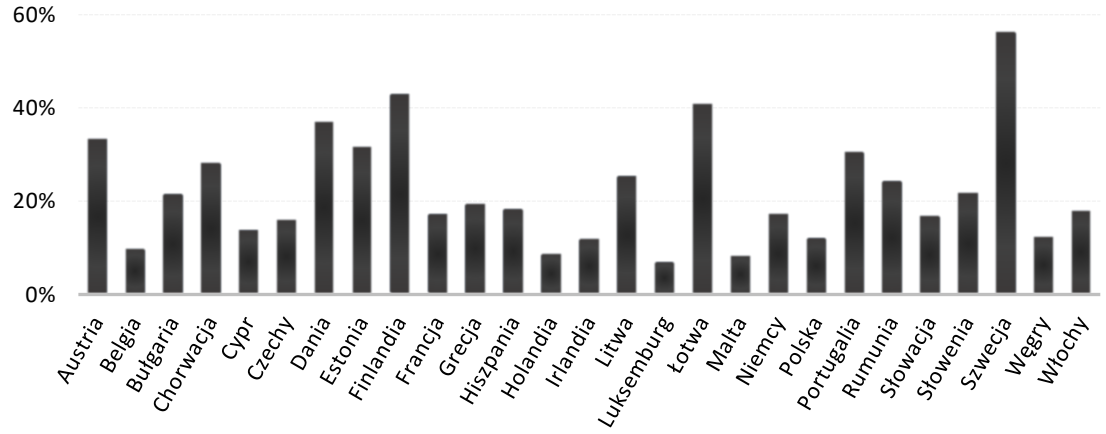
Rysunek 5. Wartości zmiennej X_2 (udział odpadów komunalnych poddanych recyklingowi)
Źródło: opracowanie własne

Udział ścieków bytowych poddanych oczyszczaniu reprezentuje zmienna X_3, która cechuje się zarówno ujemną skośnością jak i kurtozą. Świadczy to o tym, że wartości są słabo skoncentrowane wokół średniej, jak również o tym, że większość z nich jest większa niż średnia. Można zaobserwować kilka państw (zob. Rysunek 6), które znacznie tą średnią zaniżają. Jest to między innymi Cypr, dla którego wartość zmiennej jest ponad dwukrotnie niższa niż średnia. Natomiast jest kilka państw, jak Malta czy Holandia, gdzie prawie całość ścieków bytowych jest poddawana oczyszczaniu.



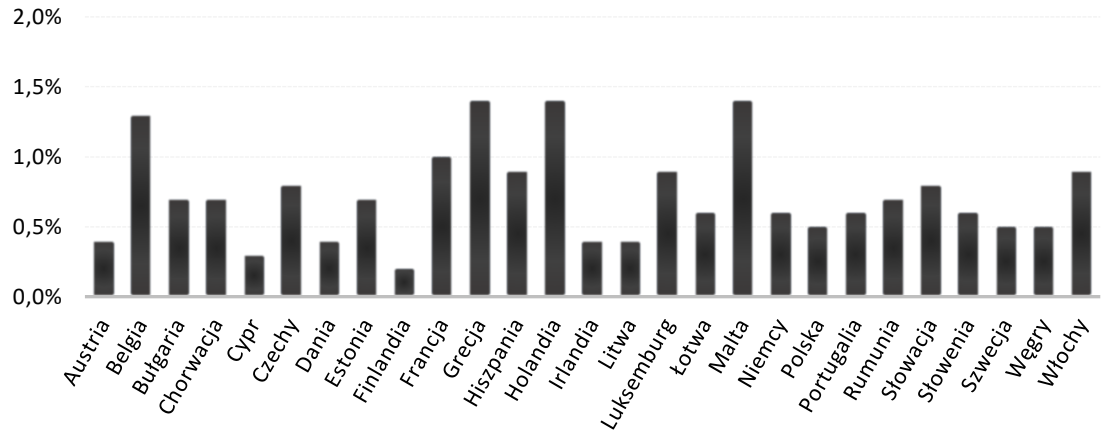
Rysunek 6. Wartości zmiennej X_3 (udział ścieków bytowych poddanych oczyszczaniu)
Źródło: opracowanie własne

Zmienną X_4, czyli udział energii odnawialnej w produkcji energii ogółem, cechuje dosyć duże odchylenie standardowe, próba jest zróżnicowana. Zarówno skośność jak i kurtoza są dodatnie, co świadczy o tym, że wartości są mocno skoncentrowane wokół średniej, jak również o tym, że większość z nich jest mniejsza niż średnia. Zdecydowanie wyróżnia się wysoka wartość zmiennej dla Szwecji, gdzie ponad połowa energii produkowana jest ze źródeł odnawialnych (zob. Rysunek 7).



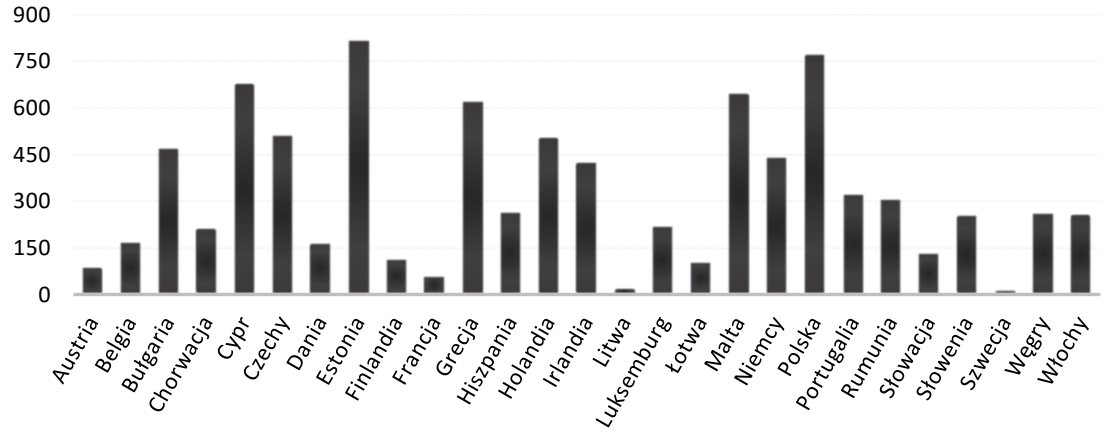
Rysunek 7. Wartości zmiennej X_4 (udział energii odnawialnej w produkcji energii ogółem)
Źródło: opracowanie własne

Wydatki na ochronę środowiska w stosunku do PKB reprezentuje zmienna X_5. Próba jest symetryczna o czym świadczą zbliżone do siebie wartości średniej i mediany. Kurtoza jest ujemna, dlatego można stwierdzić, że występuje słaba koncentracja wartości zmiennej wokół średniej. Najwięcej, bo aż 1,4% wydaje Grecja, Holandia i Malta (zob. Rysunek 8). Siedem razy mniej na ochronę środowiska przeznacza Finlandia i jest to państwo przeznaczające na ten cel najmniej środków w stosunku do swojego PKB. Niewiele więcej wydaje Polska, jest to 0,5%.



