Państwowa Wyższa Szkoła Zawodowa w Nowym Sączu

Tadeusz Kudłacz

Metody wielowymiarowej analizy porównawczej w analizie układów terytorialnych

Nowy Sącz 2021

Redaktor Naukowy
dr hab. Marek Słociński, prof. PWSZ

Redaktor Wydania
dr Marek Reichel

Recenzja
dr Krzysztof Jakóbik

Redaktor Techniczny
dr Tamara Bolanowska-Bobrek

ISBN 978-83-65575-71-5

Wydawca
Wydawnictwo Naukowe Państwowej Wyższej Szkoły Zawodowej w Nowym Sączu
ul. Staszica 1, 33-300 Nowy Sącz
tel.: +48 18 443 45 45, e-mail: wn@pwsz-ns.edu.pl

Adres Redakcji
Nowy Sącz 33-300, ul. Staszica 1
tel.: +48 18 443 45 45, e-mail: tbolanowska@pwsz-ns.edu.pl

Druk
Wydawnictwo i drukarnia NOVA SANDEC s.c.
Mariusz Kałyniuk, Roman Kałyniuk
33-300 Nowy Sącz, ul. Lwowska 143
tel.: +48 18 547 45 45, e-mail: biuro@novasandec.pl

Spis treści

1. Część pierwsza: Uwagi wprowadzające

1.1. Wstęp

W określeniu „metoda wielowymiarowej analizy porównawczej”[1] dwa słowa mają kluczowe znaczenie w wyjaśnieniu istoty problemu, w tym zwłaszcza charakteru metod oraz celu przeprowadzanej analizy, której mają one służyć. Te dwa słowa to: „wielowymiarowa” oraz „porównawcza”. Określenie „wielowymiarowa” podpowiada, że przedmiotem analizy będą zjawiska, procesy (u nas odpowiednio określone obiekty, a dokładniej jednostki terytorialne) o złożonej naturze, które charakteryzowane są wieloma właściwościami i – co bardzo ważne – ta wielość właściwości musi być jednocześnie w analizach uwzględniana. Mówiąc o wielowymiarowości i odnosząc to do bardziej konkretnego przykładu, interesować nas będą analizy odpowiednio definiowanych właściwości jednostek terytorialnych (np. miast, województw), które to właściwości opisywane wieloma charakterystykami (wskaźnikami) możliwe będą do rozpoznania, ale pod warunkiem łącznego uwzględniania wielu wskaźników jednocześnie. Określenie „porównawcza” informuje z kolei, że przedmiotem analizy nie będą właściwości pojedynczych obiektów (jednostek terytorialnych), lecz odnoszenie (porównywanie) właściwości jednych obiektów z właściwościami innych. Mogą to być inne obiekty tej samej klasy (np. inne miasta) lub też odpowiednio skonstruowane obiekty modelowe (wzorcowe).

Przedstawione powyżej wyjaśnienia uogólnić można krótkim stwierdzeniem: wielowymiarowa analiza porównawcza zajmuje się metodami i technikami porównywania obiektów wielocechowych. Metody wielowymiarowej analizy porównawczej, mocno rozwijane w Polsce w latach 80. poprzedniego stulecia, stanowią dzisiaj relatywnie obszerny dział statystyki. Są przedmiotem zainteresowań przedstawicieli różnych dyscyplin, m.in.: ekonomistów, geografów, socjologów, biologów, demografów. Liczebnie ogromna jest obecnie również literatura poświęcona tym metodom. Ograniczając się do jednej tylko pozycji, warto przywołać autora, który należy do wiodących w Polsce twórców opracowań poświęconych metodom WAP (Grabiński, 1984).

W niniejszej książce zakres omawianych metod zawężony zostanie do grupy mającej najczęstsze zastosowanie w analizach zjawisk ekonomiczno-przestrzennych. Metody te określać będziemy umownie metodami taksonomicznej klasyfikacji. Zostaną one omówione w dwóch odrębnych grupach – o czym poniżej, po uprzednim wyjaśnieniu istoty metod klasyfikacji.

W ogólnym ujęciu metody taksonomiczne służą do klasyfikacji obiektów, które charakteryzowane są wieloma cechami reprezentującymi wyróżnione kryterium (podstawę) przeprowadzanej analizy. Klasyfikacja z kolei jest czynnością zmierzającą do podziału danego zbioru obiektów w odpowiednie klasy (grupy), legitymujące się zakładanymi właściwościami. Zgodnie z tytułem podręcznika, „obiektami” podlegającymi klasyfikacji są jednostki terytorialne dowolnej skali przestrzennej. Przedmiotem klasyfikacji mogą zatem być wszystkie gminy w Polsce lub też gminy danego województwa czy nawet powiatu. Może to również dotyczyć zbioru miast, powiatów, województwa czy też krajów; inaczej ujmując, zbioru dowolnych jednostek terytorialnych danego poziomu przestrzennego. Wspomniano powyżej, że wynikiem klasyfikacji są wydzielane klasy (grupy) legitymujące się zakładanymi właściwościami.

[1] W skrócie przyjęło się określać: „metody WAP”.

„Właściwości" te, to innymi słowy kryterium, według którego klasyfikacja jest przeprowadzana. Kryterium tym może być np. zbliżony poziom rozwoju, podobieństwo określonych właściwości, np. warunków dla innowacyjnego rozwoju, lub jakość kapitału ludzkiego. Efektem klasyfikacji są grupy jednostek terytorialnych, których wyróżnikiem jest względna jednorodność podzbioru jednostek należących do poszczególnych, wydzielonych grup; „jednorodność", czyli podobieństwo z punktu widzenia przyjętego kryterium klasyfikacji.

Nasuwa się tu ważne pytanie: jaki jest cel, znaczenie zabiegów klasyfikacyjnych w badaniach naukowych? Klasyfikacja na ogół nie jest celem samym w sobie. Jest natomiast ważnym etapem w postępowaniu badawczym, pozwalającym głębiej wniknąć w istotę rozważanych właściwości. Często jest nawet koniecznym etapem poprawnego rozpoznania. Rozważmy następujący problem: projektem podlegającym rozwiązaniu jest analiza i ocena prawidłowości rozwoju turystyki w Polsce. Prawidłowość ta dotyczy zarówno przebiegu w czasie zjawisk turystycznych (np. ruchu turystycznego z podziałem na jego rodzaje, a także bazy turystycznej), jak również współzależności (np. czynników przesądzających o ruchu turystycznym, czy o rozwoju infrastruktury turystycznej). Zagadnieniem ważnym do wstępnego rozstrzygnięcia jest baza informacyjna będąca podstawą przeprowadzanych analiz. Badanie możemy bowiem prowadzić na zbiorze istniejących w Polsce gmin lub powiatów, ewentualnie na podstawie zbioru województw. Załóżmy, że podstawą badania będą powiaty. Jest ich obecnie (rok 2020) 380 łącznie z powiatami grodzkimi. Nietrudno zauważyć, że w zbiorze wszystkich powiatów są takie, które cechują się wybitnymi warunkami dla rozwoju turystki oraz takie, w których turystyka raczej nie ma szans rozwoju. Między tymi skrajnościami jest całe spektrum pośrednich uwarunkowań, w tym z koniecznym rozróżnianiem ich rodzajów (np. warunki dla turystyki górskie, nadmorskiej, rekreacyjnej, wypoczynku świątecznego itd.). Jeżeli więc wykrywanie wspominanych prawidłowości prowadzone byłoby na zbiorze *en bloc* wszystkich powiatów, wówczas otrzymywane wyniki byłyby znacząco zniekształcone wysoką niejednorodnością zbioru obiektów będących podstawą wnioskowania. O wiele poprawniejszym podejściem jest w takiej sytuacji dokonanie uprzedniego grupowania powiatów w podzbiory legitymujące się zbliżonymi uwarunkowaniami dla rozwoju turystyki, a następnie dopiero przeprowadzanie odpowiedniej analizy w ramach tak wydzielonych klas obiektów.

Głównym kryterium doboru metod omawianych w niniejszym podręczniku jest częstość ich zastosowań w analizach ekonomiczno-przestrzennych. Prezentowane będą zatem metody należące do grupy narzędzi powszechnie stosowanych w badaniach regionalnych. Zakres prezentowanego materiału dydaktycznego dostosowany jest do programu jednosemestralnego przedmiotu studiów, będącego odpowiednią kombinacją zajęć wykładowych i ćwiczeniowych na studiach pierwszego lub drugiego stopnia na kierunkach uruchamianych głównie w ramach dyscyplin *ekonomia i finanse; geografia społeczno-ekonomiczna i gospodarka przestrzenna* oraz *nauki o zarządzaniu i jakości*. Wyraźnego podkreślenia wymaga przyjęta formuła doboru treści. Kierowano się w tym względzie przede wszystkim potrzebą kształtowania umiejętności zastosowań wybranych narzędzi w rozwiązywaniu konkretnych problemów praktyki, rezygnując jednocześnie z dogłębnego charakteryzowania wszystkich kwestii związanych z formalną stroną metod. Książka stanowić zatem może ważną podstawę literaturową zajęć dydaktycznych, dotyczących analiz ekonomiczno-przestrzennych, w których jako wiodący cel dydaktyczny przyjęto: *wyposażyć studenta w wiedzę pozwalającą na dokonywanie doboru właściwych metod do rozwiązywania problemów praktyki związanej z kształtowaniem społeczno-gospodarczego, przestrzennego oraz ekologicznego rozwoju w układach terytorialnych*.

Lektura książki nie powinna sprawiać problemów osobom posiadającym podstawową wiedzę z teorii ekonomii, m.in. w zakresie takich kategorii, jak:

- liberalny oraz interwencyjny mechanizm rozwoju gospodarczego;
- istota, czynniki i mierniki wzrostu oraz rozwoju;
- rynki pracy;
- procesy gospodarcze w skali makro-, mezo- oraz mikro-.

Wskazana jest także elementarna wiedza ze statystyki:

- rozkłady zmiennych, w tym elementarne parametry rozkładu;
- istota rachunku korelacji;
- podstawowe modele regresji.

W realizacji procesu dydaktycznego, któremu podporządkowane są treści niniejszej książki, wielce użyteczna byłaby również wiedza dotycząca polityki rozwoju, w tym zwłaszcza poziomu regionalnego i lokalnego.

Zauważyć należy, że przekazywaniu wiedzy dotyczącej omawianych w książce metod z konieczności towarzyszyć musi kształtowanie umiejętności formowania problemów, do rozwiązania których przydatne są przedstawiane metody. Konieczna jest również biegłość w realizacji procedur obliczeniowych. Zdecydowanie rekomendować należy korzystanie w osiąganiu tego celu z dostępnych pakietów statystycznych. Na ogół na I roku studiów studenci realizują przedmioty związane z technikami obliczeniowymi. Jako przykładowe można wymienić komputerowe wspomaganie obliczeń statystycznych czy też komputerowe wspomaganie projektowania przestrzeni.

Mają zatem podstawową wiedzę dotyczącą niektórych pakietów statystycznych, w tym związanych z wykorzystanie arkusza Excel. Realizacja przedmiotu z treściami nawiązującymi do omawianych metod jest więc okazją ugruntowania umiejętności obliczeniowych. Między innymi w arkuszu Excel jest dostępny zbiór metod w zakładce „formuły". Odrębnymi narzędziami pomocnymi w realizacji obliczeń są specjalistyczne pakiety statystyczne, m.in. wspominana Statistica, a także pakiet statystyczny R. Nawiązując do pierwszego z nich (Statistica), warto zauważyć, że w pakiecie tym jest m.in. zakładka „modele wielowymiarowe", a dalej zakładka „analiza skupień", w ramach której dostępne są obliczenia wprost związane z zastosowaniem aglomeracyjnych metod taksonomicznych oraz metody k-średnich. Tym między innymi metodom poświęcono w książce sporo miejsca.

Książka składa się z dwóch części. Pierwsza z nich dotyczy metod oceny jednostek terytorialnych różnej skali przestrzennej. Najczęściej spotykanym w praktyce przedmiotem oceny jednostek terytorialnych jest ich rozwój; zarówno uogólniony (np. miast danego terytorium), jak też „branżowo" ukierunkowany (np. rozwój turystyki, przedsiębiorczości, rolnictwa, przemysłów wysokiej techniki). Użyteczność tych metod adresować można do rozwiązywania problemów, w których tytułową ocenę nie można w sposób wiarygodny przeprowadzić na podstawie jednego wskaźnika o dostępnych danych liczbowych, lecz konieczne jest użycie szerszego ich zespołu celem naświetlenia różnych aspektów przedmiotu badania. Warto zauważyć, że wraz z przechodzeniem na coraz niższe poziomy przestrzennej dezagregacji zmniejsza się liczba wskaźników syntetyzujących procesy rozwoju, dla których dostępne są dane liczbowe. Przykładowo, w przypadku krajów operować można wskaźnikami pochodnymi PKB. PKB i związane z nim wskaźniki pochodne traktowane są jako najbardziej syntetyzujące miary osiągnięć rozwojowych. Pomimo swoich pewnych ułomności stosowane

są w analizach porównawczych rozwoju różnych krajów. Ale już na poziomie regionów ((NUTS-2) dane w tym względzie są co prawda udostępniane, jednak z dużym na ogół opóźnieniem. Dla niższych poziomów jednostek terytorialnych miary te nie są w ogóle osiągalne. Musimy wtedy posługiwać się pewnym pakietem wskaźników szczegółowych, które dopiero łącznie stosowane opisywać mogą w sposób zadowalający założone kryterium oceny (np. rozwój turystyczny powiatów w Polsce). W sytuacjach takich pojawia się jednak problem, a mianowicie – jak na podstawie szeregu cech szczegółowych opisujących kryterium oceny, o rozbieżnych najczęściej ocenach, sformułować można wskaźnik uogólniający przeprowadzane oceny. Tego rodzaju zagadnieniom poświęcona jest ta właśnie grupa metod. Jak to przedstawialiśmy na wstępie, w literaturze są one często określne jako „metody wielowymiarowej analizy porównawczej". „Wielowymiarowość" uzasadnia wielość wskaźników szczegółowych charakteryzujących kryterium oceny. Inaczej ujmując, ocena każdej rozważanej jednostki terytorialnej uwzględniać musi wielowymiarowość kryterium tej oceny; przykładowo, w ocenie rozwoju miast uwzględnić trzeba wymiar gospodarczy, społeczny, ekologiczny, przestrzenny. Omawianą grupę metod zaliczyliśmy do metod taksonomicznych, w sensie przedstawionej wcześniej ich interpretacji, tzn. „służą do klasyfikacji obiektów, które charakteryzowane są wieloma cechami reprezentującymi wyróżnione kryterium". Jeżeli bowiem dokonamy oceny jednostek terytorialnych danego ich zbioru (np. województw w Polsce), wówczas porządkując monotonicznie badane obiekty (województwa) ze względu na wartości otrzymanych ocen i przyjmując określone granice tych wartości, jesteśmy w stanie wydzielić grupę (klasę) jednostek najwyżej rozwiniętych, nieco niżej rozwiniętych, aż do grupy jednostek o najniższych ocenach. Oczywiście ilość wydzielanych klas, również zasady wydzielania, to pochodna przyjmowanego z góry założenia w tym względzie.