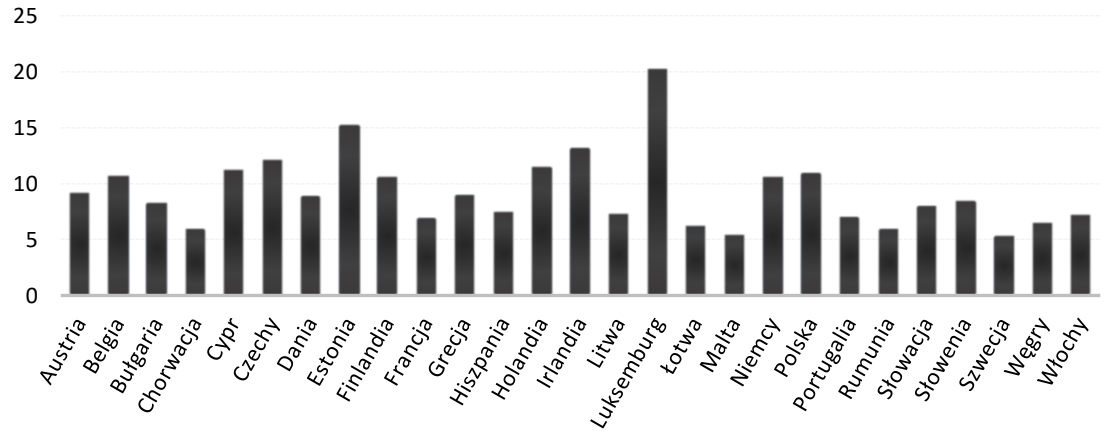
Rysunek 8. Wartości zmiennej X_5 (wydatki na ochronę środowiska w stosunku do PKB)
Źródło: opracowanie własne

Intensywność emisji dwutlenku węgla, czyli zmienna X_6, jest bardzo zróżnicowana – odchylenie standardowe przekracza 230 g CO_2 / kWh, przy średniej równej około 327 g CO_2 / kWh. Największą intensywność emisji można przypisać Estonii i Polsce, zaś najmniejszą Szwecji. Różnica pomiędzy tymi państwami jest ogromna. Można zauważyć (zob. Rysunek 9), że niewiele jest państw skupionych blisko średniej, co potwierdza ujemna kurtoza.



Rysunek 9. Wartości zmiennej X_6 (intensywność emisji dwutlenku węgla)
Źródło: opracowanie własne

Emisja gazów cieplarnianych, czyli zmienna X_7, jest najwyższa w Luksemburgu, czyli drugim najmniejszym państwie Unii Europejskiej (zob. Rysunek 10). W dużej mierze przyczyną tego jest rozbudowany sektor transportu. Zaś najmniejsza emisja jest na Malcie, najmniejszym państwie. Zarówno skośność jak i kurtoza są dodatnie, co świadczy o tym, że wartości są mocno skoncentrowane wokół średniej oraz, że większość z nich jest mniejsza niż średnia wynosząca delikatnie ponad 9 t / osobę.



Rysunek 10. Wartości zmiennej X_7 (emisja gazów cieplarnianych)
Źródło: opracowanie własne

3.2. Analiza skupień dbałości o środowisko w 2019 roku

Podrozdział został poświęcony przeprowadzeniu analizy skupień ze względu na dbałość o środowisko państw Unii Europejskiej w 2019 roku.

Wybór zmiennych, formuły normalizacji i miary odległości

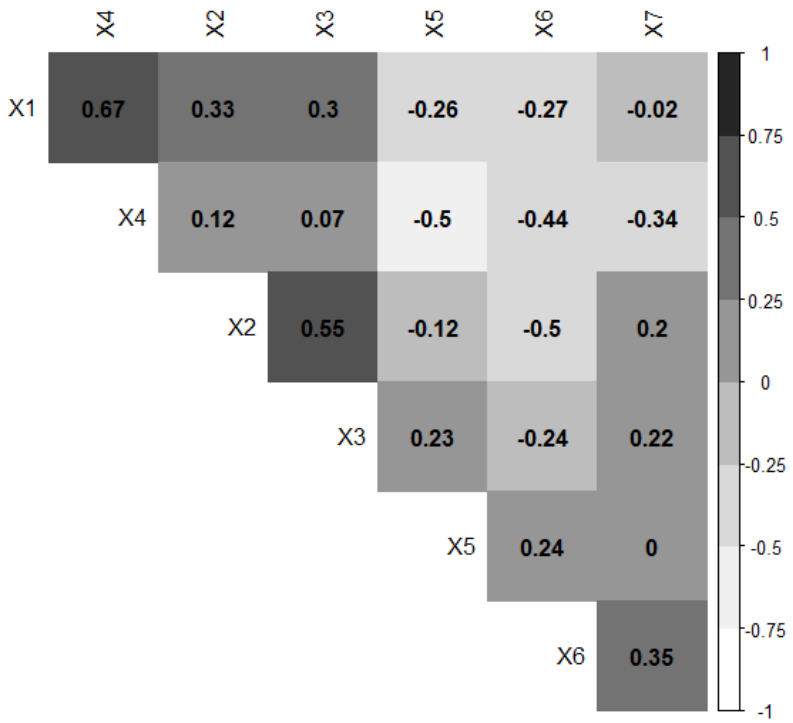
Po pierwszej, merytorycznej części doboru zmiennych przeprowadzonej w rozdziale 1.2, została przeprowadzona ta formalna. Zbadano zmienność przyjętego zestawu zmiennych (zob. Tabela 4) oraz korelacje pomiędzy zmiennymi (zob. Rysunek 11).

Tabela 4. Wartości współczynników zmienności dla przyjętego zestawu zmiennych

Zmienna	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Współczynnik zmienności	0,68	0,38	0,23	0,54	0,47	0,71	0,36

Źródło: opracowanie własne

Wszystkie wartości współczynnika zmienności są większe niż 0,1, dlatego można stwierdzić, iż w przyjętym zestawie zmiennych nie ma takich, które charakteryzują się niską zmiennością. Z uwagi na to kryterium wybór zmiennych został podtrzymany.



Rysunek 11. Współczynniki korelacji dla przyjętego zestawu zmiennych
Źródło: opracowanie własne

Każda wartość współczynnika korelacji jest na moduł mniejsza niż 0,90, dlatego też wszystkie zmienne nie są od siebie silnie zależne i co za tym idzie nie występuje zjawisko powielania informacji. Największa korelacja wystąpiła pomiędzy ekologicznymi użytkami rolnymi a energią odnawialną i pomiędzy recyklingiem odpadów komunalnych a oczyszczanymi ściekami bytowymi. O ile zależność pomiędzy udziałem ścieków bytowych poddanych oczyszczaniu a udziałem odpadów komunalnych poddanych recyklingowi nie budzi większego zdziwienia, ze względu na oczywiste powiązanie, o tyle zależność pomiędzy udziałem całkowitej powierzchni użytków rolnych zajmowanych przez rolnictwo ekologiczne a udziałem energii odnawialnej w produkcji energii ogółem nie jest już tak oczywista. Z uwagi na kryterium niezależności zmiennych ich wybór został podtrzymany. Należy podkreślić, iż wybrane zmienne przeszły pozytywną weryfikację zarówno pod względem merytorycznym jak i formalnym.

W analizie jako formułę normalizacji wybrano standaryzację klasyczną, jako tę najpopularniejszą metodę. Spośród dwóch podstawowych miar odległości dla danych na skali ilorazowej lub przedziałowej, czyli odległości miejskiej i euklidesowej, zdecydowano się wykorzystać tylko tę drugą.

Analiza skupień – metody hierarchiczne

Analizę metodami hierarchicznymi przeprowadzono z wykorzystaniem trzech metod klasyfikacji: pojedynczego połączenia, pełnego wiązania i Warda. Do wyznaczenia liczby skupień została zastosowana reguła Mojena. Przyjęto wartość $k = 1{,}25$ (zob. Tabela 5).