Kolejna część podręcznika poświęcona jest typowym metodom taksonomicznym, tzn. takim, których efektem końcowym jest grupowanie obiektów na zasadzie największego podobieństwa z punktu widzenia przyjętego kryterium klasyfikacji. Metodami tymi są: dendryt wrocławski, metoda diagramu Czekanowskiego, grupa metod aglomeracyjnych oraz metoda k-średnich. Ich dobór dyktowany był częstością wykorzystywania w praktyce (metoda dendrytu wrocławskiego, metody aglomeracyjne oraz metoda k-średnich) lub też z uwagi na ważne wartości dydaktyczne w wyjaśnianiu istoty metod taksonomicznych (metoda diagramu Czekanowskiego).

Inspiracją do przygotowania niniejszego podręcznika były realizowane przez autora cykle wykładów na dwóch różnych kierunkach studiów[2], w ramach których omawiane w podręczniku metody stanowiły zasadniczą ich część. Były to (są to) przedmioty: *metody i techniki analizy regionalnej* oraz *metody analizy przestrzennej*. Decyzję o opracowaniu podręcznika podjęto mimo dostępności książek poświęconych charakteryzowanym metodom[3]. Przesłankami tej decyzji były:

1) Brak książki dokładnie dostosowanej do stopnia szczegółowości oraz zakresu poruszanych zagadnień stosownie do wytycznych wynikających z opracowanego dla obu kierunków efektów uczenia się. W następstwie tego zachodziła konieczność polecania wielu pozycji literatury podstawowej, z jednoczesnym rekomendowaniem wybranych jedynie fragmentów książek.

[2] Są to: Ekonomia (specjalność *Gospodarka i finanse sektora publicznego* oraz kierunek *Gospodarka przestrzenna.*
[3] Wykaz wielu z nich ujmuje zestawienie bibliograficzne na końcu książki.

2) Przygotowanie opracowania, w którym wyraźnie preferowane jest doskonalenie umiejętności formułowania problemów, dla rozwiązania których zastosować można przedstawiane metody, niejako „kosztem" szczegółowości (głębi) charakteryzowania strony formalnej (statystycznej) metod).
3) Eksperckie doświadczenia Autora związane z opiniowaniem przygotowywanych przez praktykę opracowań wykorzystujących specjalistyczne narzędzia analizy, pozwalające na dzielenie się ze słuchaczami wiedzą o najczęściej popełnianych błędach w stosowaniu przedmiotowych metod.

1.2. Spostrzeżenia ogólne dotyczące metod badawczych (metod analizy)

Metodyczny charakter opracowania uzasadnia potrzebę – krótkiego chociażby – zaprezentowania podstawowych kategorii związanych z badaniami naukowymi; m.in. z ich istotą, celem i procesem przebiegu, a także dotyczących kwestii metodycznych. Poniżej przedstawione zostaną odpowiednie wyjaśnienia. Celem przejrzystości ujęto je w formę krótkich, wręcz definicyjnych, określeń. Zainteresowanych pogłębieniem wiedzy w zakresie przedstawianych kategorii odsyłam do specjalistycznej literatury (Apanowicz, 2002; Krajewski, 2010; Łobocki, 2000).

1) Metodologia, metodyka, metoda, technika

Metodologia – nauka o metodach badań naukowych stosowanych w danej dziedzinie wiedzy.

Metodyka – zbiór zasad wykonywania określonej pracy lub zbiór zasad wykonywania jakichś czynności zmierzających do danego celu (rozwiązania problemu).

Metoda – świadomie stosowane możliwe sposoby postępowania, mające prowadzić do osiągnięcia zamierzonego celu (rozwiązania problemu).

Technika badawcza – jeden konkretny sposób postępowania, mający prowadzić do osiągnięcia zamierzonego celu (rozwiązania problemu).

Metodologia – jest szeroko rozumianą teorią dotyczącą dostępnych w danej dziedzinie metod i ich cech funkcjonalnych (zastosowań). Metodyka ogranicza tę teorię do zasad właściwego doboru i stosowania metod. Metoda jest gotowym algorytmem[4] działania/postępowania prowadzącego do rozwiązania danego problemu/zadania wraz ze zbiorem koniecznych założeń oraz interpretacją (szerszym opisem) poszczególnych etapów algorytmu. Jeżeli w ciągu rozważanych kategorii: *metodologia – metodyka – metoda* jest jeszcze *technika badawcza*, wówczas metoda najczęściej rozumiana jest nie jako jeden algorytm postępowania, lecz zespół możliwych (alternatywnych) algorytmów. Technika badawcza jest wtedy jednym z takich algorytmów postępowania, prowadzącym do rozwiązania problemu. Między metodą a techniką badawczą nie ma jakiejś zasadniczej różnicy. Techniki badawcze są bliżej skonkretyzowanymi sposobami postępowania badawczego. Stanowią jakby „ostatni akord" danej metody badań, która jest dla nich zawsze istotnym punktem odniesienia i obejmuje kilka ich odmian (Łobocki, 2000, s. 29). Najogólniej można powiedzieć, że zarówno metody, jak i techniki badań to sposoby postępowania naukowego, mające na celu rozwiązanie sformułowanego uprzednio problemu. Różnice między nimi upatrywać można również w tym, że:

[4] Algorytm: *przepis postępowania prowadzący do rozwiązania ustalonego problemu, określający ciąg zdefiniowanych czynności elementarnych, które należy w tym celu wykonać.*

> metody są raczej ogólnie zalecanymi (postulowanymi) sposobami rozwiązywania nurtujących badacza problemów. Techniki natomiast odnoszą się do bardziej uszczegółowionych sposobów postępowania badawczego i faktycznie stosowanych w danej nauce. Są one więc także metodami badań, lecz nie w ogólnym, a węższym znaczeniu tego słowa (Ibidem, s. 28).

Jako przykład możemy podać badanie zależności w zbiorze odpowiednio zdefiniowanych zjawisk metodą „rachunku korelacji". W ramach takiej metody wyróżnić możemy cały szereg sposobów wyznaczania konkretnych współczynników korelacji: współczynniki korelacji prostej, współczynniki korelacji krzywoliniowej, współczynniki korelacji wielorakiej, współczynniki korelacji cząstkowej itd. Wymieniane rodzaje współczynników, w stosunku do ogólnie rozumianego rachunku korelacji, nazwać możemy właśnie technikami badań. Wybór konkretnej techniki należy oczywiście do badacza podejmującego się rozwiązania danego problemu i musi być dokonywany w oparciu o należytą wiedzę dotyczącą właściwości, zarówno danych technik, jak i metody, do której techniki te należą.

Należy wyraźnie podkreślić, że w prezentowaniu metod kolejno dalej omawianych nie będziemy wyraźnie rozróżniali: to są metody, a to są techniki analizy. Jak dowodzą tego powyżej przedstawiane interpretacje, rozróżnianie to jest zawsze relatywne (technika również jest metodą). Oznacza to, że w pakiecie wiedzy z obszaru tu prezentowanego rozróżnianie to nie będzie konieczne.

2) Badania naukowe

„Badania naukowe to poznawanie świata we wszystkich jego przejawach. Przebiega jako wieloetapowy, świadomy i celowy proces zróżnicowanych działań poznawczych" (Apanowicz, 2002, s. 19).

W nieco bardziej rozwiniętej formie twierdzić można, że badania naukowe kreuje „zespół zabiegów poznawczych, działań i czynności ludzi zajmujących się nauką prowadzących do wykrywania prawd o obiektywnej rzeczywistości metodami naukowymi, ich uzasadniania i przedstawiania w postaci pojęć, twierdzeń i teorii naukowych" (Wiśniewski, 1983, s. 14).

3) Proces i etapy badań naukowych

a) Proces badań naukowych: Celowo, logicznie uporządkowany przebieg przedsięwzięć badawczych, prowadzących do rozwiązania problemu naukowego.

b) Etapy procesu badań naukowych (por. Apanowicz, 2002, s. 97):
   - Etap koncepcji badań:
     - ✓ sformułowanie problemu badawczego[5], w tym celów i zakresów projektowanych analiz;

[5] Bardzo często „problem badawczy" jest definiowany przez zestaw kilku ogólnych pytań badawczych, których odpowiedzi mogą być inspiracją do podjęcia analiz prowadzących do pozyskania bardziej oryginalnej wiedzy. Niekiedy zestaw ogólnych pytań badawczych jest rozwijany pytaniami szczegółowymi.

- ✓ zdefiniowanie hipotez badawczych[6];
- ✓ określenie podstaw informacyjnych badań; np. w badaniach ilościowych (patrz poniżej) dobór zmiennych i ich wskaźników;
- ✓ dobór metod (technik) badawczych.

➢ Etap realizacji badań:

- ✓ gromadzenie informacji, ich porządkowanie, przeprowadzanie wstępnych analiz rozpoznawczych;
- ✓ właściwy proces badań, z wykorzystaniem zaprojektowanych metod;
- ✓ weryfikacja hipotez;
- ✓ konkluzje końcowe (spostrzeżenia, wnioski, rekomendacje).

4) Ważniejsze rodzaje badań naukowych

a) Ze względu na charakter, cel i przeznaczenie:

➢ Badania podstawowe.

Mają one na celu zdobycie nowej wiedzy oraz umiejętności bez nastawienia na bezpośrednie zastosowanie komercyjne; przykładem może być opracowanie diagnozy w danym obszarze problemowym (por. *Poradnik kwalifikowania zadań*…, s. 7-9).

➢ Badania stosowane.

Mające na celu zdobycia nowej wiedzy ukierunkowanej na zastosowanie w praktyce (Ibidem).

b) Ze względu na liczbę osób prowadzących badanie:

➢ Badania indywidualne.

➢ Badania zespołowe.

c) Ze względu na liczbę dyscyplin naukowych, które wchodzą w skład badania:

➢ Badania monodyscyplinarne.

➢ Badania interdyscyplinarne.

d) Ze względu na czas, jakiego badanie dotyczy:

➢ Dotyczące przeszłości (historyczne).

➢ Dotyczące współczesnych (aktualnych) problemów.

➢ Dotyczące przyszłości (badania prognostyczne).

[6] Jedną z podstawowych kategorii procesu badawczego jest hipoteza badawcza (hipotezy badawcze). W badaniach naukowych **hipoteza** jest prawdopodobnym założeniem (twierdzeniem), którego zgodność lub niezgodność z rzeczywistością powinna być dowiedziona w trakcie prowadzonych czynności badawczych. Stawianie hipotez i dowodzenie ich racji bądź błędu jest podstawą rozwoju nauki. Z pojęciem „hipoteza" ściśle koresponduje termin „teza". **Teza** jest twierdzeniem, które zawsze jest prawdziwe. Teza może być wynikiem hipotezy, która została udowodniona jako prawdziwa i nie wymaga przeprowadzenia dowodu. Tezą jest więc stwierdzenie, które uważamy za prawdziwe, w hipotezie tej pewności nie ma.

e) Ze względu na charakter wykorzystywanych podstaw informacyjnych oraz stosowane metody analizy:

- Badania (metody) ilościowe.

  Zjawisko (lub zespół zjawisk/procesów) opisywany jest pakietem charakterystyk (cech) liczbowych, które pozwalają na wykorzystanie metod mniej lub bardziej specjalistycznego przetwarzania informacji liczbowych. Istotą tej grupy metod jest wykrywanie prawidłowości i współzależności w grupie badanych zjawisk, także analizy prowadzące do uogólniania/syntetyzowania charakterystyk liczbowych. Przeprowadzane analizy w istotnym zakresie ukierunkowane są na weryfikację przyjętych hipotez; jest to podejście o mniejszym na ogół ładunku subiektywizmu, w porównaniu z badaniami typu jakościowego. **Przedmiotem rozważań w niniejszym opracowaniu są właśnie metody tej grupy, a więc metody analizy ilościowej.**

- Badania (metody) jakościowe.

  Istota tej grupy, w ogólnej interpretacji, sprowadza się do wykorzystania pakietu wiedzy, jaką dysponuje podmiot (zespół badawczy) podejmujący się rozwiązania postawionego zadania. Metody jakościowe określane są często mianem metod heurystycznych. Prowadzone analizy zorientowane są głównie na poszukiwanie odpowiedzi na pytania: „jak?", „dlaczego?". Jest to podejście badawcze o większym na ogół ładunku subiektywizmu w stosunku do grupy pierwszej.

5) Cele badań naukowych

a) cel poznawczy – badania naukowe prowadzimy po to, aby: zbudować teorię, wytworzyć nową wiedzę, wzbogacić wiedzę istniejącą, dokonać zmian (reinterpretacji) w wiedzy, która wcześniej została wytworzona,

b) cel praktyczny – występuje, gdy usiłujemy aktywnie i twórczo wykorzystać istniejącą wiedzę do wprowadzenia zmian w obszarze praktyki.

Warto zauważyć, że skrótowo ujęte powyżej cele badań naukowych odpowiednio korespondują z wymienionymi wcześniej dwoma rodzajami badań naukowych: badania podstawowe i badania stosowane.