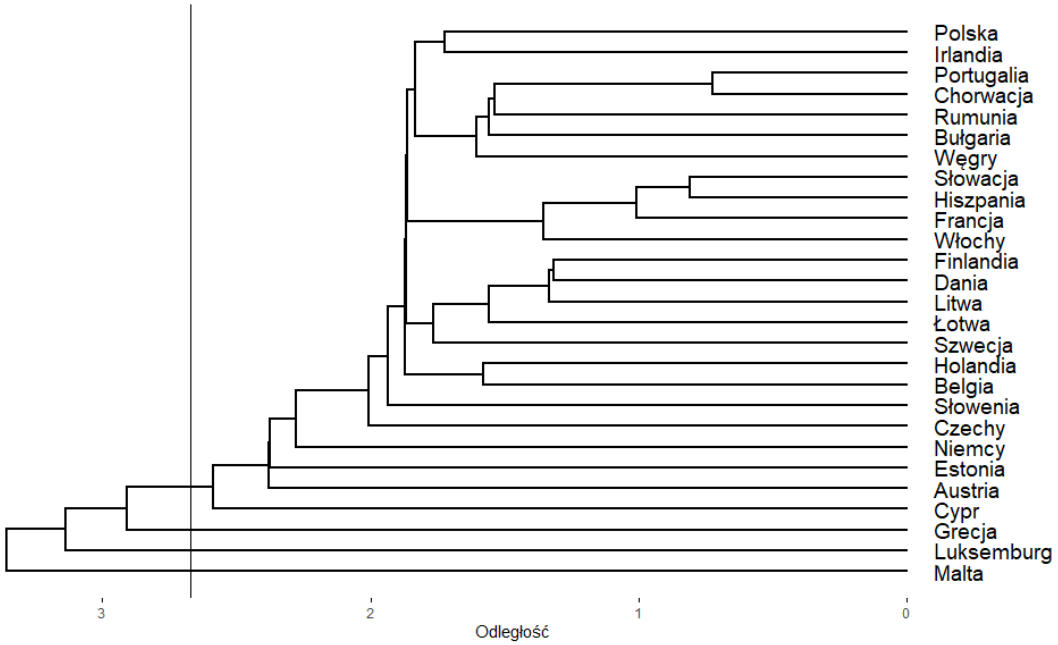
Tabela 5. Optymalna odległość odcięcia dendrogramu według reguły Mojena

Metoda	pojedynczego wiązania	kompletnego połączenia	Warda
Wartość	2,67	4,75	5,78

Źródło: opracowanie własne

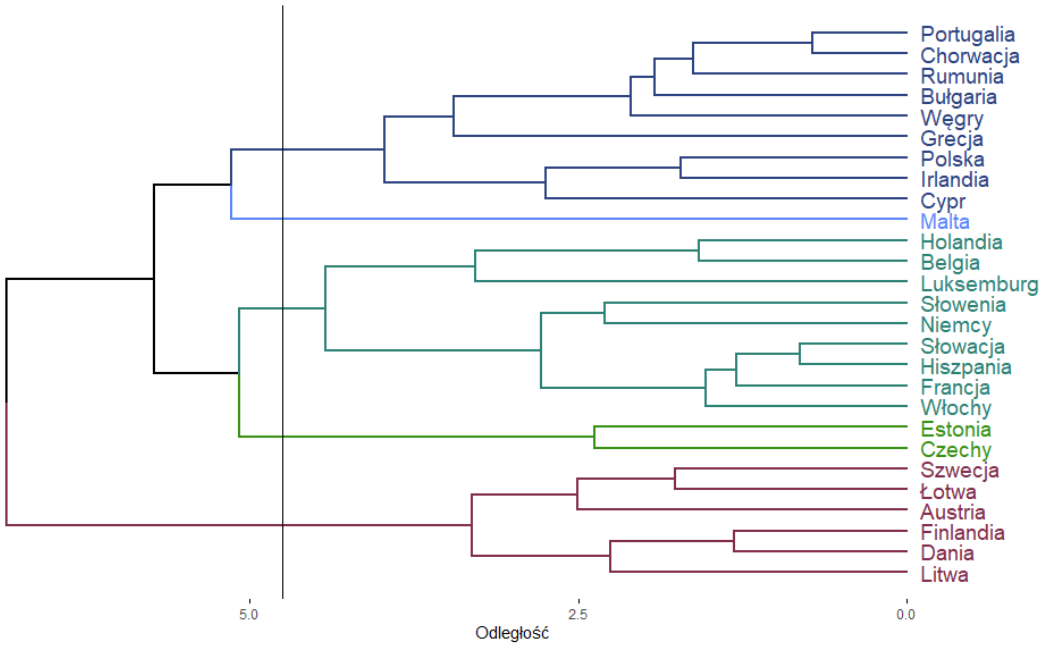
Dendrogram uzyskany metodą pojedynczego połączenia charakteryzuje się widocznym łańcuchem (zob. Rysunek 12). Optymalna liczba klas wyznaczona regułą Mojena wynosi cztery. Natomiast aż trzy klasy zawierają tylko po jednym obiekcie, co jest następstwem tego, iż w tej metodzie o wiele częściej do klasy już istniejącej są dołączane nowe obiekty, niż tworzone są nowe klasy. Najbardziej niepasującym do innych obiektem zdaje się być

Malta, która nawet przy podziale na dwie klasy, dalej byłaby osobnym, jednoobiektowym skupieniem. Wyniki zdają się być mało przydatne i niewiele wnoszące do analizy badanego problemu ze względu na powstawanie klas mało do siebie podobnych.



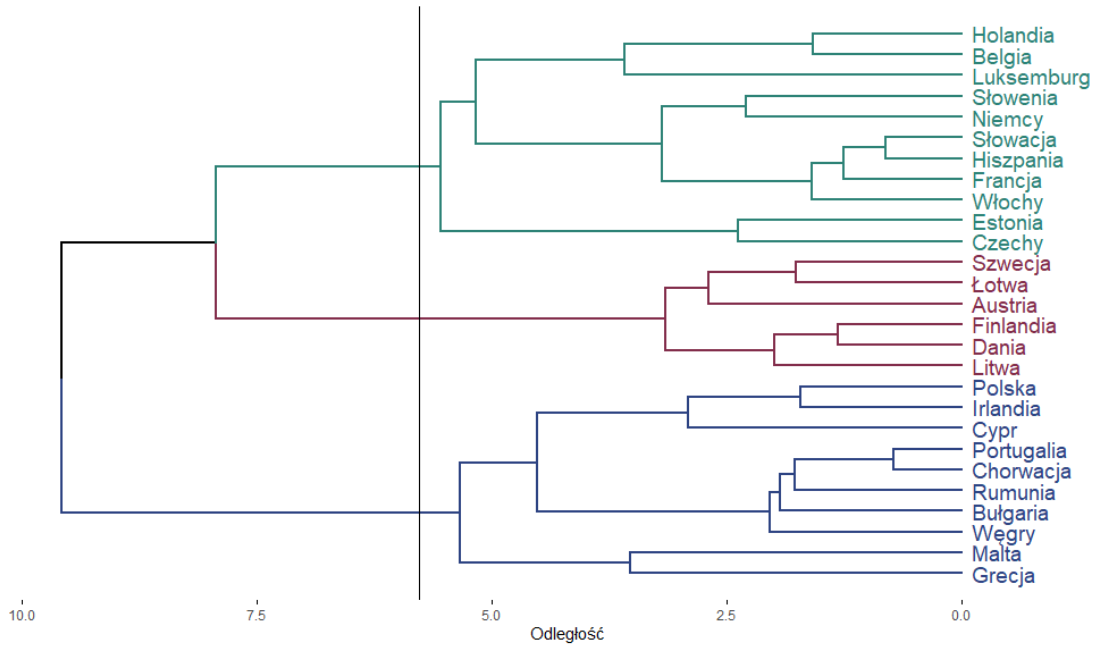
Rysunek 12. Dendrogram - metoda pojedynczego połączenia
Źródło: opracowanie własne

Wyniki uzyskane metodą kompletnego połączenia są oczywiście pozbawione efektu łańcucha. Oceniając dendrogram (zob. Rysunek 13) można stwierdzić, iż właściwe byłoby jego odcięcie, wtedy kiedy podział następuje na trzy lub pięć klas. Dodatkowo reguła Mojena wskazuje optymalną liczbę skupień jako pięć. Wyniki zdają się być o wiele lepsze niż te uzyskane metodą pojedynczego połączenia, mimo, że wyodrębniono dwie małoliczne klasy. Pierwsza jest jednoobiektowa i została zaklasyfikowana do niej Malta, druga natomiast jest dwuobiektowa i w jej skład weszły Czechy i Estonia. Natomiast nie zawsze stan, w którym występują jedno lub dwuobiektowe klasy jest niewłaściwy. Zdarza się, że obiekt, który sam tworzy osobną klasę, jest po prostu bardzo niepodobny do innych.