6) Układy terytorialne

Użyte w tytule podręcznika określenie „...analizie układów terytorialnych" oznacza, że obszarem naszych rozważań i prowadzonych analiz będzie poznawanie (poszukiwanie) określonych właściwości jednostek terytorialnych, m.in. ich ocenianie i klasyfikowanie pod względem wyróżnionych kryteriów czy też ustalanie stopnia zróżnicowania analizowanych ich zbiorów (np. województw w Polsce).

2. Część druga: Metody analizy regionalnej (przestrzennej)

2.1. Metoda wielocechowej klasyfikacji (oceny) jednostek przestrzennych – istota, wymogi stawiane wskaźnikom oceny

Zgodnie z zapowiedzią przedstawioną we wstępie, przechodzimy do omówienia metody o nazwie ujętej powyższym tytułem. Nieco inna jej nazwa – którą będziemy się również posługiwali – to: „metody[7] syntetycznej oceny jednostek terytorialnych". Metody te będą przedmiotem rozważań kilku kolejnych podrozdziałów, w których omawiane będą odpowiednio uporządkowane etapy prac związanych ze stosowaniem wspominanych metod. Pełna procedura/ algorytm[8] postępowania związanego z konstruowaniem wskaźnika syntetycznego (wskaźnik syntetyczny jest finalnym celem poszukiwania) ujmuje zestawienie prezentowane w tabeli 1.

Niniejszy podrozdział ujmuje dwa pierwsze etapy (punkt 1 i 2 przedstawionej procedury; tabela 1).

Etap pierwszy dotyczy istoty problemu, do rozwiązania którego służą metody syntetycznej oceny jednostek terytorialnych. Wyobraźmy sobie, że stajemy przed zadaniem dokonania oceny zespołu jednostek terytorialnych (np. miast w Polsce, gmin powiatu nowosądeckiego, powiatów województwa małopolskiego, województw w Polsce, regionów w Europie, krajów UE itp.) z punktu widzenie określonego (zadanego) kryterium. Może nim być np. poziom społeczno--gospodarczego rozwoju, poziom rozwoju przedsiębiorczości, warunków dla rozwoju turystyki, zagospodarowania infrastrukturalnego itd. Zauważmy jednocześnie, że dla wiarygodnego zobrazowania każdego z przykładowo podawanych kryteriów oceny posłużyć się musimy pewnym pakietem wskaźników[9] szczegółowych. Weźmy pod uwagę, jako przykład, warunki dla rozwoju turystyki gmin powiatu nowosądeckiego. Dla oceny tych warunków trudno byłoby znaleźć jeden wskaźnik (cechę), na którym moglibyśmy w pełni polegać w dokonywanej ocenie gmin. Naturalną wydaje się sytuacja, że szukalibyśmy szerszego zestawu wskaźników, jak np. liczby miejsc noclegowych, być może z rozróżnieniem w odrębnych cechach, kategorii obiektów turystycznych, liczby szlaków turystycznych, odpowiednio mierzonych, szczególnie ważnych atrakcji środowiska przyrodniczego, odpowiednio mierzonych, szczególnie ważnych dla danego segmentu turystyki urządzeń infrastruktury technicznej (np. wyciągi narciarskie dla turystyki zimowej), dostępności komunikacyjnej mierzonej np. gęstością dróg. **Ważnym wnioskiem, nasuwającym się z powyżej prezentowanej istoty metod oceny, jest to, że do przeprowadzania ocen z punktu widzenia założonego kryterium musimy posłużyć się wieloma wskaźnikami szczegółowymi**[10].

[7] Użyta liczba mnoga wynika z tego, że w niektórych etapach przedstawianego poniżej algorytmu pojawiać się będą propozycje alternatywnych rozwiązań.

[8] W dalszej części podręcznika, odwołując się do zamieszczanego zestawienia ujmującego procedurę postępowania związanego z konstruowaniem wskaźnika syntetycznego, często używane będzie określenie „algorytm". Przypomnijmy jego interpretację: w ogólnym znaczeniu, algorytm – to przepis; zestawienie kolejnych kroków (czynności) prowadzących do wykonania określonego zadania lub rozwiązania problemu; uporządkowany sposób postępowania przy rozwiązywaniu zadania.

[9] Dalej posługiwać się będziemy zamiennie trzema określeniami: wskaźnik, miernik, cecha.

[10] Omawiane metody tracą sens w sytuacji, gdy do oceny wystarczyłby jeden wskaźnik.

Tabela 1
Metody syntetycznej oceny jednostek terytorialnych – główne etapy procedury

1. **Zdefiniowanie zadania – zjawisko (kryterium) oceny:**
    a) Terytorialny system społeczno-gospodarczy [TSSG] (np. kraj, region)
    b) Jednostki tworzące ten system
    c) Kryterium oceny (zjawisko podlegające ocenie w ramach TSSG).
2. **Dyskusja cech – mierników szczegółowych (właściwości cech).**
3. **Wybór cech diagnostycznych – metody:**
    a) Metoda grafu
    b) Metoda dendrytu.
4. **Standaryzacja cech diagnostycznych – metody:**
    a) *Zero-jedynkowa*
    b) Uproszczona
    c) *Min-max*.
5. **Agregacja cech standaryzowanych – metody wyznaczania wskaźnika syntetycznego:**
    a) Metoda sumy cech standaryzowanych
    b) Metoda wzorca rozwoju (modelowa).
6. **Interpretacja wyników.**

Źródło: opracowanie własne.

Warto zwrócić uwagę na dwa ważne, kluczowe słowa w nazwie omawianych metod, czyli: „ocena" i „syntetyczna". Zacznijmy od wyjaśnień dotyczących znaczenia pierwszego z nich (drugie będzie omawiane nieco później). W ogólnym ujęciu mówić możemy o ocenie bezwzględnej i o ocenie względnej. Z oceną bezwzględną mamy do czynienia wówczas, gdy staramy się poznać wartości ocenianego przedmiotu/obiektu na podstawie analizy jego składowych, ich stanu, jakości współdziałania, bez odwoływania się porównawczego do innych podobnych. Ocena z kolei względna sprowadza się właśnie do porównywania się z innymi, podobnymi obiektami. Do przeprowadzania tego rodzaju ocen (względnych) służą omawiane tu metody. Całe zadanie wydaje się stosunkowo proste. Mamy bowiem ustalić, jak w relacji do innych sytuuje się dana, oceniana jednostka terytorialna; nieco inaczej ujmując, zadaniem naszym jest dokonanie pewnego rodzaju rankingu ocenianych jednostek, z jednoczesną możliwością wnioskowania, jaki dystans – z punktu widzenia zadanego kryterium oceny – dzieli każdą jednostkę terytorialną od innych jednostek należących do ocenianego ich zbioru.

Wspominaną wcześniej prostotę istotnie mąci jednak wymóg w miarę jednoznacznej oceny, w sytuacji gdy podstawą jest nie jedna cecha, lecz szerszych ich zestaw. Nie mamy przecież prawa oczekiwać zbieżności ocen dokonywanych oddzielnie na podstawie różnych wskaźników szczegółowych. Pod względem niektórych z nich, dane np. województwo cechować się może relatywnie niezłą sytuacją w stosunku do pozostałych, podczas, gdy pod względem innych wskaźników zajmować może bardziej odległe pozycje. Tutaj właśnie

dochodzimy do kwestii zasadniczej omawianych metod, a mianowicie – jak na podstawie wielu wskaźników szczegółowych charakteryzujących daną rzeczywistość (kryterium oceny) dokonać jednoznacznej oceny względnej zespołu branych pod uwagę jednostek terytorialnych, czyli ustalić, na ile dana jednostka jest lepsza lub gorsza od każdej innej. Inaczej ujmując, musimy doprowadzić do sytuacji, która pozwoli na jednoznaczne uszeregowanie monotoniczne naszych ocenianych jednostek terytorialnych (od najlepszej do najgorszej lub odwrotnie), z jednoczesną możliwością oceny dystansu ich dzielącego. To ostatnie sprowadza się do przypisania każdej jednostce terytorialnej jednego (odpowiednio uogólnionego) wskaźnika oceny. Ten właśnie jeden wskaźnik oceny jest przedmiotem poszukiwania omawianych metod. **Najogólniej rzecz biorąc, istota tych metod sprowadza się do konstruowania wskaźnika syntetyzującego (uogólniającego) informacje o ocenianych obiektach, prezentowane przez cechy szczegółowe.**

Powyższa konkluzja naświetla istotę metod syntetycznej oceny, sygnalizowaną przez punkt pierwszy przedstawionej na wstępie procedury. Zauważmy, że w ujętym tam zdefiniowaniu zadania rozróżnione zostały trzy składowe (podpunkty a, b, c, punktu 1). Wymagają one odpowiednich wyjaśnień. Warto zwrócić uwagę, że najczęstszym celem dokonywania ocen jednostek terytorialnych jest chęć rozpoznania stopnia ich zróżnicowania z punktu widzenia określonych zjawisk/procesów. Wiedza w tym zakresie jest przykładowo niezbędna dla potrzeb budowy strategii rozwoju danej jednostki terytorialnej. Trudno byłoby sobie wyobrazić np. budowę strategii rozwoju województwa małopolskiego bez dostatecznej wiedzy o stopniu zróżnicowania już osiągniętego poziomu rozwoju jednostek składowych województwa (przyjmijmy, że będą to gminy, chociaż możliwa byłaby również analiza poprzez pryzmat struktur powiatowych). To właśnie tego rodzaju wiedza jest podstawą projektowania w strategii określonych przedsięwzięć interwencyjnych, mających wzmacniać szanse szybszego rozwoju jednostek nieco słabszych. Przykład ten posłuży nam do zidentyfikowania wspominanych trzech składowych każdego problemu, do rozwiązania którego służą omawiane metody. Terytorialnym systemem społeczno-gospodarczym (TSSG) jest w naszym przykładzie województwo małopolskie; ogólniej – jest nim jednostka terytorialna, dla której przeprowadzana jest analiza stopnia wewnętrznego zróżnicowania pod względem określonego kryterium ocen. Jednostkami tworzącymi TSSG podlegającymi ocenie są w naszym przykładzie gminy; w ogólnym ujęciu, jednostki terytorialne, poprzez ocenę których obrazowane jest wewnętrzne zróżnicowanie TSSG. Jako kryterium oceny w przykładzie wybraliśmy poziom rozwoju. Każdorazowo kryterium oceny dobierane jest przez badacza stosownie do celów poznawczych, jakie sobie stawia.

Wiarygodność, a w efekcie praktyczna przydatność wskaźnika syntetycznego zależeć będzie od dwóch głównych przesłanek: po pierwsze – poprawności procedury syntetyzowania, której to procedurze poświęcimy kilka dalszych podrozdziałów; a po drugie – od wiarygodności wskaźników szczegółowych, będących podstawą konstruowania wskaźnika syntetycznego. Do tej ostatniej kwestii odnosi się punkt 2 algorytmu konstruowania wskaźnika syntetycznego (*dyskusja cech – mierników szczegółowych*). Problemowi właściwego doboru cech szczegółowych warto nadać wysokie znaczenie nie tylko z uwagi na jego przesądzający wpływ na wiarygodność, a zatem i użyteczność otrzymywanych wyników, ale również ze względu na bardzo częste lekceważenie w praktycznym zastosowaniu metod, potrzeby refleksji nad cechami przyjmowanymi do odpowiednich wyliczeń. Doświadczenia praktyki dowodzą, że często wskazywanym kryterium doboru cech szczegółowych do oceny określonych zjawisk jest jedynie ich dostępność.

Jest oczywistym, że dostępność odpowiednich statystyk zawsze zakreślać będzie pole możliwych wyborów. Nie może ona jednak być żadnym usprawiedliwieniem przyjmowania cech o wątpliwej wartości ocennej (stwierdzenie w postaci: „innych nie można było pozyskać" jest mało przekonującym argumentem).

Wskazywana wyżej „dyskusja cech" to inaczej rozważanie wymogów, jakie stawiamy cechom szczegółowym, na bazie których budowany będzie wskaźnik syntetyczny. Ograniczając się do ważniejszych wymogów, zwracamy uwagę, że cechy powinny być:

1) **Adekwatne w świetle kryterium oceny**

   Jak już wiemy, kryterium oceny charakteryzowane jest wieloma cechami szczegółowymi. Czymś naturalnym jest, że powinny one w sposób najbardziej trafny opisywać najistotniejsze aspekty zjawiska/procesu będącego kryterium przeprowadzanej oceny. Rzeczą najważniejszą dla tego wymogu jest zachowanie względnej „równowagi znaczenia cech" w wyjaśnianiu kryterium, które mają reprezentować. Jak można się już teraz domyślać, każda cecha ma określony wpływ na wynik końcowy wskaźnika syntetycznego. Rzecz w tym, że procedura, którą będziemy dalej prezentować nie uwzględnia ważenia cech[11] (różnicowania ich znaczenia w kształtowaniu rezultatu końcowego). Każda zatem z cech w równym stopniu partycypować będzie w otrzymywanych wartościach budowanej syntezy.

2) **Mierzalne (kwantyfikowalne, czyli możliwe do wyrażenia w liczbach)**

   Wymóg ten jest dosyć oczywisty. Jeżeli bowiem wyznaczanie wskaźnika syntetycznego sprowadza się do odpowiedniego ciągu wyliczeń, to ich podstawą musi być liczbowa reprezentacja cech. Nie ta jednak oczywistość jest motywem podnoszenia wskazanego wymogu. Powodem, dla którego warto rozważyć spełnienie wymogu mierzalności cech, są spotykane praktyki szacowania/nadawania liczbowego wymiaru właściwościom o charakterze jakościowym, a więc trudno mierzalnym. W tym tkwi niebezpieczeństwo zbyt dużego ładunku subiektywizmu w przeprowadzanych ocenach; zwłaszcza dokonywanych z ukierunkowaniem na oczekiwania „zleceniodawcy" badań. Biorąc to pod uwagę, unikajmy raczej cech, których wartości liczbowe nadawane muszą być poprzez zabiegi zbyt uproszczonych procedur szacowania. Ich użycie z całą pewnością obniżać będzie wiarygodności wyników.

3) **Kompletne (dotyczące wszystkich ocenianych jednostek terytorialnych)**

   Celem finalnym omawianych metod jest wyznaczanie wskaźnika syntetycznej oceny dla wszystkich jednostek terytorialnych danego ich zbioru (województw w Polsce, miast województwa małopolskiego itd.). Chcąc wykonać takie zadanie, dysponować musimy wartościami liczbowymi odnoszącymi się do każdej ocenianej jednostki terytorialnej. Czyli, jeżeli mamy „m" ocenianych jednostek terytorialnych, to dla każdej z cech szczegółowych dysponować musimy „m" liczbowymi charakterystykami. Jakiekolwiek braki w tym względzie stawiają w wątpliwość możliwość wykonania zadania.

[11] W literaturze pojawiają się propozycje wprowadzania wag cech. Z uwagi na daleko posunięty subiektywizm tego rodzaju zabiegu, w praktycznych zastosowaniach omawianych metod najczęściej jest to jednak pomijane.

4) Wykazujące dostateczne wewnętrzne zróżnicowanie (istotny współczynnik zmienności)

Wyjaśnienie tego wymogu zawiera się w tym, że przedmiotem naszych rozważań są oceny względne. Weźmy pod uwagę skrajny przypadek, tj. cechę, która – opisując dane kryterium oceny – przyjmuje te same wartość dla każdego ocenianego obiektu. Nietrudno zauważyć, że z punktu widzenia oceny względnej nie ma ona żadnej wartości poznawczej. Można nawet sformułować wniosek, że im większe będą różnice w wartościach danej cechy, tym większa jest jej wartość ocenna.

5) Wykazujące dostateczną stabilność w czasie

Wymóg ten jest ważny w sytuacji, kiedy przeprowadzana ocena nie jest jednoznacznie adresowana do ściśle określonego momentu czasu (np. sytuacji w zakresie danego zjawiska na koniec roku). Najczęściej przeprowadzane są oceny z myślą o zapoznaniu się z najbardziej aktualną sytuacją[12]. Wtedy to koniecznej refleksji poddana powinna być decyzja, czy obok cech o wysoce stabilnym liczbowo obrazie (cechy o charakterze kumulatywnym) brane mogą być pod uwagę cechy znacząco mniej stabilne (o charakterze niekumulatywnym)[13]. Musimy mieć świadomość tego, że bezkrytyczne uwzględnienie jednych i drugich nasunie wątpliwość dotyczącą zakresu stabilności otrzymanych w finale rezultatów.

6) Wiarygodne źródło

Również ten wymóg jest bardzo oczywisty, nieco nawiązując do powyższego, ujętego w punkcie 2. Najczęstszym źródłem pozyskiwanych informacji są dane Głównego Urzędu Statystycznego, urzędów jednostek samorządu terytorialnego, różnych organizacji, w tym przedsiębiorstw. Źródła te nie nasuwają żadnych wątpliwości związanych ich wiarygodnością. Z należytą ostrożnością podchodzić powinniśmy natomiast do liczbowej faktografii pochodzącej z różnego rodzaju badań terenowych, nie wykluczając także niektórych badań ankietowych. Konieczne w tym względzie jest uprzednie rozpoznanie profesjonalizmu tego rodzaju badań. Zauważa się jednak w praktyce stosowanie zbyt uproszczonych procedur różnego rodzaju wycen.

7) Cechy przydatne do oceny (stymulanty, destymulanty)

Ogół cech z punktu widzenia siły i kierunku powiązania z kryterium oceny możemy podzielić na stymulanty, destymulanty i nominanty. Stymulanta to taka cecha, której wyższe wartości świadczą pozytywnie o zjawisku którego dotyczą, zaś niższe wartości świadczą, że jest gorzej (PKB, dochody budżetu, stan zainwestowania itd.). Destymulanta z kolei, przyjmując wartości wyższe, dokumentuje gorszą sytuację, zaś przyjmując wartości niższe, oznacza, że jest lepiej w dziedzinie, której dotyczy (stopa bezrobocia, liczba ludności w przeliczeniu na lekarza, śmiertelność niemowląt itd.). Nominantą jest cecha taka, co do której trudno jest jednoznacznie orzekać, że więcej to lepiej, czy też więcej to gorzej.

Nietrudno zauważyć, że niekwestionowaną wartość ocenną mają tylko stymulanty oraz destymulanty. Przyjmijmy zasadę, że do pakietu cech szczegółowych odzwierciedlających kryterium oceny nie będziemy rekomendować cech o niejednoznacznej sile i kierunku powiązania z tym kryterium, czyli nominat.

---

[12] Oczywiście, że aktualność ta determinowana będzie dostępnością odpowiednich danych liczbowych.

[13] Cecha o charakterze kumulatywnymi to taka cecha, której wartości liczbowe dla danego momentu czasu w przemożny zakresie zależą od wartości wcześniejszych. Przykładem może być liczba ludności województw, wartość zainwestowanego majątku w powiatach województwa małopolskiego itd. Cechy niekumulatywne legitymują się przeciwną właściwością, to jest znaczniej mniejszą ich zależnością w danym momencie czasie od tego co było wcześniej (stopa bezrobocia, wskaźniki migracji, także stopa inwestycji).

Na zakończenie przedstawiamy zestawienie ujmujące wyjściowe oznaczenia, którymi będziemy się posługiwać w omawianiu wszystkich metod uwzględnionych w niniejszej książce. Oznaczenia te prezentowane są w postaci tabeli (tabela 2), jak również w zapisie macierzowym [X].

Tabela 2
Zestaw wskaźników do oceny jednostek terytorialnych z punktu widzenia zadanego kryterium

Cecha wskaźnik) / Jedn. terytorialna	1	2	3	...	n
1	x_{11}	x_{12}	x_{13}	...	x_{1n}
2	x_{21}	x_{22}	x_{23}	...	x_{2n}
3	x_{31}	x_{32}	x_{33}	...	x_{3n}
...	...	...	...	...	...
m	x_{m1}	x_{m2}	x_{m3}	...	x_{mn}

m – liczba badanych/ocenianych jednostek przestrzennych
n – liczba cech charakteryzujących kryterium oceny
x_{ij} – wartość j-tej cechy dla i-tej jednostki terytorialnej (kraju województwa, gminy, miasta).

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \dots x_{1n} \\ x_{21} & x_{22} & x_{23} & x_{24} & \dots x_{2n} \\ x_{31} & x_{32} & x_{33} & x_{34} & \dots x_{3n} \\ x_{41} & x_{42} & x_{43} & x_{44} & \dots x_{4n} \\ \dots & & & & \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \dots x_{nn} \end{bmatrix} \quad (1)$$

2.2. Metody syntetycznej oceny jednostek terytorialnych – wybór cech diagnostycznych z zastosowaniem metody grafu

W niniejszym podrozdziale przechodzimy do omówienia punktu 3a algorytmu zamieszczonego w tabeli 1, czyli wyboru cech diagnostycznych za pomocą metody grafu.

Celem lepszego osadzenia problemu podejmowanego w tym fragmencie opracowania w strukturze całości zagadnień metody syntetycznej oceny, warto istotę tej metody przypomnieć (patrz pierwszy punkt powyższego algorytmu):

1) Stajemy przed zadaniem dokonania oceny zespołu jednostek terytorialnych (np. miast, regionów, krajów) z punktu widzenie określonego (zadanego) kryterium (np. poziomu ogólnego społeczno-gospodarczego rozwoju, poziomu rozwoju przedsiębiorczości, warunków dla rozwoju turystyki, zagospodarowania infrastrukturalnego itd.).
2) Należy zauważyć, że kryterium oceny nie da się wiarygodnie opisać jednym wskaźnikiem. Gdybyśmy wymyślili bardzo szczegółowe kryterium oceny, np. poziom bezrobocia, wtedy rzeczywiście jeden wskaźnik wystarczałby (np. stopa bezrobocia). Do takich przypadków zastosowanie omawianej metody nie ma oczywiście sensu. Należy wyraźnie podkreślić, że sens stosowania charakteryzowanej j metody dotyczy wyłącznie sytuacji, gdy kryterium oceny jest opisywane wieloma szczegółowymi cechami/wskaźnikami/miernikami[14]. Z takimi właśnie przypadkami spotykamy się najczęściej w praktyce.
3) Koniecznej uwagi wymaga fakt, że naszym celem jest ocena względna, czyli ustalenie, na ile dana jednostka jest lepsza lub gorsza od każdej innej. Inaczej ujmując, musimy doprowadzić do sytuacji, która pozwoli na jednoznaczne uszeregowanie monotoniczne naszych ocenianych jednostek terytorialnych (od najlepszej do najgorszej lub odwrotnie), z jednoczesną możliwością oceny dystansu ich dzielącego. To ostatnie sprowadza się do przypisania każdej jednostce terytorialnej jednego wskaźnika oceny. Wskaźnik ten jest syntezą (pewnym uogólnieniem) zbioru wartości cech szczegółowych. Ta właśnie „synteza" (uogólnienie) jest celem/istotą omawianej metody i temu będą podporządkowane dalsze rozważania.

Tyle przypomnienia. Zakładam, że punkty 1 oraz 2 powyższego algorytmu zostały już odpowiednio przemyślane; tu jedynie przypomniane zostały niektóre kwestie odnoszące się jedynie do istoty metody. Nieco bardziej szczegółowo było to omawiane wcześniej.

Poniżej, zgodnie z tytułem podrozdziału, podejmujemy do omówienia problem „wyboru cech diagnostycznych – metoda grafu". Zanim jednak przejdziemy do zaprezentowania konkretnej procedury wyboru, najpierw relatywnie obszernie wyjaśnić należy istotę problemu, a więc jakie są przesłanki (podstawy) wyboru cech, który to wybór sprowadza się do odpowiedniej ich selekcji podporządkowanej pewnym wymogom.

Podstawą wyboru cech, inaczej przyczyną odpowiedniej ich selekcji, jest pewien dodatkowy wymóg[15], jaki stawia się zespołowi cech służących ocenie jednostek terytorialnych: wymogiem tym jest to, że cechy nie mogą lub nie powinny być ze sobą wysoko skorelowane. W rozwinięciu tego zagadnienia odwoływać się po trochu będziemy do znanej już wiedzy ze statystyki.

[14] Określenia: cecha, wskaźnik, miernik stosowane dalej będą zamiennie, co było już przedmiotem podanej informacji.
[15] Dodatkowy w stosunku do wymogów sformułowanych w punkcie 2 („dyskusja cech").

Wyobraźmy sobie, że zostajemy postawieni przed zadaniem dokonania oceny „m" jednostek terytorialnych z punktu widzenia jakiegoś zadanego kryterium. Wstępnie ustalone zostało, że kryterium to charakteryzowane jest „n" cechami[16]. Przykładowo ustaliliśmy, że do oceny warunków dla rozwoju turystyki w gminach województwa małopolskiego uwzględnić należy zestaw „n" cech spełniających wspominane wymogi (patrz punkt 2 algorytmu). W ujęciu ogólnym ilustruje to tabela 2 i zestawienie (1) w zapisie macierzowym.

Ponownie przypomnijmy sformułowany wyżej, „dodatkowy" wymóg stawiany zespołowi cech służących ocenie jednostek terytorialnych: cechy nie mogą lub nie powinny być ze sobą wysoko skorelowane. Zostawiając do dalszego wyjaśnienia, co to znaczy wysoko skorelowane, teraz wyjaśnienia wymagają dwa użyte określenia, tj. „nie mogą" i „nie powinny". Pierwsze wskazuje, że jest to wymóg konieczny, a drugie, że jest to jedynie rekomendacja, sugestia. Respektowanie jednego czy drugiego zależy od użytej/wybranej techniki/procedury agregowania cech szczegółowych w jeden syntetyczny wskaźnik oceny. Procedury te będą przedmiotem omawiania w dalszych fragmentach książki. W tym miejscu jedynie warto zasygnalizować, że jedna z procedur agregowania/syntetyzowania odwołuje się do formuł matematycznych, które wymagają, aby cechy były niezależne, a więc nieskorelowane. W przeciwnym razie otrzymywane wyniki wyliczeń mogą być obarczone błędem. Nie można jednak sformułować wymogu: „cechy wzajemnie nieskorelowane", gdyż wówczas zaistnieć może wysoce prawdopodobna sytuacja, iż należałoby odrzucić większość z nich, a niekiedy nawet wszystkie oprócz jakiejś jedynej. Dotyczyć to może zwłaszcza przypadków, gdy korelacje liczone są na dużej liczbie obiektów[17] i wtedy nawet niewielkie wartości współczynnika korelacji wskazywać mogą na istotność badanego związku. Stąd więc przyjmuje się: „nie są ze sobą wysoko skorelowane". Wskazuje to na zgodę, że przy relatywnie niedużym skorelowaniu cech wynik będzie na tyle nieznacznie zniekształcony, iż wiarygodność przeprowadzanej oceny będzie wystarczająca.

Druga grupa procedur nie stawia wymogu w zakresie skorelowania cech. Pomimo tego w odniesieniu do tej grupy procedur użyliśmy określenia „nie powinny". Jest to więc ogólna rekomendacja, które nie wynikają z istoty techniki agregowania. W rekomendacji tej odwołujemy się do dosyć oczywistego rozumowania. Jak już wiadomo, naszym zadaniem jest dokonanie jednoznacznej oceny względnej określonego zespołu jednostek terytorialnych. W tym miejscu warto bardzo wstępnie i ogólnie wyjaśnić, że istotą wspominanego agregowania/ syntetyzowania cech w jeden wskaźnik syntetyczny jest użycie takich procedur, które pozwalają na przekształcenie zbioru wielu cech w jeden ciąg/wektor ocen. Uświadamiamy sobie zatem, że każda z cech szczegółowych ma określony wpływ na wynik końcowy, czyli na wartość wskaźnika syntetycznego[18]. Wyobraźmy sobie teraz dwie cechy (oznaczmy je przez 1 oraz 2), pozostające w maksymalnym względem siebie skorelowaniu, czyli mamy wówczas do czynienia z jednoznaczną, prostą zależnością funkcyjną. Nietrudno zauważyć, że w takiej sytuacji, gdybyśmy dokonywali oceny (względnej) naszych jednostek terytorialnych, raz w oparciu o cechę (1), drugi raz w oparciu o cechę (2), wynik byłby identyczny. Przyjmijmy przykładowo, że w zbiorze cech oceniających poziom rozwoju województw w Polsce znalazły się dwie następujące: (1) wskaźnik zatrudnienia; (2) stopa bezrobocia. Abstrahując od realiów, przyjmijmy założenie,

[16] W tym miejscu należy pamiętać o kwestiach, które były już wcześniej omawiane, a dotyczące punktu 2 („dyskusja cech"), czyli wymogów stawianych cechom opisującym kryterium oceny.
[17] Co jest równoznaczne z dużą liczbą obiektów podlegających ocenie.
[18] Na marginesie prowadzonego rozważania, przypomnieć należy, że przyjmowane będzie założenie, iż siła wpływu każdej cech na wartość wskaźnika syntetycznego jest taka sama, tzn. że nie będziemy wprowadzać wag cech (wspominano o tym już wcześniej).