Rysunek 13. Dendrogram - metoda kompletnego połączenia
Źródło: opracowanie własne

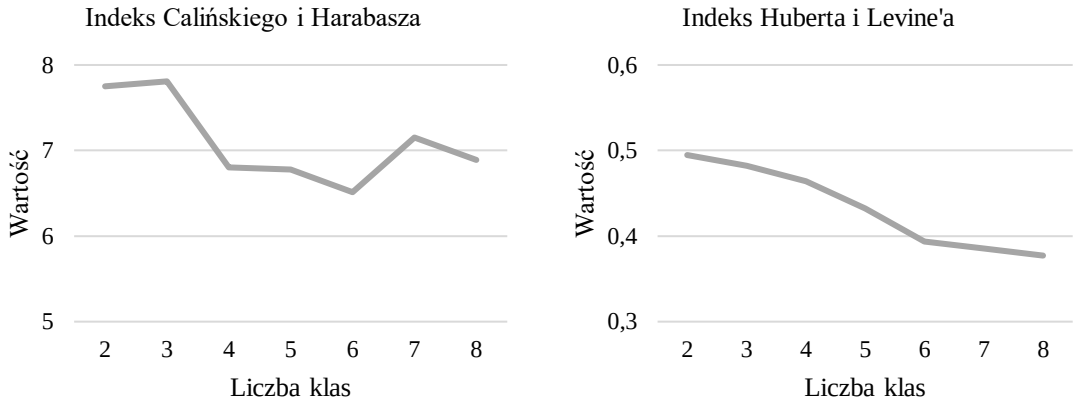
Analizując dendrogram uzyskany metodą Warda (zob. Rysunek 14) można zauważyć, iż optymalna liczba klas zdaje się być równa trzy. Tak samo wskazuje również reguła Mojena. Gdyby dla metody najdalszego sąsiedztwa zastosować podział na trzy skupienia, to podział obiektów na skupienia byłby identyczny jak ten z wykorzystaniem metody Warda.



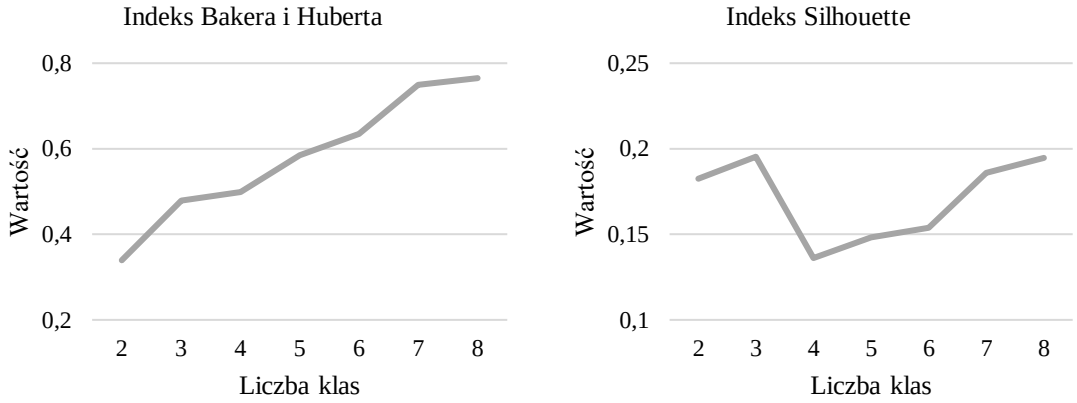
Rysunek 14. Dendrogram - metoda Warda
Źródło: opracowanie własne

Analiza skupień – metody optymalizujące

Analizę w oparciu o metody optymalizujące przeprowadzono wykorzystując metodę *k*-średnich. Wymaga ona aby liczbę skupień znać a priori, dlatego też wykorzystano cztery wskaźniki do wyboru optymalnej liczby skupień. Jednocześnie zdecydowano, że maksymalna liczba klas, bez względu na wskazania indeksów, może wynieść nie więcej niż 30% liczebności zbioru, a więc nie może być większa niż osiem.



Rysunek 15. Wartości indeksu Calińskiego i Harabasza oraz Huberta i Levine'a
Źródło: opracowanie własne



Rysunek 16. Wartości indeksu Bakera i Huberta oraz Silhouette
Źródło: opracowanie własne

Indeksy Silhouette oraz Calińskiego i Harabasza wskazują optymalną liczbę klas jako trzy. Natomiast według indeksu Bakera i Huberta należałoby podzielić zbiór na osiem skupień, tak samo według indeksu Huberta i Levine'a. Biorąc to pod uwagę zdecydowano, że podział metodą *k*-średnich będzie dzielił zbiór obiektów na trzy klasy. Należy podkreślić, iż podział państw na klasy dokonany metodą *k*-średnich jest identyczny jak ten dokonany metodą Warda i kompletnego połączenia (zob. Tabela 6).

Tabela 6. Podział obiektów na klasy metodą *k*-średnich

Klasa A	Klasa B	Klasa C
Austria Dania Finlandia Litwa Łotwa Szwecja	Belgia Czechy Estonia Francja Hiszpania Holandia Luksemburg Niemcy Słowacja Słowenia Włochy	Bułgaria Chorwacja Cypr Grecja Irlandia Malta Polska Portugalia Rumunia Węgry