że poziom skorelowania tych cech jest maksymalny (współczynnik korelacji = -1). Jeżeli dokonalibyśmy oceny województw, najpierw w oparciu o cechę (1), później w oparciu o cechę (2), zauważylibyśmy, że ustalone dwa rankingi województw są identyczne. W pełni zbieżny powinien też być wynik rozumowania dotyczący „dystansu" dzielącego dane województwo od każdego innego. Co prawda, w pierwszym przypadku byłyby to wartości wskaźnika zatrudnienia, zaś w drugim – stopy bezrobocia; proporcje dzielące dowolne województwo od pozostałych nie uległyby jednak zmianie.

Powyższe rozumowanie prowadzi więc do wniosku, że nie ma sensu uwzględniać cech pozostających w tak dużym, wzajemnym skorelowaniu, ponieważ powtarzają te same informacje dotyczące rozkładu różnic między ocenianymi jednostkami terytorialnymi. Wniosek ten, rozciągniemy też na sytuację, w której dwie cechy, co prawda nie pozostają w zależności funkcyjnej (maksymalna wartość skorelowania), ale są wysoko ze sobą skorelowane. Również wówczas takie dwie cechy w wysokim stopniu powielają wspominane wyżej informacje dotyczące rozkładu różnic między ocenianymi jednostkami terytorialnymi.

Sumując, przyjmujemy ostatecznie, że cechy służące ocenie jednostek terytorialnych nie będą wysoko ze sobą skorelowane. Zajmijmy się teraz kwestią zdefiniowania określenia „wysoko skorelowane"[19]. Nazywać to będziemy poszukiwaniem wartości krytycznej współczynnika korelacji (r_k), tzn. takiej, dla której wartość powyżej r_k taktowane będą jako korelacja wysoka oraz tak samo, wartość poniżej $-r_k$ taktowane będą jako korelacja wysoka[20]. Przyjmujemy zatem, że o korelacji dopuszczalnej świadczą wartość współczynnika mieszczące się w przedziale $r_{xy} \in [-r_k, r_k]$.

Dla ustalenia wartości krytycznej (r_k), w praktyce wykorzystuje się dwa podejścia. **Pierwsze** jest niezwykle proste i polega na przyjęciu z góry określonej wartość, np. $r_k = 0{,}7$. Jest to możliwe i stosowane w sytuacji doświadczonego zespołu analityków prowadzących przedmiotowe badania, którzy – znając specyfikę przyjętego kryterium oceny – są kompetentni rozważać, na ile wynikająca z wyliczeń siła skorelowania dwóch cech jest, a na ile nie jest, wysoka. Należy w tym miejscu zauważyć, że ocena w tym względzie zależy od specyfiki rozważanej pary cech. Dla niektórych sytuacji skorelowanie rzędu, powiedzmy nawet 0,8 nie wydaje się wysokie[21], a w innym, przykładowo 0,4, jest już traktowane jako wysokie[22].

---

[19] Koniecznego przypomnienia (z przedmiotu statystyka) wymaga współczynnik korelacji, na bazie którego prowadzić będziemy dalsze rozważania. Odwoływać się będziemy do tzw. współczynnika korelacji prostej, z oznaczeniem r_{xy}. Mierzy on poziom skorelowania cechy „x" z cechą „y". Jak wiadomo, przyjmuje wartości z przedziału [-1, 1]. Użyte w nazwie określenie „korelacja prosta" informuje, że jego wyliczenie bazuje na założeniu, iż zależność między cechami jest liniowa, tzn. taka sama w całym zbiorze ich wartości. Podstawowy wzór definicyjny współczynnika korelacji prostej jest następujący:

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i - x_{śr})(y_i - y_{śr})}{\sqrt{\sum_{i=1}^{m}(x_i - x_{śr})^2}\sqrt{\sum_{i=1}^{m}(y_i - y_{śr})^2}}$$

gdzie:

r_{xy} – wartość korelacji między cechą „x" oraz „y" $r_{xy} \in [-1, 1]$.

[20] Należy pamiętać, że o korelacji wysokiej świadczą wartości współczynnika bliskie -1 oraz 1; korelacja niska to wartości bliskie 0, po obu stronach zera na osi liczbowej.

[21] Np. korelacja pomiędzy wskaźnikiem zatrudnienia w województwach a wartością PKB tych województw.

[22] Np. korelacja pomiędzy wprowadzanymi przez województwa przedsięwzięciami wspierającymi innowacyjność gospodarki a efektami tych przedsięwzięć.

Drugie podejście nawiązuje do badania statystycznej istotności współczynnika korelacji, i jest znacząco bardziej skomplikowane, ale musimy sobie z tym „skomplikowaniem" procedury poradzić. Dla ustalenia uwagi rozważmy taki przykład. Każdy z nas – a więc pracownik oraz student danej uczelni/wydziału (powiedzmy łącznie 10 tys. osób) – oddzielnie otrzymał ogólne zadanie, tj. dokonania oceny zależności korelacyjnej pomiędzy zjawiskiem „x" a zjawiskiem „y" (dla przykładu niech to będzie: x – udział ludności z wyższym wykształceniem; y – stopa bezrobocia). Każdy z nas, postępując oddzielnie, dla potrzeb realizacji zadania, na swój sposób zdefiniuje próbkę, na podstawie której wyliczy współczynnik korelacji r_{xy}. Ktoś może przyjąć, że wyliczenia będzie robił na podstawie województw w Polsce (m=16), inny na podstawie powiatów ziemskich województwa wielkopolskiego (m=31), jeszcze inny na podstawie miast w Polsce (dla stycznia 2021 r, m=954), a jeszcze ktoś inny na podstawie państw UE (w 2020, m=27) itd. Trzeba zauważyć, że oczywistym jest, iż nie mamy prawa oczekiwać dokładnie takich samych wartości r_{xy} otrzymywanych na podstawie różnie definiowanych próbek. Każdy z nas otrzyma nieco inny wynik. Niekiedy może to być różnica na dalszych miejscach po przecinku, ale również nie należy wykluczać większych różnic. Gdybyśmy teraz otrzymane wyniki nanieśli na oś liczbową w znanym przedziale [-1, 1], to z całą pewnością zauważylibyśmy, że rozkład naniesionych punktów na długości osi nie jest równomierny. W pewnym wąskim przedziale pojawia się bowiem wysoka ich koncentracja, zaś im dalej od tego zagęszczenia na prawo oraz na lewo, tym mniejsze na ogół jest zagęszczenie. Z obserwacji tej wyciągamy wniosek, że współczynnik korelacji ma jakiś rozkład, tzn. że dla danej (znanej) pary zjawisk (x, y) punkty reprezentujące współczynniki korelacji pojawiać się będą na osi liczbowej z określonym prawdopodobieństwem. Podkreślić należy, że rozkład ten jest znany na gruncie statystyki jako nauki i stablicowany. Jest to tzw. rozkład *t Studenta*. Konkretne wartości tego rozkładu, możliwe są do odczytania z odpowiednio opracowanych tablic statystycznych. Wartości te zależą od dwóch parametrów:

a) liczebności próby, na podstawie której wyliczany był współczynnik korelacji,
b) tzw. poziom istotności.

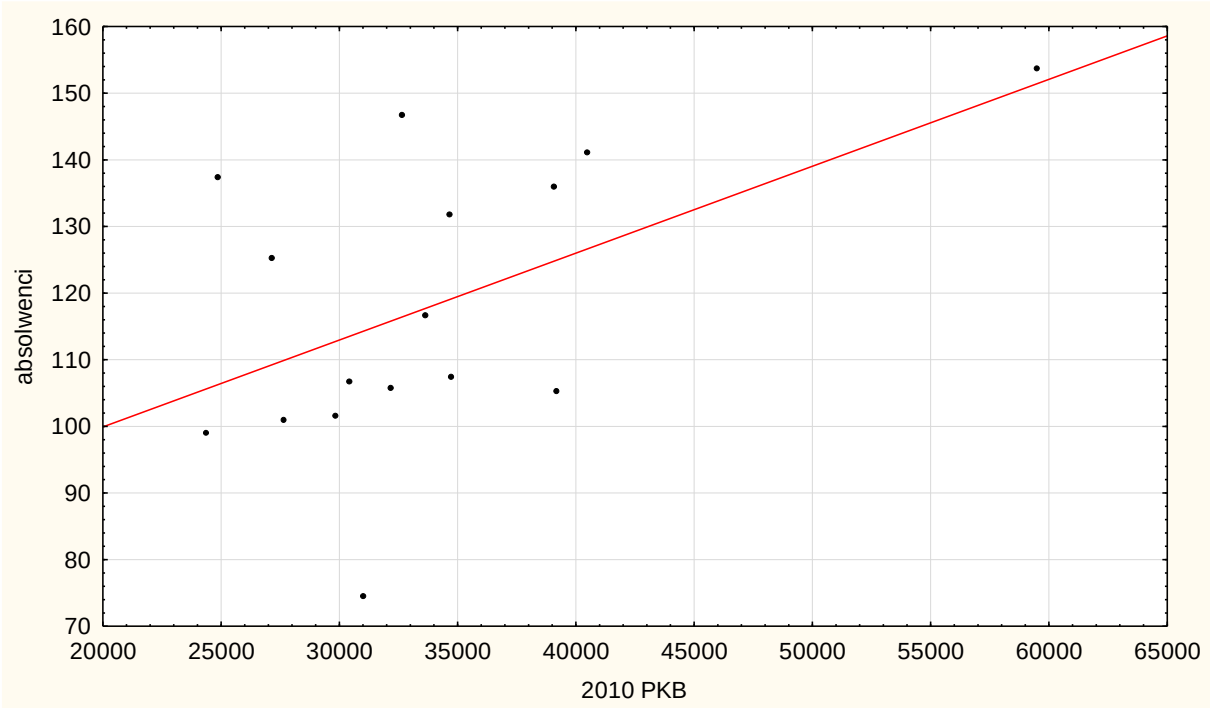
Rozważmy pokrótce jeden i drugi parametr.

Ad. a) Liczebność próby w tablicach rozkładu *t Studenta* tradycyjnie oznaczana jest przez ss (stopnie swobody). W przypadku tego rozważanego rozkładu ss = m-2. Przykładowo, jeżeli współczynnik korelacji wyliczany był na podstawie danych dotyczących województw, to ss = 14 (16-2). Warte uwagi jest dosyć oczywiste spostrzeżenie – im bardziej liczna jest próba, na podstawie której wyliczony został współczynnik korelacji, tym większa wydaje się wiarygodność otrzymanego wyniku mierzącego zależność pomiędzy dwoma zmiennymi. Nieco inaczej – im większa jest ss, tym bardziej jesteśmy skłonni przypisać wiarygodność wnioskowania o istniejącym związku korelacyjnym niższym nawet wartościom współczynnika. Obrazowo potwierdzają to wykresy ilustrowane rysunkiem[23] 1 i 2, sporządzone na podstawie danych tabeli 3.

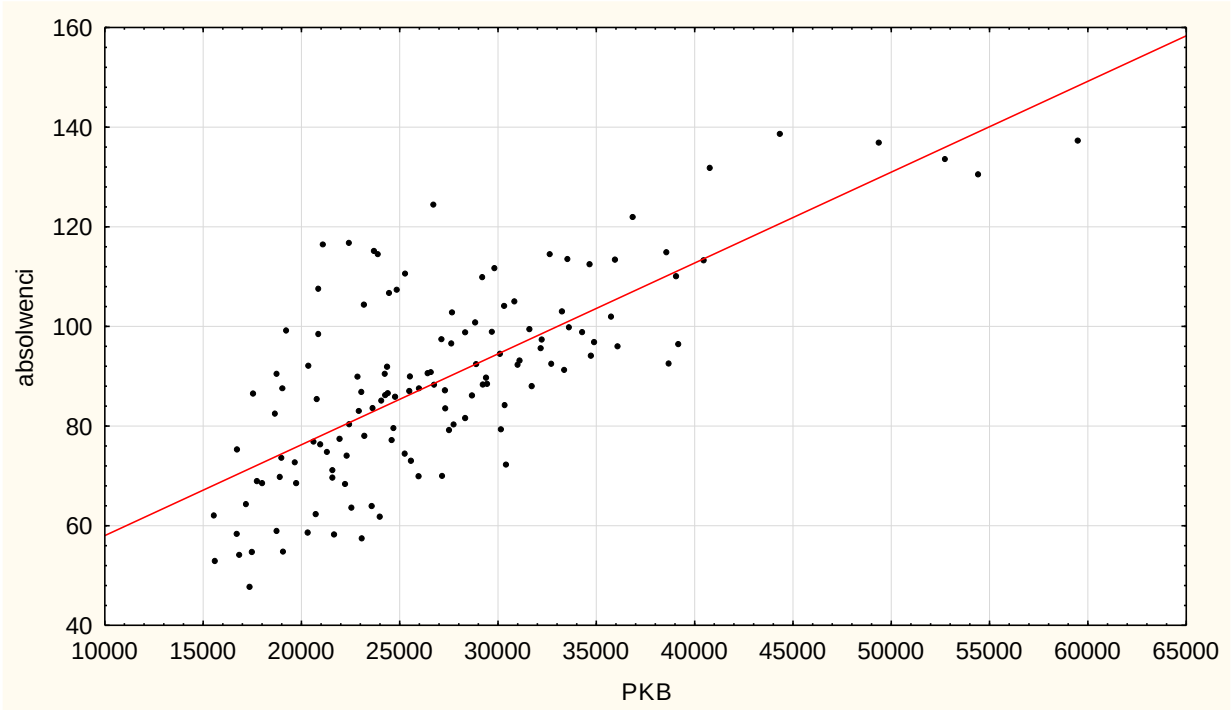
[23] Rysunki te ilustrują graficznie zależność tych samych zjawiska, ale na podstawi różnej liczebności próbki.

Tabela 3
Dane liczbowe do wykresu korelacyjnego (dla jednego roku)

Jednostka terytorialna	PKB na mieszkańca	Absolwenci uczelni/mieszkańca
Polska	33 867	405,19
dolnośląskie	40 510	394,14
kujawsko-pomorskie	32 213	376,41
lubelskie	24 892	457,56
lubuskie	31 043	422,48
łódzkie	34 695	479,82
małopolskie	32 681	381,32
mazowieckie	59 519	444,73
opolskie	30 455	351,49
podkarpackie	24 405	397,10
podlaskie	27 170	397,75
pomorskie	34 761	360,10
śląskie	39 220	387,84
świętokrzyskie	29 871	480,18
warmińsko-mazurskie	27 677	384,57
wielkopolskie	39 104	394,24
zachodniopomorskie	33 662	333,03



Rysunek 1. Wykres korelacyjny PKB – absolwenci szkół (dane dla jednego roku).
Źródło: opracowanie własne.



Rysunek 2. Wykres korelacyjny PKB – absolwenci szkół (dane dla 4 kolejnych lat).
Źródło: opracowanie własne.

Ad. b) Poziom istotności, na ogół oznaczany jako α, przyjmuje wartości z przedziału (0,1). Nie wchodząc zbyt głęboko w zawiłości jego istoty i upraszczając nieco rozumowanie, mierzy on przyjęte z góry prawdopodobieństwo popełnienia błędu związanego z wnioskowaniem dotyczącym weryfikacji hipotezy o istotności współczynnika korelacji. Rozsądek podpowiada więc, że zakładane z góry α nie może być relatywnie wysokie; a w praktyce na ogół przyjmuje się $\alpha \leq 0{,}1$.

Sformułujmy więc odpowiednie hipotezy:

$$H_0: \rho = 0; \qquad H_1: \rho \neq 0; \qquad (2)$$

Wyjaśnić należy, że ρ oznacza prawdziwą wartość współczynnika korelacji w nieznanej nam populacji generalnej[24]. Zastanówmy się chwilę, co oznacza powyższy zapis hipotez. Do wyjaśnienia posłuży nam wspominany wcześniej przykład: *mamy dokonać oceny zależności korelacyjnej pomiędzy udziałem ludności z wyższym wykształceniem (cecha x) a stopą bezrobocia (cecha y)*. Należy zwrócić uwagę, że zadanie to jest żądaniem oceny przedmiotowej zależności w ogóle (populacja generalna), a nie zależności np. w zbiorze województw w Polsce czy też powiatów ziemskich województwa wielkopolskiego. Hipoteza H_0 zakłada, że nie ma zależności (statystycznie istotnej) między zjawiskami x i y. Natomiast hipoteza alternatywna H_1 zależność taką zakłada. Rozstrzygnięcie, która z nich jest prawdziwa jest możliwe, ale jedynie na podstawie którejś z przyjętych próbek (nie znamy bowiem populacji generalnej). Wynik tego rozstrzygnięcia zależny będzie od wspominanych dwóch parametrów: liczebności próbki (ss) oraz założonego współczynnika istotności (α).

[24] Należy pamiętać, że wyliczenia współczynnika korelacji dokonywane były na podstawie odpowiedniej próby.

Do zweryfikowania hipotezy zerowej służy statystyka t postaci:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{m-2} \qquad (3)$$

Przy znanej liczbie stopni swobody (liczba ta jest zawsze znana) i przy założonym z góry poziomie istotności α, z tablic rozkładu *t Studenta* odczytujemy wartość statystyki t_α. Jeżeli okaże się, że $|t| \geq t_\alpha$, to hipotezę H_0 o braku korelacji między zmiennymi x i y odrzucamy na rzecz hipotezy alternatywnej. Gdyby natomiast $|t| < t_\alpha$, wówczas nie ma podstaw do odrzucenia tej hipotezy.

Tak w skrócie przebiega testowanie hipotezy o istotności współczynnika korelacji. Do ustalania wspominanej wyżej wartości krytycznej r_k, definiującej określenie „wysoka korelacja", wprost z tej procedury nie możemy skorzystać. Wprowadzimy zatem pewne rozumowanie modyfikujące.

Zauważmy, że jeżeli dla charakterystyki kryterium oceny przyjmiemy „n" cech i jeśli chcemy wyliczyć współczynniki korelacji wszystkich możliwych ich par, wówczas otrzymujemy macierz korelacji R o postaci[25]:

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} & \ldots. & r_{1n} \\ r_{21} & r_{22} & r_{23} & r_{24} & \ldots. & r_{2n} \\ r_{31} & r_{32} & r_{33} & r_{34} & \ldots. & r_{3n} \\ r_{41} & r_{42} & r_{43} & r_{44} & \ldots. & r_{4n} \\ \multicolumn{6}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\ r_{n1} & r_{n2} & r_{n3} & r_{n4} & \ldots. & r_{nn} \end{bmatrix} \qquad (4)$$

Skorzystajmy z odpowiedniego przekształcenia powyższej formuły „t" (wzór 3) w ten sposób, aby po jednej stronie równości otrzymać „r". Pomnóżmy więc obie strony tego równania przez $\sqrt{m-2}$, a więc:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{m-2} \Big/ \sqrt{1-r^2}$$

Dokonajmy kolejno odpowiednich wyliczeń:

$$t\sqrt{1-r^2} = r\sqrt{m-2}$$

[25] Należy zauważy, że jest to macierz symetryczna, gdyż korelacja zjawiska „x" ze zjawiskiem „y" jest tożsama z korelacją „y" z „x". Biorąc pod uwagę, że współczynnik korelacji dowolnej zmiennej z nią samą zawsze wynosi 1 (współczynniki na główne przekątnej macierzy R), to do wyliczenia pozostaje mam zawsze [n(n-1)]/2 różnych współczynników korelacji.

Po podniesieniu stronami do potęgi 2, a następnie dokonując prostych przekształceń, otrzymujemy:

$$t^2(1-r^2) = r^2(m-2)$$
$$t^2 - t^2r^2 = mr^2 - 2r^2$$
$$t^2 = t^2r^2 + mr^2 - 2r^2$$
$$t^2 = r^2(t^2 + m - 2)$$
$$r^2 = \frac{t^2}{t^2 + m - 2}$$

Ostatecznie wzór przybiera pożądaną postać:

$$r = \frac{t}{\sqrt{t^2+m-2}} \qquad (5)$$

Współczynnik r wyznaczony powyższą formuła jest poszukiwaną wartością krytyczną współczynnika korelacji (r_k). Pamiętając, że definiować on powinien wysoką wartość korelacji, wyznaczany musi być przy zdecydowanie niskim poziomie istotności α; najczęściej α = 0,05; 0,01 lub nawet 0,001.

Mając w pełni wyjaśnione określenie: *cechy nie powinny (nie mogą) być ze sobą wysoko skorelowane*, przejdziemy obecnie do rozważań związanych z procedurą takiego doboru zbioru cech, iż spełniać będą wskazany wymóg.

Z algorytmu przytoczonego na początku podrozdziału (punkt 3 algorytmu) wynika, że wybór cech diagnostycznych spełniających powyższy wymóg oprzeć możemy na dwóch metodach (technikach):

a) metodzie grafu,
b) metodzie dendrytu.

Zaczynamy od pierwszej z nich (metoda grafu). Przypomnijmy, co to jest graf. Graf jest konstrukcją graficzną, ujmującą zbiór wierzchołków (inaczej, węzłów), połączonych krawędziami (inaczej, wiązadłami lub łukami) dla zobrazowania określonej relacji w zbiorze obiektów. W naszym przypadku węzłami będą cechy, zaś wiązadłami związki korelacyjne, w jakich cechy wzajemnie pozostają. Przyjmujemy, że graficznie „węzły" obrazowane będą kółeczkiem z numerem, który odpowiadał będzie konkretnej cesze, zaś „związki korelacyjne"[26] obrazowane będą linią łączącą dany wierzchołek z innym wierzchołkiem, które to wierzchołki reprezentują dwie cechy o ustalonym współczynniku korelacji.

[26] Związki korelacyjne mierzone będą współczynnikiem korelacji prostej.

Algorytm postępowania przy stosowaniu omawianej metody wyboru cech jest następujący:

1) Wyznaczenie macierzy korelacji cech wyjściowych.
2) Ustalenie wartości krytycznej współczynnika korelacji:
    a) wybór wynikający ze z góry przyjętego założenia,
    b) na podstawie testu istotności.
3) Wyznaczenie macierzy przejścia.
4) Budowa grafu na podstawie macierz przejścia.
5) Ustalenie podgrafów i wybór węzłów reprezentantów.
6) Ustalenie ostatecznej listy cech diagnostycznych.

Ad. 3) Macierz przejścia jest macierzą binarną z wartościami „1", „0". Powstaje ona z przekształcenia macierzy korelacji w ten sposób, że tam, gdzie wartości współczynników korelacji co do bezwzględniej ich wartości są wyższe od r_k, wpisujemy 1, a w przeciwnym wypadku wpisujemy „0", czyli[27]:

$$\left.\begin{array}{ll} |r_{xy}| \le r_k & \text{wpisujemy } 0 \\ |r_{xy}| > r_k & \text{wpisujemy } 1 \end{array}\right\} \qquad (6)$$

W macierzy tej na przecięciu i-tego wiersza (i=1,2,…n) z j-tą kolumną (j=1,2,…n) znajdujemy więc albo „1", albo „0".

Ad. 4) Na podstawie macierzy przejścia budowany jest graf w ten sposób, że jeżeli na przecięciu i-tego wiersza (jest to cecha oznaczona nr „i") z j-tą kolumną (jest to cecha oznaczona nr „j") znajdujemy „1", wówczas łączymy te dwa wierzchołki linią (łukiem, wiązadłem); jeżeli natomiast znajdujemy „0", to te dwa wierzchołki nie są łączone.

Ad. 5) Postępując w powyższy sposób w odniesieniu do każdej cechy (kolejne wiersze macierzy przejścia), zbudujemy graf, który nie jest spójny, tzn. składa się z części (zespołów wierzchołków połączonych wiązadłami), które nie łączą się z innymi częściami grafu[28] (są względem siebie izolowane). Części te nazywać będziemy podgrafami. Liczba podgrafów wskazuje na liczbę cech, które zostaną wyselekcjonowane, jako spełniające warunek: *nie są ze sobą wysoko skorelowane*. Czyli z każdego podgrafu wybieramy jedną cechę „reprezentanta", według alternatywnej zasady[29]:

[27] Zauważmy, że z formuł 6 wynika, iż w przypadku gdyby któryś ze współczynników korelacji (r_{xy}) wyznaczonej macierzy był dokładnie równy r_k, wówczas w macierzy przejścia wpisujemy „0".

[28] Należy zauważyć, że teoretycznie ujmując, graf zbudowany na podstawie macierzy przejścia może okazać się spójny. Taka sytuacja nas jednak nie interesuje, dlatego z góry zostało przyjęte założone, że „nie jest spójny". Jak się przekonamy później, graf spójny oznaczałby sytuację, że ze zbioru „n" cech możemy wybrać tylko jedną z nich, co zaprzeczałoby istocie omawianej metody „syntetycznej oceny jednostek terytorialnych".

[29] Alternatywność zasad oznacza, że na którąś z nich musimy się zdecydować. Jednocześnie trzeba zauważyć, że trudno byłoby wskazać, która jest ważniejsza, a która mniej ważna. Zależy to od konkretnego rozważanego przypadku. Decyzję w tym względzie musi podjąć zespół (osoba) dokonujący badania.

a) najwyższa liczba łuków

W oparciu o tę zasadę dokonujemy wyboru tej cechy reprezentanta, dla której reprezentujący ją wierzchołek w danym podgrafie cechuje się największą liczbą krawędzi (łuków) połączeń.

b) najwyższe wartości skorelowania

Ta zasada wymaga ustalenia stopnia skorelowania każdej cechy podgrafu z wszystkimi pozostałymi cechami tego podgrafu. Wykorzystując więc tę zasadę, należałoby na podstawie wyznaczonej macierzy R dokonać zestawienia współczynników korelacji cech danego podgrafu, czyli sporządzić/zestawić dla każdego podgrafu oddzielną macierz korelacji odpowiednio „wykrojoną" z pełnej macierzy korelacji R. Pewną pomocą w rozstrzygnięciu zagadnienia wyboru cechy-reprezentanta według tego właśnie kryterium, może okazać się zsumowanie bezwzględnych wartość korelacji każdej kolumny (lub wiersza) tak ustalonej macierzy. Suma o najwyższej wartości może sugerować wybór.

c) w wyniku oceny merytorycznej przydatności cech

Ta zasada (kryterium wyboru) sprowadza się do z góry przyjętego rozstrzygnięcia, którą cechę uważamy za ważną z merytorycznego punktu widzenia; „merytorycznego", tzn. jej przydatności (znaczenia) w charakteryzowaniu kryterium dokonywanej oceny jednostek terytorialnych.