Źródło: opracowanie własne

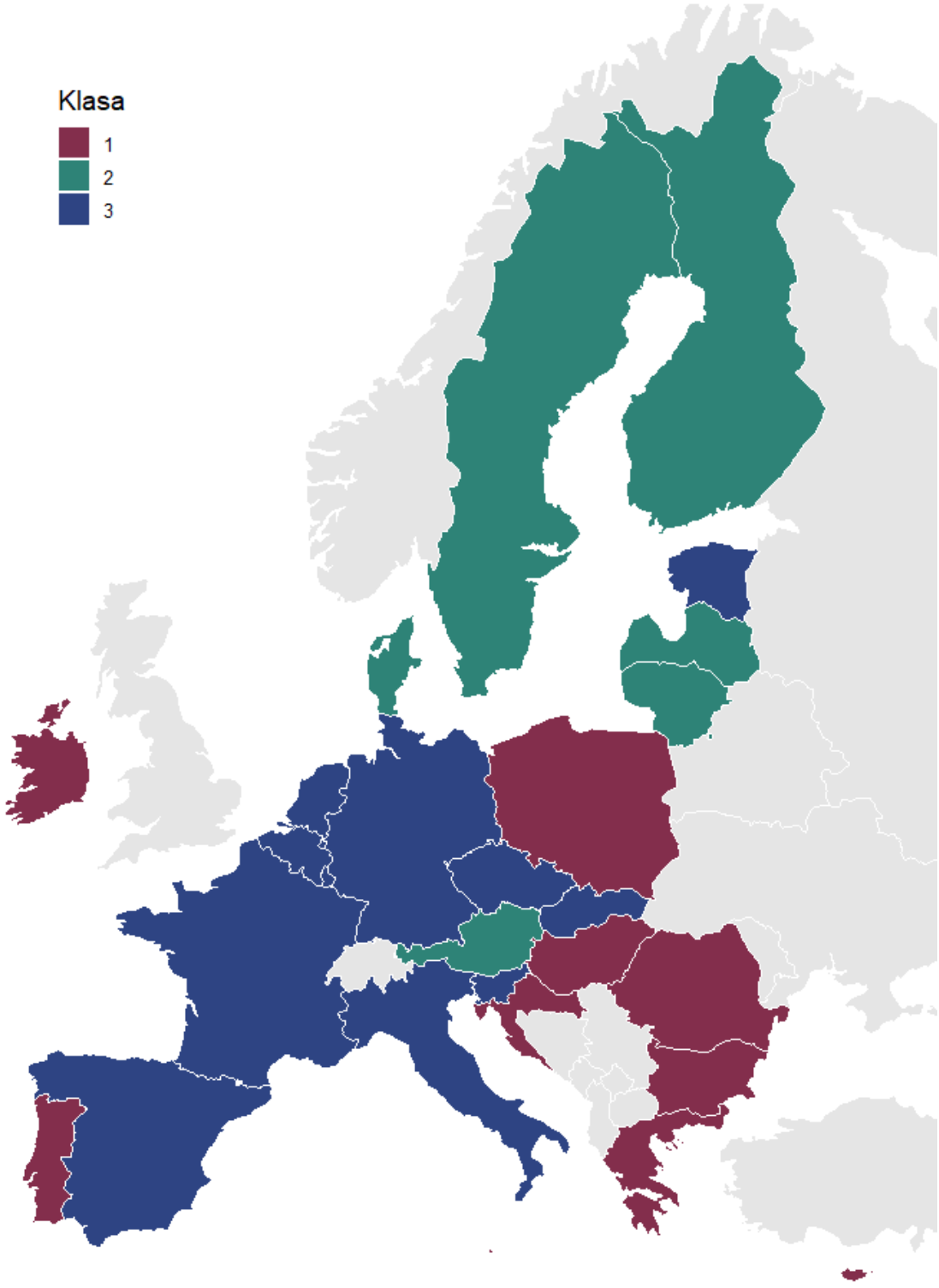
3.3. Zestawienie wyników analizy skupień

Realizując badanie trzema metodami: *k*-średnich, Warda i kompletnego połączenia stwierdzono, iż optymalna liczba klas to odpowiednio trzy, trzy i pięć. Finalnie zdecydowano się dokonać podziału państw na trzy skupienia (zob. Rysunek 17). Bez względu na wybraną metodę, dla liczby klas równej trzy, te same państwa tworzą dane skupienie. Wyznaczono również charakterystyki dla każdego ze skupień (zob. Tabela 7).

Tabela 7. Charakterystyki poszczególnych klas

	Zmienna	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Klasa 1	Średnia	4,71	25,46	63,37	18,37	0,72	**471,74**	8,39
	Odchylenie standardowe	3,13	10,54	17,47	7,67	0,38	197,61	2,66
Klasa 2	Średnia	**15,51**	**48,37**	**91,53**	**39,46**	0,42	83,37	7,98
	Odchylenie standardowe	6,34	6,14	6,41	10,36	0,13	58,92	1,98
Klasa 3	Średnia	10,32	47,39	89,99	16,71	**0,90**	330,27	**10,83**
	Odchylenie standardowe	5,43	11,78	10,31	6,83	0,26	219,44	4,06

X_1 – ekologiczne użytki rolne; X_2 – recykling odpadów komunalnych; X_3 – oczyszczane ścieki bytowe; X_4 – energia odnawialna; X_5 – wydatki na środowisko; X_6 – intensywność emisji; X_7 – emisja gazów cieplarnianych
Źródło: opracowanie własne



Rysunek 17. Podział na klasy uzyskany metodą Warda i *k*-średnich
Źródło: opracowanie własne

Do klasy 1 zakwalifikowano dziesięć państw. Są to: Bułgaria, Chorwacja, Cypr, Grecja, Irlandia, Malta, Polska, Portugalia, Rumunia i Węgry. Klasę można określić jaką tą najgorszą ze względu na badany zestaw zmiennych. Pierwszym z powodów jest najniższa

średnia wartość odsetka ekologicznych użytków rolnych równa 4,71%. Jest to około dwukrotnie mniej niż średnia w grupie 3 i trzykrotnie mniej niż w klasie 2. Ponadto odchylenie standardowe jest dość niewielkie, co świadczy o tym, że dla wszystkich państw w tej klasie ta wartość jest niska. Biorąc pod uwagę recykling odpadów komunalnych to państwa wchodzące w skład klasy 1 poddają recyklingowi tylko czwartą część odpadów, kiedy w pozostałych dwóch skupieniach jest to prawie dwa razy więcej. Oczyszczane ścieki bytowe również są na niższym poziomie, bo w krajach nie wchodzących do skupienia 1 oczyszcza się ich prawie o połowę więcej. Energia odnawialna stanowi nieco ponad 18% energii wytwarzanej ogółem i jest to średnia podobna do tej dla państw ze skupienia 3, natomiast ponad dwa razy mniejsza w stosunku do krajów z klasy 2. Przekłada się to w jakiejś części na intensywność emisji, która w przypadku klasy 1 jest ogromna i wynosi około 470 g CO_2 / kWh, kiedy w klasie 2 jest to około sześć razy mniej.

Polska, która znalazła się w klasie 1, cechuje się niskimi wartościami zmiennych typu stymulanty i wysokimi dla zmiennych typu destymulanty, nawet w obrębie swojej klasy. To oznacza, iż Polska wypada gorzej niż przeciętne państwo z najgorszego skupienia. Szczególnie zauważalne są niskie nakłady finansowe, czyli wydatki na środowisko w stosunku do PKB. Grupa 1, jako ta najgorsza, winna się charakteryzować wysoką średnią klasową tej właśnie zmiennej, która często rzutuje na poprawę pozostałych. W wypadku Polski takie zjawisko nie występuje. Państwo to wydaje mniej niż większość państw Unii Europejskiej, zwłaszcza z grupy najgorszej. Drugą cechą, która miała znaczący wpływ na sklasyfikowanie Polski w klasie 1 jest fakt, iż Polska to drugi najgorszy kraj w Unii pod względem intensywności emisji. Podobnie zły wynik został odnotowany w kontekście emisji gazów cieplarnianych. Właściwie trudno jest wskazać cechę, której nie należałoby poprawić, aby Polska znalazła się w innej, lepszej grupie. Jedynymi dwiema są recykling odpadów komunalnych i oczyszczane ścieki bytowe, gdzie Polska wypada nieznacznie lepiej niż średnia klasy, do której została zaklasyfikowana. Miejsce Polski należy również odnieść do klasyfikacji jej sąsiadów. Największe podobieństwo Polski do państw ościennych występuje przy porównaniu emisji gazów cieplarniach. Natomiast pozostałe cechy sąsiadów Polski przyjmują wartości bardziej pożądane, największa różnica występuje dla ekologicznych użytków rolnych oraz intensywności emisji. To powoduje, iż Polska znalazła się w grupie innej niż kraje z nią sąsiadujące.