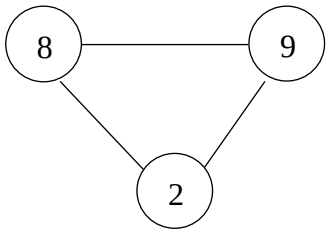
Warto zauważyć, że w pewnych przypadkach dwie pierwsze zasady wyboru cechy-reprezentanta zawodzą. Dotyczy to zwłaszcza sytuacji, kiedy podgraf składa się z dwóch lub trzech wierzchołków. Rozważmy poniższy przykład:

Podgraf jednoelementowy:

Podgraf dwuelementowy:

Podgraf trzyelementowy (wersja 1):

Podgraf trzyelementowy (wersja 2):



Jak zauważamy, w piewszym przypadku nie ma żadnych wątpliwości. Podgraf jest jednoelementowy, a więc jedyną cechą spełniającą waunek, jest cecha nr 1. W podgrafie dwuelementowym wybór cechy reprezentatna nastąpić musi w opaciu o zasadę merytorycznej oceny przydatności cechy (zasada c). W trzyelementowych podgrafach także, częściowo przynajmniej, zawodzą dwie pierwsze zasady.

Powyżej operowaliśmy określeniem „wybór cechy reprezentanta". Warto wyjaśnić istotę tego określenia. Przede wszystkim należy przypomnieć, że dwa dowolne wierzchołki w grafie są łączone wiązadłem tylko wtedy, gdy skorelowanie cech, które reprezentują, jest powyżej ustalonej wartości krytycznej współczynnika korelacji (r_k) – zob. procedurę sporządzania macierzy przejścia. Oznacza to, że w ramach każdego podgrafu cechy wprost lub pośrednio skorelowane są powyżej wartości krytycznej. Każdy podgraf cechuje się więc taką właściwością, że skupia cechy niespełniające wymogu *nie są ze sobą wysoko skorelowane*. Dlatego też z każdego podgrafu możemy wybierać jedną tylko cechę. Teoretycznie ujmując, każda cecha danego podgrafu spełnia warunek, że nie jest wysoko skorelowana z żadną cechą innych podgrafów (jak już zauważaliśmy, każdy podgraf jest izolowanym zespołem wierzchołków i łączących je łuków). Uświadamiając sobie z kolei, że wybierana cecha ma reprezentować wszystkie pozostałe cechy tego podgrafu, prowadzi to do wniosku, że powinna ona wykazywać jak najwyższe powiązania (skorelowanie) z tym cechami (zasada (a) oraz (b) powyższego wykazu zasad wyboru cech reprezentantów). W wyjaśnieniu tym zawiera się również uzasadnienie wcześniejszego stwierdzenia, że liczba podgrafów wskazuje na liczbę cech spełniających przytaczany warunek *nie są ze sobą wysoko skorelowane*.

Prezentowane powyżej wyjaśnienia dotyczą kluczowego etapu w procedurze omawianej metody wyboru cech (metodzie grafu). Dla ugruntowania i koniecznego zrozumienia istoty tego zagadnienia prześledzimy konkretny przykład. Pozwoli on również na naświetlenie wszystkich wcześniejszych etapów omawianego algorytmu.

Przykładowym zadaniem jest: Należy dokonać oceny poziomu rozwoju województw.

Warto zauważyć, że – w nawiązaniu do algorytmu postępowania w metodzie syntetycznej oceny jednostek terytorialnych, zaprezentowanego na wstępie podrozdziału 2.1 – w naszym przykładzie: terytorialnym systemem społeczno-gospodarczym jest Polska (punkt 1a); jednostkami podlegającymi ocenie są województwa (punkt 1b); kryterium oceny jest poziom rozwoju (punkt 1c).

Przyjmijmy, że w wyniku dyskusji cech – mierników szczegółowych (punkt 2 algorytmu), ustalono następujący zbiór wskaźników szczegółowych opisujących kryterium oceny:

1) PKB *per capita*.
2) Stopa bezrobocia.
3) Wskaźnik zatrudnienia (pracujący do ludności w wieku 15 lat i więcej).
4) Odsetek pracujących w usługach rynkowych.
5) Nakłady na B+R na mieszkańca.
6) Zatrudnienie w B+R na 1 000 osób aktywnych zawodowo.
7) Saldo migracji na 1 000 ludności (krajowe i zagraniczne).
8) Szkoły wyższe ogółem.
9) Studenci ogółem na 10 000 ludności.
10) Nauczyciele akademiccy na 1 0000 ludności.
11) Liczba studentów przypadająca na jednego profesora.
12) Teatry i instytucje muzyczne.
13) Plony zbóż.
14) Miejsca noclegowe ogółem na 100 tys. ludności.
15) Korzystający z noclegów na 1 000 ludności.
16) Turyści zagraniczni korzystający z noclegów na 1 000 ludności.
17) Odsetek gospodarstw domowych posiadających więcej niż jeden samochód osobowy.

18) Odsetek osób korzystających z Internetu z dostępem przez stałe łącze szerokopasmowe.
19) Odsetek osób korzystających z usług administracji publicznej za pomocą Internetu.
20) Przeciętny miesięczny dochód rozporządzalny na 1 osobę w wieloosobowych gospodarstwach domowych.

Wartości liczbowe, pochodzące z BDL GUS, ilustruje tabela 4.

Mając ustalony zestaw cech[30], przechodzimy teraz do rozwiązania punktu 3 algorytmu syntetycznej oceny jednostek terytorialnych, tj. *wyboru cech diagnostycznych*. Przyjmujemy, że wyboru tych cech dokonywać będziemy w oparciu o omówioną wyżej metodę grafu. Jak pamiętamy, zadaniem tego punktu algorytmu jest wybór takiego zespołu cech, które spełniają warunek: *nie są ze sobą wysoko skorelowane*.

Dla powyższego zbioru cech wyznaczona więc została macierz korelacji. Ilustruje ją tabela 5. Na marginesie głównego problemu rozważań, warto zauważyć symetryczność tej macierzy oraz jedynki na główne przekątnej.

Kolejnym krokiem jest wyznaczanie wartości krytycznej współczynnika korelacji r_k (punkt 2b algorytmu metody grafu). Liczba stopni swobody wynosi w naszym przypadku ss = 14. Jak już wiadomo, wartość parametru α (poziom istotności) musi być z góry przyjmowana przez badacza/analityka. Rozważymy cztery różne wersje parametru α: 0,1; 0,05; 0,01; 0,001.

Tabele 6 oraz 7 ujmują skopiowane z tablic statystycznych wartości statystyki t_α. Warto przypomnieć, że boczek tabel pokazuje stopnie swobody, zaś w główce ujęte są kolejne wartości poziomu istotności α. Dla powyższych, wybranych wartości α, odczytane z tablic rozkładu *t Studenta* wartości statystyki t_α wynoszą odpowiednio:

$t_{0,1} = 1,76$

$t_{0,05} = 2,145$

$t_{0,01} = 2,977$

$t_{0,001} = 4,140$

Podstawiając do wzoru (5) znane już wartości parametrów (tj. ss = 14 oraz kolejne wartości t_α), otrzymujemy wariantowe wartości krytyczne współczynnika korelacji[31]:

dla $t_{0,1} = 1,76$ $r_k = 0,426$

dla $t_{0,05} = 2,145$ $r_k = 0,497$

dla $t_{0,01} = 2,977$ $r_k = 0,623$

dla $t_{0,001} = 4,14$ $r_k = 0,742$

[30] Dla uproszczenia rozważań przyjmujemy, że wymienione cechy spełniają wymogi stawiane w punkcie 2 algorytmu syntetycznej oceny cech.

[31] Zdecydowanie rekomenduję przeprowadzenie przez każdą osobę samodzielnych wyliczeń potwierdzających prezentowane wartości.

Tabela 4

Dane liczbowe dotyczące cech charakteryzujących poziom rozwoju województw

Województwa	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
dolnośląskie	24,6	17,3	42,0	39,8	120	3,8	-0,6	34	573	30,5	86,3	17	43,9	1,58	601	163	41	62	22,7	778
kujawsko-pomorskie	21,6	19,4	44,7	34,0	55	3,2	-1,0	21	412	20,3	81,2	7	32,4	1,27	320	42	46	67	16,8	664
lubelskie	16,8	15,5	48,6	25,2	84	3,2	-2,2	20	478	28,7	84,8	6	29,6	0,89	249	42	50	62	16,4	672
lubuskie	21,6	20,0	44,6	38,9	35	1,9	-0,6	8	329	16,7	112,6	3	32,9	1,77	537	152	47	62	24,5	691
łódzkie	22,3	15,3	45,7	32,5	124	3,5	-0,7	27	508	27,7	75,2	14	27,2	0,60	239	37	47	64	18,4	769
małopolskie	20,7	11,6	47,4	36,4	224	6,8	1,0	34	611	36,4	78,8	20	33,3	1,89	765	273	49	67	27,7	732
mazowieckie	36,6	12,2	47,8	46,6	451	10,5	2,9	101	715	31,8	77,2	34	27,0	0,72	420	155	48	68	26,9	938
opolskie	20,8	16,6	44,7	32,9	27	2,4	-3,0	6	355	15,5	92,1	3	48,6	0,71	180	31	50	63	19,3	796
podkarpackie	16,9	16,4	45,2	28,2	53	1,6	-1,1	17	353	16,1	108,5	3	29,7	0,88	261	30	53	59	22,0	619
podlaskie	18,1	13,4	48,0	27,1	51	2,6	-1,6	19	441	25,1	86,9	5	26,8	0,98	331	72	44	55	19,7	725
pomorskie	23,6	16,0	43,5	40,8	132	5,2	0,5	28	443	26,6	82,8	15	32,3	3,89	613	134	42	68	24,8	801
śląskie	27,2	13,5	42,3	39,7	93	3,5	-1,9	43	420	21,1	100,0	24	34,7	0,75	326	56	47	67	27,3	793
świętokrzyskie	18,7	18,0	44,0	27,7	15	1,2	-1,7	14	498	14,1	116,6	3	27,2	0,64	229	19	48	60	13,5	621
warmińsko-mazurskie	18,8	23,6	41,6	33,0	46	2,0	-2,2	9	396	18,5	84,4	4	27,9	2,62	530	153	45	64	17,8	701
wielkopolskie	26,0	12,1	47,2	34,7	129	3,6	0,7	36	523	26,2	73,8	11	35,4	1,11	365	79	56	70	20,4	739
zachodniopomorskie	22,5	21,5	41,8	41,9	41	2,9	-1,0	21	516	24,9	99,0	12	34,9	6,21	911	281	40	66	20,3	763

Źródło: opracowanie własne na podstawie danych GUS (BDL).

Tabela 5

Macierz korelacji cech charakteryzujących rozwój społeczno-gospodarczy województwa

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	1,00	-0,39	0,03	0,82	0,79	0,78	0,71	0,89	0,62	0,40	-0,34	0,86	0,08	-0,03	0,15	0,24	-0,05	0,63	0,58	0,85
2	-0,39	1,00	-0,69	-0,03	-0,58	-0,56	-0,48	-0,57	-0,49	-0,56	0,40	-0,54	0,04	0,49	0,24	0,13	-0,46	-0,17	-0,46	-0,39
3	0,03	-0,69	1,00	-0,33	0,42	0,38	0,35	0,30	0,32	0,41	-0,38	0,09	-0,33	-0,50	-0,35	-0,20	0,56	-0,10	0,02	0,03
4	0,82	-0,03	-0,33	1,00	0,57	0,63	0,62	0,61	0,43	0,36	-0,17	0,74	0,25	0,43	0,60	0,62	-0,39	0,66	0,71	0,76
5	0,79	-0,58	0,42	0,57	1,00	0,97	0,85	0,94	0,81	0,69	-0,53	0,86	-0,19	-0,14	0,17	0,33	0,10	0,48	0,58	0,73
6	0,78	-0,56	0,38	0,63	0,97	1,00	0,84	0,90	0,79	0,74	-0,57	0,88	-0,11	-0,01	0,28	0,42	-0,03	0,55	0,62	0,77
7	0,71	-0,48	0,35	0,62	0,85	0,84	1,00	0,78	0,72	0,64	-0,42	0,71	-0,23	0,08	0,38	0,45	0,07	0,53	0,59	0,53
8	0,89	-0,57	0,30	0,61	0,94	0,90	0,78	1,00	0,79	0,59	-0,42	0,91	-0,18	-0,16	0,09	0,21	0,05	0,47	0,55	0,75
9	0,62	-0,49	0,32	0,43	0,81	0,79	0,72	0,79	1,00	0,79	-0,50	0,76	-0,14	0,02	0,32	0,44	-0,08	0,38	0,28	0,55
10	0,40	-0,56	0,41	0,36	0,69	0,74	0,64	0,59	0,79	1,00	-0,72	0,69	-0,06	0,12	0,44	0,52	-0,15	0,37	0,45	0,51
11	-0,34	0,40	-0,38	-0,17	-0,53	-0,57	-0,42	-0,42	-0,50	-0,72	1,00	-0,43	0,00	0,04	-0,08	-0,15	-0,03	-0,49	-0,12	-0,48
12	0,86	-0,54	0,09	0,74	0,86	0,88	0,71	0,91	0,76	0,69	-0,43	1,00	-0,03	0,01	0,30	0,39	-0,14	0,58	0,70	0,80
13	0,08	0,04	-0,33	0,25	-0,19	-0,11	-0,23	-0,18	-0,14	-0,06	0,00	-0,03	1,00	0,08	0,11	0,10	-0,05	0,17	0,16	0,23
14	-0,03	0,49	-0,50	0,43	-0,14	-0,01	0,08	-0,16	0,02	0,12	0,04	0,01	0,08	1,00	0,83	0,72	-0,62	0,27	0,11	0,10
15	0,15	0,24	-0,35	0,60	0,17	0,28	0,38	0,09	0,32	0,44	-0,08	0,30	0,11	0,83	1,00	0,96	-0,57	0,34	0,45	0,22
16	0,24	0,13	-0,20	0,62	0,33	0,42	0,45	0,21	0,44	0,52	-0,15	0,39	0,10	0,72	0,96	1,00	-0,46	0,34	0,51	0,33
17	-0,05	-0,46	0,56	-0,39	0,10	-0,03	0,07	0,05	-0,08	-0,15	-0,03	-0,14	-0,05	-0,62	-0,57	-0,46	1,00	0,09	-0,06	-0,24
18	0,63	-0,17	-0,10	0,66	0,48	0,55	0,53	0,47	0,38	0,37	-0,49	0,58	0,17	0,27	0,34	0,34	0,09	1,00	0,39	0,49
19	0,58	-0,46	0,02	0,71	0,58	0,62	0,59	0,55	0,28	0,45	-0,12	0,70	0,16	0,11	0,45	0,51	-0,06	0,39	1,00	0,58
20	0,85	-0,39	0,03	0,76	0,73	0,77	0,53	0,75	0,55	0,51	-0,48	0,80	0,23	0,10	0,22	0,33	-0,24	0,49	0,58	1,00

Źródło: obliczenia własne.