Klasa 2 jest klasą najmniej liczną, do której zakwalifikowano sześć z dwudziestu siedmiu państw. Są to: Austria, Dania, Finlandia, Litwa, Łotwa i Szwecja. Na podstawie dendrogramu dla metody kompletnego połączenia (zob. Rysunek 13) i Warda (zob. Rysunek 14) wydaje się, iż jest to klasa, w której obiekty są najbardziej podobne do siebie, ponieważ odległości na dendrogramie są najmniejsze. Potwierdza to również fakt, iż odchylenie standardowe dla pięciu z siedmiu wybranych zmiennych jest najmniejsze w klasie 2. Skupienie 2 cechują najwyższe średnie wartości dla zmiennych typu stymulanty i najniższe dla zmiennych typu destymulanty. Za wyjątkiem jednej zmiennej – wydatków na środowisko, gdzie średnia wartość w tej klasie jest najniższa spośród badanych trzech skupień, mimo iż zmienna ta jest stymulantą. Nasuwa się pytanie, dlaczego tak jest. Otóż wydatki na środowisko wpływają na wartości pozostałych zmiennych. Jest dość oczywiste, że dbanie o środowisko wymaga sporych nakładów finansowych, chociażby na zmiany w sektorze energetyki, aby jak najwięcej energii było produkowanej z wykorzystaniem odnawialnych źródeł energii. Dlatego też państwa z klasy 2, którą można określić tą najlepszą, prawdopodobnie nie muszą już wydawać tak dużych środków na poprawę jakości ekosystemu, jak państwa gdzie to środowisko nie jest jeszcze w tak dobrej kondycji. To tłumaczy dlaczego średnia wartość tej zmiennej jest najniższa spośród badanych klas, a mimo wszystko klasę 2 można nazwać najlepszą ze względu na badany zestaw zmiennych.

Do klasy 3 zakwalifikowano jedenaście państw. Są to: Belgia, Czechy, Estonia, Francja, Hiszpania, Holandia, Luksemburg, Niemcy, Słowacja, Słowenia i Włochy. Ekologiczne użytki rolne stanowią średnio 10,32% i jest to wartość lepsza niż ta w klasie 1 i gorsza niż ta w klasie 2. Recykling odpadów komunalnych i oczyszczane ścieki bytowe są tylko nieznacznie gorsze niż w klasie najlepszej, czyli 2. W przypadku intensywności emisji średnia wartość dla państw skupienia 3 wynosi około 330 g CO_2 / kWh, natomiast należy podkreślić bardzo duże odchylenie standardowe równe około 220 g CO_2 / kWh, co świadczy o tym, że państwa grupy 2 są bardzo zróżnicowane pod tym względem. Podobne zjawisko występuje dla emisji gazów cieplarniach – średnia wartość dla państw tego skupienia jest najwyższa spośród trzech badanych skupień, ale również odchylenie standardowe jest spore. Może to świadczyć o tym, że państwa skupienia 3 znalazły się w tym skupieniu bardziej ze względu na wartości innych cech niż ze względu na intensywność emisji czy emisję gazów cieplarnianych. Skupienie to określono jako przeciętne.

Warto również odnieść uzyskane wyniki analizy skupień do kwestii geograficznych (zob. Rysunek 17) i historycznych. Klasa uznana za najlepszą to kraje skandynawskie i Austria, o których mówi się, że są to państwa bogate. Ale również do tego skupienia zaklasyfikowano Litwę i Łotwę, które to są państwami postradzieckimi, raczej nie postrzeganymi jako zamożne. Można więc się zastanawiać, dlaczego zostały połączone z Austrią i państwami skandynawskimi. Otóż Litwa, a jeszcze bardziej Estonia, produkują sporą część energii elektrycznej ze źródeł odnawialnych. Estonia pod tym względem zajmuje trzecie miejsce w Unii Europejskiej. Ponadto oba kraje cechują niskie wartości emisji gazów cieplarniach i intensywności emisji. Skupienie 3 uznane za przeciętne to państwa głównie zachodniej części Europy, w większości będące w Unii kilkadziesiąt lat. Wiele z nich to państwa bogate, jak Niemcy, Francja czy kraje Beneluksu. Widać więc, że nie zawsze zamożność i długi staż przynależności do Unii Europejskiej wpływają na dbałość o środowisko w danym kraju tak istotnie, by państwo to znalazło się w najlepszej klasie. Klasa 1 uznana za najgorszą to Polska, państwa południowo-wschodniej części Europy, dwa kraje najbardziej wysunięte za zachód oraz dwa wyspiarskie państwa, czyli Malta i Cypr, które jednocześnie są jednymi z najmniejszych krajów Unii Europejskiej. Żadnego z tych państw nie zalicza się do krajów bogatych, a większość z nich jest członkiem Unii Europejskiej stosunkowo od niedawna.

3.4. Porządkowanie liniowe ze względu na dbałość o środowisko

Jako uzupełnienie klasyfikacji przedstawionej w poprzednim podrozdziale zostało przeprowadzone porządkowanie liniowe metodą Hellwiga i TOPSIS ze względu na wszystkie siedem wykorzystanych w pracy zmiennych (zob. Tabela 8). Zdecydowano, iż wszystkim zmiennym zostanie przypisana ta sama waga.

Obie metody zgodnie wskazały, że najlepszym państwem ze względu na badany zestaw zmiennych jest Szwecja. Drugie miejsce zajęła Austria, a trzecie Łotwa. Wszystkie trzy kraje należą do klasy 2, uznanej za najlepszą. Warto podkreślić, że różnica pomiędzy wartością zmiennej syntetycznej dla Szwecji a wartościami dla kolejnych państw jest znacząca. Natomiast już różnica pomiędzy państwem, które zajęło drugie i trzecie miejsce jest minimalna dla metody Hellwiga (około 0,0008) i delikatnie większa dla

metody TOPSIS. Dlatego może stwierdzić, iż stan środowiska na Łotwie i w Austrii jest bardzo podobny.

Pozostałe trzy państwa należące do najlepszej grupy nie zajęły kolejnych trzech miejsc w rankingu, ale dalej zajmowały wysokie z nich. Najgorsze zajęte miejsce w rankingu dla państwa z grupy 2 to miejsce 8, które przypisano Litwie. Tak wysokie miejsca potwierdzają, iż sklasyfikowanie Szwecji, Łotwy, Austrii, Danii, Finlandii i Litwy do jednej klasy było właściwe. Państwa te cechuje podobna, wysoka dbałość o środowisko naturalne.

Dwa państwa z grupy przeciętnej, które zajęły miejsca wyższe niż te z grupy najlepszej, to Włochy i Francja. Pozostałe państwa z grupy przeciętnej plasują się w środku rankingu, za wyjątkiem Luksemburga, który to zajął aż dwudzieste czwarte miejsce w badaniu przeprowadzonym metodą Hellwiga, a dwudzieste drugie w badaniu z wykorzystaniem metody TOPSIS. Należy odpowiedzieć na pytanie, dlaczego ranking nie wygląda w ten sposób, że pierwsze miejsca zajmują państwa z grupy najlepszej, dalej z przeciętnej, a na końcu z ostatniej, tylko zdarzają się sytuacje, kiedy te granice się zacierają i państwo z grupy przeciętnej zajmuje czwarte miejsce od końca, mimo że grupa najgorsza liczy aż jedenaście państw, więc wydawałoby się, że żadne państwo z grupy przeciętnej nie powinno spaść poniżej szesnastego miejsca. Takie zjawisko może być konsekwencją występowania wartości odstających. Należy również pamiętać, że obiekty w porządkowaniu liniowym są szeregowane ze względu na jedną zmienną agregatową, która z kolei jest wynikiem zależnym od odległości pomiędzy parametrami danego obiektu a wzorcem lub wzorcem i antywzorcem. Dlatego też nawet jedna mocno odstająca cecha danego obiektu może spowodować, że jego miejsce w rankingu będzie odbiegać od miejsca innych państw z jego grupy.

Niskie miejsce w rankingu zajęła Polska (dwudzieste piąte w porządkowaniu metodą Hellwiga i dwudzieste szóste z wykorzystaniem metody TOPSIS). Wpływ na to przede wszystkim miała wysoka intensywność emisji równa aż 773,3 g CO_2 / kWh, co jest drugim najwyższym wynikiem w Unii Europejskiej. Dla porównania w Szwecji jest to zaledwie 13,3 g CO_2 / kWh. Wszystkie badane zmienne dla Polski przyjmują wartości poniżej średniej dla zmiennych typu stymulanty i powyżej dla zmiennych typu destymulanty, co dowodzi mało efektywnego dbania o środowisko. Porównując miejsce

w rankingu dla Polski z miejscami jej sąsiadów należy podkreślić sporą różnicę. Polsce najbliżej do Czech, które zajmują szesnaste lub siedemnaste miejsce (w zależności od metody). Różnica pomiędzy pozostałymi sąsiadami jest jeszcze większa. Uwarunkowania takiej sytuacji są bardzo podobne do tych, które zdecydowały, iż Polska nie jest w tej samej klasie z żadnym ze swoich sąsiadów.

Niskie miejsca w rankingu obsadzone zostały przez inne państwa z grupy najgorszej. Warto zaznaczyć, iż Chorwacja i Portugalia, mimo przynależności to tej właśnie grupy, przeskoczyły o kilka miejsc niektóre państwa z grupy przeciętnej. Należy zaznaczyć, że te dwa państwa łączą się w jedną klasę na dendrogramie jako pierwsze – zarówno dla metody Warda jak i kompletnego połączenia (zob. Rysunek 13, Rysunek 14).

Ostatnie miejsce (uwzględniając obie metody) zajął Cypr. Wszystkie wartości zmiennych dla tego państwa są znacznie poniżej średniej dla zmiennych typu stymulanty i powyżej dla zmiennych typu destymulanty. Taka sama zależność występuje w odniesieniu do średniej dla grupy najgorszej. Bezapelacyjnie można stwierdzić, iż jest to państwo, które najmniej efektowanie dba o środowisko naturalne.

Omawiając różnice w wynikach dla wybranych dwóch metod porządkowania liniowego należy zaznaczyć wysoką zależność pomiędzy nimi, czego dowodzi współczynnik Kendalla równy około 0,90. Innymi słowy rankingi są do siebie bardzo podobne, choć istnieją różnice. Największą z nich jest różnica aż pięciu miejsc dla jednego z państw, ale w większości przypadków różnica ta jest równa jeden lub w ogóle nie występuje. Świadczy to również o stabilności uzyskanych wyników.

Tabela 8. Wyniki porządkowania liniowego metodą Hellwiga i TOPSIS

Państwo	Metoda Hellwiga		Metoda TOPSIS	
	Wynik	Ranga	Wynik	Ranga
Austria	0,5366	2	0,6874	2
Belgia	0,3556	12	0,5161	11
Bułgaria	0,2358	19	0,3829	24
Chorwacja	0,3250	14	0,5079	13
Cypr	-0,0251	27	0,2337	27
Czechy	0,3153	16	0,4525	17
Dania	0,4587	5	0,5777	6
Estonia	0,2652	17	0,4731	15
Finlandia	0,4042	7	0,5824	4
Francja	0,4142	6	0,5519	7
Grecja	0,2196	20	0,4414	18
Hiszpania	0,3846	9	0,5088	12
Holandia	0,2631	18	0,4392	19
Irlandia	0,1136	23	0,3069	25
Litwa	0,3887	8	0,5484	8
Luksemburg	0,1098	24	0,3993	22
Łotwa	0,5358	3	0,6543	3
Malta	0,0927	26	0,3907	23
Niemcy	0,3267	13	0,4566	16
Polska	0,1038	25	0,2696	26
Portugalia	0,3196	15	0,4831	14
Rumunia	0,1695	22	0,4212	21
Słowacja	0,3843	10	0,5282	9
Słowenia	0,3754	11	0,5219	10
Szwecja	0,6145	1	0,7462	1
Węgry	0,1900	21	0,4258	20
Włochy	0,4801	4	0,5812	5

Źródło: opracowanie własne

Zakończenie

Dbanie o ekosystem z dnia na dzień staje się coraz ważniejsze, a ochrona środowiska jest jednym z największych, o ile nie największym globalnym problemem. Przeciwdziałanie globalnemu ociepleniu, redukcja ilości produkowanych odpadów czy rozważne gospodarowanie wodą to tylko nieliczne wyzwania przed którymi stoi ludzkość. Ocena państw ze względu na dbałość o środowisko może być wskazówką lub sygnałem alarmowym do wprowadzania zmian i poprawy stanu ekosystemu.

Celem niniejszej pracy była analiza państw członkowskich Unii Europejskiej ze względu na dbałość o środowisko. Wykorzystana w tym celu analiza skupień bez wątpienia jest bardzo efektywnym narzędziem wykorzystywanym do klasyfikacji obiektów. Liczba metod, zarówno wyznaczania odległości jak i dokonywania klasyfikacji czy wyboru optymalnej liczby skupień jest ogromna. Badanie zostało uzupełnione o stworzenie rankingu ze względu na dbałość o środowisko z wykorzystaniem dwóch metod porządkowania liniowego. Oczywiste jest, iż niniejszą pracę można by rozszerzać o kolejne metody, niemniej spośród całej ich gamy wybrano te najpopularniejsze lub te uznane jako najlepsze w przytoczonych publikacjach naukowych.

Przeprowadzone badania umożliwiły podział dwudziestu siedmiu państw członkowskich Unii Europejskiej na trzy klasy. Skupienie 1 to Polska, państwa południowo-wschodniej Europy oraz Portugalia i Irlandia. Klasa ta została uznana za najgorszą ze względu na dbałość o środowisko. Należy podkreślić niskie miejsce Polski w rankingu oraz stwierdzenie, iż jest to kraj, dla którego zestaw przyjętych zmiennych przyjmuje wartości znacznie mniej pożądane niż w przypadku większości państw. Jest to najbardziej

widoczne w kontekście intensywności emisji. W klasie 2 znalazły się państwa skandynawskie, Austria, Litwa oraz Łotwa. Jest to skupienie uznane za najlepsze, średnie wartości prawie wszystkich zmiennych są najwyższe dla zmiennych typu stymulanty i najniższe dla zmiennych typu destymulanty. Potwierdzeniem jest fakt, iż najlepsze miejsce w rankingu ze względu na dbałość o środowisko zajęła Szwecja, głównie z uwagi na odsetek produkowanej zielonej energii oraz intensywność emisji. Natomiast klasa 3 to głównie państwa środkowej i zachodniej części Europy. Skupienie to określono jako przeciętne, średnie wartości zmiennych są mniej pożądane, niż te w klasie 2, ale bardziej niż te w skupieniu 1. Odzwierciedlenie wyników analizy skupień znalazło się w stworzonym rankingu z wykorzystaniem metod porządkowania liniowego.

Uzyskane w badaniu wyniki znajdują potwierdzenie w literaturze. Oczywiście nie są identyczne, na co wpływ bez wątpienia ma przyjęty inny zestaw zmiennych oraz okres, z którego dane pochodzą. Niemniej badanie potwierdziło powtarzanie się większości zależności, jak na przykład tendencje niektórych państw do bycia klasyfikowanymi we wspólnym skupieniu. Będące uzupełnieniem wyniki uzyskane metodami porządkowania liniowego również znalazły odzwierciedlenie w przytoczonej literaturze. Największe podobieństwo wystąpiło w porównaniu ze wskaźnikiem wydajności środowiskowej.

Analizę można by rozwinąć i przeprowadzić kolejne badania w oparciu o dane z innych lat, a następnie takie wyniki porównać. Drugą perspektywą rozwinięcia pracy zdaje się być zbadanie nie tyle dbałości o środowisko w danym roku, ale też progresu (lub regresu) jakie dane państwo wykonało w ujęciu pięciu czy dziesięciu lat. Niemniej takie badanie powinno być odniesione do ogólnej pozycji danego kraju w kontekście dbania o ekosystem. Państwo, w którym dbałość o ekosystem była na dobrym poziomie już kilkanaście lat temu nie ma za bardzo możliwości znaczącej poprawy, a jeśli nawet, to poprawa ta nie będzie wyglądała tak efektywnie, jak w przypadku państwa będącego kiedyś ocenianym nisko ze względu na dbałość o środowisko. Reasumując, badanie byłoby jeszcze bardziej kompleksowe, kiedy zrealizowano by je w oparciu o progres ze względu na dbałość o ekosystem. Natomiast taka analiza winna być również uzupełniona o ocenę dbałości o ekosystem w kilku, podobnie odległych względem siebie okresach, na przykład w ujęciu dekady. Tak przeprowadzone badanie nakreśliłoby jeszcze szerszy obraz badanego problemu niż ten zrealizowany w niniejszej pracy.

Bibliografia

Austin, E., Coull, B. A., Zanobetti, A. i Koutrakis, P. (2013). A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition. *Environment International, 59*, strony 244-254.

Balicki, A. (2009). *Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomicze.* Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.

Bąk, A. (2016). Porządkowanie liniowe obiektów metodą Hellwiga i TOPSIS - analiza porównawcza. *Praca naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, strony 22-31.

Bąk, A. (2018). Zastosowanie metod wielowymiarowej analizy porównawczej do oceny stanu środowiska w województwie dolnośląskim. *Wiadomości statystyczne*, strony 7-20.

Bilgen, S., Keles, S., Kaygusuz, A., Sari, A. i Kaygusuz, K. (2008). Global warming and renewable energy sources for sustainable development: A case study in Turkey. *Renewable and Sustainable Energy Reviews, 12*, strony 372–396.

Bin, S. i Ang, B. (2016). Carbon emission intensity in electricity production: A global analysis. *Energy Policy, 94*, strony 56-63.

Ćwik, J. i Koronacki, J. (2005). *Statystyczne systemy uczące się.* Warszawa: Wydawnictwo Naukowo-Techniczne.

Dołęga, W. (2016). Ekologia w wytwarzaniu. *Energia Gigawat, 5*.

ECOTEC Research & Consulting Limited. (2008). *Renewable Energy Sector in the EU: its Employment and Export Potential.* Birmingham.

European Commission. (2007). *Life and waste recycling. Innovative waste management options in Europe.* .

European Commission. (2013). *Facts and figures on organic agriculture in the European Union.*

European Commission. (2018). *Renewable Energy Prospects for the European Union.*

European Commission. (2019). *Towards a sustainable Europe by 2030.*

Frodyma, K. (2017). Energia ze źródeł odnawialnych a stan środowiska naturalnego w Unii Europejskiej. *Studia Ekonomiczne. Zeszyty naukowe Uniwersytetu Ekonomicznego w Katowicach*, strony 38-52.

Gatnar, E. i Walesiak, M. (2009). *Statystyczna analiza danych z wykorzystaniem programu R.* Warszawa: Wydawnictwo naukwe PWN.

Geller, A. (2016). *Calculating Kendall's tau with multiple measurements.* Ozarks: Texas A&M University-Corpus Christi.

Główny Urząd Statystyczny. (2019). *Ekonomiczne aspekty ochrony środowiska 2019.* Warszawa.

Gordon, A. D. (1999). *Classification.* London: Champan & Hall/CRC.

HDRA. (1998). *What is Organic Farming?*

Hellwig, Z. (1981). Wielowymiarowa analiza porównawcza i jej zastosowanie w badaniach wielocechowych obiektów gospodarczych. (W. Welfe, Red.) *Metody i modele ekonomiczno-matematyczne w doskodoskonaleniu*, strony 46–68.

Hulme, K. (2010). Taking care to protect the environment against damage: a meaningless obligation? *International Review of the Red Cross, 92*.

Hwang, C.-L. i Yoon, K. (1981). *Multiple Attribute Decision Making: Methods and Applications.* New York: Springer-Verlag Berlin Heidelberg.

International Energy Agency. (2019). *CO2 emissions from fuel combustion.*

James, G., Witten, D., Hastie, T. i Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

Jaworska, M. (2016). Analiza dystansu Polski do krajów Unii Europejskiej pod względem ochrony środowiska naturalnego. *Metody Ilościowe w Badaniach Ekonomicznych*, strony 46-53.

Jóźwiak, J. i Podgórski, J. (2012). *Statystyka od podstaw.* Warszawa: Polskie Wydawnictwo Ekonomiczne.

Kaufman, L. i Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis.* New York: John Wiley & Sons, Inc.

Kharitonova, A. E. i Ulyankin, A. E. (2020). Analysis of International Eco-Economic Systems Using Data Mining Methods. *Advances in Economics, Business and Management Research*, strony 207-212.

Kijewska, A. i Bluszcz, A. (2016). Analysis of greenhouse gas emissions in the European Union member states with the use of an agglomeration algorithm. *Journal of Sustainable Mining, 15*, strony 133-142.

Kłopotek, M. i Wierzchoń, S. (2015). *Algorytmy analizy skupień.* Warszawa: Wydawnictwo WNT.

Krzyśko, M., Wołyński, W., Górecki, T. i Skorzybut, M. (2008). *Systemy uczące się.* Warszawa: Wydawnictwa Naukowo-Techniczne.

Milligan, G. W. (1996). Clustering validation: results and implications for applied analyses. (P. Arabie, L. J. Hubert i G. de Soete, Redaktorzy) *Clustering and Classification*, strony 341-375.

Milligan, G. W. i Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika,, 50*, strony 59-179.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal, 20*, strony 359–363.

Ríos, A.-M. i Picazo-Tadeo, A. J. (2021). Measuring environmental performance in the treatment of municipal solid waste: The case of the European Union-28. *Ecological Indicators, 123*.

Shahzad, U. (2015). Global Warming: Causes, Effects and Solutions. *Durreesamin Journal*.

Walesiak, M. (2004). Problemy decyzyjne w procesie klasyfikacji zbioru obiektów. *Prace naukowe Akademii Ekonomicznej we Wrocławiu, 1010*, strony 52-71.

Walesiak, M. (2011). *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R.* Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu.

Wang, Q. i Yang, X. (2020). Investigating the sustainability of renewable energy - An empirical analysis of European Union countries using a hybrid of projection pursuit fuzzy clustering model and accelerated genetic algorithm based on real coding. *Journal of Cleaner Production*.

Wendling, Z. A., Emerson, J. W., de Sherbinin, A. i Esty, D. C. (2020). *Environmental Performance Index 2020.* New Haven: Yale Center for Environmental Law & Policy.

World Health Organization i UN-Habitat. (2018). *Progress on Safe Treatment and Use of Wastewater.*

Źródła internetowe

EEA, 2019. <https://www.eea.europa.eu/pl> [dostęp 07.03.2021]

Eurostat, 2019. <https://ec.europa.eu/eurostat/data/database> [dostęp 07.03.2021]

Spis rysunków

Spis tabel