Tabela 6

Tablice rozkładu t Studenta – str. 1

ss \ α	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001	α \ ss
1	0,158	0,325	0;510	0,727	1,000	1,376	1,963	3,078	6,314	12;706	31,821	63,657	636,62	**1**
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	6,925	31,598	**2**
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941	**3**
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610	**4**
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,859	**5**
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959	**6**
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,405	**7**
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041	**8**
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781	**9**
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587	**10**
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437	**11**
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318	**12**
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221	**13**
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140	**14**
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073	**15**

Źródło: *Statystyka Matematyczna. Modele i zadania* (s. 311), J. Greń, 1982, Warszawa: Państwowe Wydawnictwo Naukowe.

Tabela 7
Tablice rozkładu t Studenta – str. 2

ss \ α	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001	α \ ss
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2.120	2,583	2,921	4,015	**16**
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965	**17**
18	0,127	0,257	0,392	0,534	0,688	0,8ó2	1,067	1,330	1,734	2,101	2,552	2,878	3,922	**18**
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883	**19**
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850	**20**
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819	**21**
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792	**22**
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767	**23**
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064.	2,492	2,797	3,745	**24**
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725	**25**
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707	**26**
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,522	2,473	2,771	3,690	**27**
28	0,127	0,256	0,389	0,530	0,683	0,855	1,036.	1,313	1,701	2,048	2,467	2,763	3,674	**28**
29	0,127	0,256	0,389	0,530	0,683	0,854	1,033	1,311	1,699	2,045	2,462	2,756	3,659	**29**
30	0,127	0,256	0,389	0,530	0,683	0,854	1,033	1,310	1,697	2,042	2,457	2,750	3,646	**30**
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,701	3,551	**40**
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460	**60**
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373	**120**
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291	∞

Źródło: *Statystyka Matematyczna. Modele i zadania* (s. 312), J. Greń, 1982, Warszawa: Państwowe Wydawnictwo Naukowe.

Załóżmy, że do dalszych rozważań wybieramy wartość krytyczną, wyliczoną ze wzoru (5) dla poziomu istotności $\alpha = 0{,}01$[32]. Z powyższej zamieszczonych wyliczeń wynika, że wartość krytyczna współczynnika korelacji wynosi wówczas $\mathbf{r_k = 0{,}623}$. Przyjmując tę wartość, na podstawie macierzy korelacji skonstruować należy macierz przejścia według zasad wcześniej prezentowanych[33]. Macierz tę ujmuje tabela 8. Dysponując macierzą przejścia, konstruujemy graf powiązań w zbiorze przyjętych cech. Budowę grafu zaczynamy od cechy nr „1" (pierwszy wiersz macierzy przejścia), a więc rysujemy kółeczko z nr „1" w środku. Następnie kółeczko to łączymy z wszystkimi tymi cechami (graficznie, z kółeczkami z odpowiednimi numerami w środku), na które wskazują jedynki w pierwszym wierszu macierzy przejścia. Dalej, przechodzimy do cechy nr „2", nr „3" itd., aż do cechy ostatniej (kolejne wiersze macierzy przejścia), wykonując analogiczne czynności. Uwaga: przechodząc do rysowania połączeń kolejnych cech ze sobą, należy bezwzględnie pamiętać, że dla danej cechy jest tylko jeden wierzchołek (kółeczko) ją reprezentujący.

Poniżej zamieszczony jest rysunek (3) obrazujący łączenia pierwszych siedmiu wierzchołków. Stwarza on możliwość prześledzenia budowy grafu dla naszego przykładu, począwszy od cechy nr 1, do cechy nr 7. Z uwagi na zbyt duże zagęszczenie wierzchołków i łuków, a w konsekwencji małą czytelność sieci powiązań, zrezygnowano z budowy grafu w pełnym składzie wierzchołków.

Abstrahując od realiów naszego przykładu, gdyby założyć, że było tylko 7 cech wejściowych (tzn. gdyby przedstawiony graf był kompletny), wtedy do dalszych wyliczeń przyjąć należałoby dwie cechy-reprezentanty. Biorąc pod uwagę mniejszy liczebnie podgraf, jedną z wybranych cech byłaby wtedy cecha nr 2 albo nr 3. Decyzja może być podjęta jedynie na podstawie zasady merytorycznego doboru cech (zob. punkt c przedstawianych wyżej zasad doboru cech-reprezentantów). Ponieważ pierwsza cecha to PKB *per capita*, zaś druga to stopa bezrobocia, z punktu widzenia kryterium oceny, którym jest poziom rozwoju województw, ważniejszy wydaje się wskaźnik PKB. Z drugiego podgrafu (bardziej licznego) należałoby wybrać cechę nr 6, ponieważ wierzchołek ją reprezentujący legitymuje się największą liczbą krawędzi, a więc należy sądzić, że cecha ta najlepiej będzie reprezentować w dokonywanej ocenie województw wszystkie inne cechy należące do tego podgrafu.

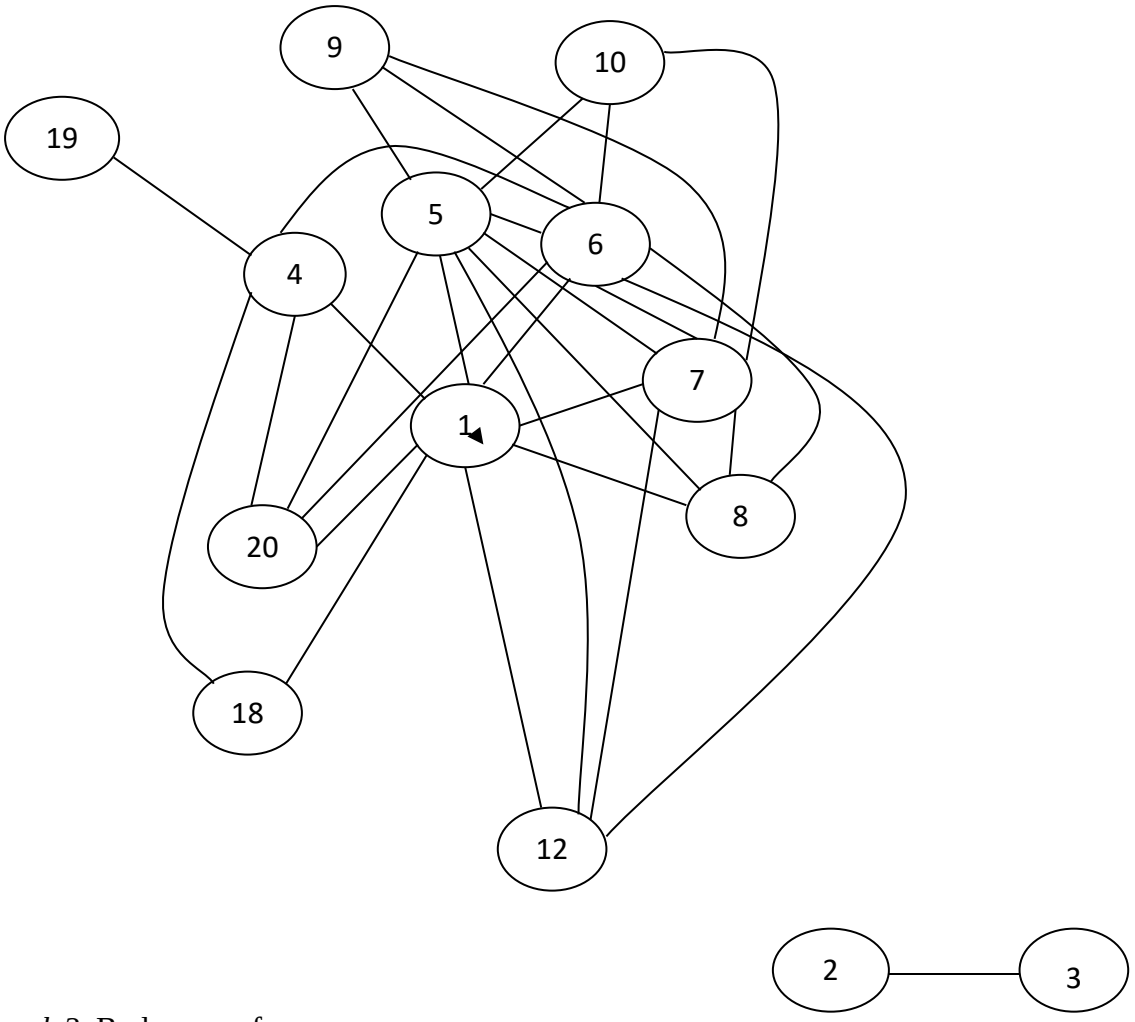
[32] Należy przypomnieć, że każdorazowo zespół realizujący omawiane zadanie dokonać musi wyboru wartości parametru α. Wiemy już, że powinna to być wartość równa lub raczej poniżej 0,1.

[33] Koniecznie należy pamiętać, że o wysokiej korelacji informują zarówno wysokie, dodatnie wartości współczynnika korelacji, jak i wysokie, co do modułu wartości ujemne (patrz: zaznaczona wartość bezwzględna współczynnika r_{xy} w ujętych powyżej formułach 6).

Tabela 8
Macierz przejścia

Cechy	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	0	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	1	0	1
2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	1
5	1	0	0	0	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	1
6	1	0	0	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	1
7	1	0	0	0	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0
8	1	0	0	0	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	1
9	0	0	0	0	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0
10	0	0	0	0	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
12	1	0	0	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
18	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
19	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
20	1	0	0	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1

Źródło: opracowanie własne.



Rysunek 3. Budowa grafu.
Źródło: opracowanie własne.

Pytania/zadania kontrolne

1. Dlaczego cechy będące podstawą dokonywanej oceny nie powinny być ze sobą wysoko skorelowane?
2. Co to znaczy „cechy wysoko skorelowane”?
3. Jak należy rozumieć ocenę względną badanych obiektów (jednostek terytorialnych?
4. Jakie dwa parametry przesądzają o statystycznej istotności współczynnika korelacji?
5. O czym przesądza „wartość krytyczna współczynnik korelacji” w naszej procedurze doboru cech diagnostycznych?
6. Graf niespójny składa się z kilku podgrafów. Jaką ma to wartość informacyjną?; czyli o czym informują części grafu?
7. Jakie przesłanki przesądzają o wyborze z danego podgrafu cechy, która reprezentować ma cechy pozostałe tego podgrafu?
8. Przyjmując przykładowo, że rk = 0,5, jakie wartości współczynników korelacji cech przesądzać będą o wyborze cech diagnostycznych?

2.3. Metody syntetycznej oceny jednostek terytorialnych – wybór cech diagnostycznych z zastosowaniem metody dendrytu

Przedmiotem rozważań w tym podrozdziale jest punkt 3b algorytmu syntetycznej oceny jednostek terytorialnych (patrz tabela 1 w podrozdziale 2.1). Poprzedni podrozdział traktował o procedurze wyboru cech diagnostycznych w oparciu o metodę grafu. Teraz zajmiemy się procedurą alternatywną, a więc wyborem cech diagnostycznych z wykorzystaniem nieco innej metody, a mianowicie metody dendrytu. Przypomnijmy właściwość, jaką legitymować się mają cechy będące podstawą konstrukcji wskaźnika syntetycznego: *cechy nie mogą lub nie powinny być ze sobą wysoko ze sobą skorelowane*[34]. Zadanie stojące przed nami jest zatem następujące: ze zbioru „n" cech wyjściowych należy wyselekcjonować podzbiór cech legitymujących się tym, że żadna para z nich nie będzie wykazywać względem siebie wysokiego skorelowania.

Dendryt jest szczególną postacią grafu, a więc taką, która nie posiada pętli (sprzężeń zwrotnych). Zatem dendryt, podobnie jak graf, jest konstrukcją graficzną ujmującą zbiór wierzchołków (inaczej, węzłów) połączonych krawędziami (inaczej, wiązadłami, krawędziami lub łukami) obrazującymi określone relacje w zbiorze obiektów. W naszym przypadku węzłami będą cechy, zaś wiązadłami związki korelacyjne, w jakich cechy wzajemnie pozostają. Również analogicznie jak w grafie przyjmujemy, że graficznie „węzły" obrazowane będą kółeczkiem z numerem, który odpowiadał będzie konkretnej cesze, zaś „związki korelacyjne"[35] obrazowane będą linią łączącą dany wierzchołek z innym wierzchołkiem, które to wierzchołki reprezentują dwie cechy o ustalonym współczynniku korelacji.

Algorytm postępowania z wykorzystaniem metody dendrytu dla wyboru cech spełniających powyżej sformułowany warunek jest następujący:

1) Wyznaczenie macierzy korelacji cech wyjściowych.
2) Ustalenie najwyższych wartości współczynników korelacji każdej cechy.
3) Interpretując każdą cechę jako wierzchołek dendrytu, zaś współczynnik korelacji jako jego wiązadło (łuk), połączyć wierzchołki na podstawie największego skorelowania cech.
4) W przypadku, gdy otrzymany dendryt nie jest spójny, należy postępowanie powtórzyć w ten sposób, aby dla danego, izolowanego fragmentu dendrytu, wyszukać jego połączenie łukiem z innym fragmentem na zasadzie najwyższej wartości współczynnika korelacji.
5) Ustalamy wartość krytyczną współczynnika korelacji (jak przy grafie).
6) Wszystkie wiązadła reprezentujące współczynniki korelacji o wartości niższej od krytycznej zostają usunięte.
7) Powstałe w ten sposób części dendrytu reprezentują odpowiednie grupy cech.
8) Z każdej grupy należy wybrać cechę reprezentanta, według alternatywnej zasady[36]:
    a) najwyższa liczba łuków,
    b) najwyższe wartości skorelowania,
    c) w wyniku oceny merytorycznej przydatności cech.
9) Ustalić ostateczną listę cech.

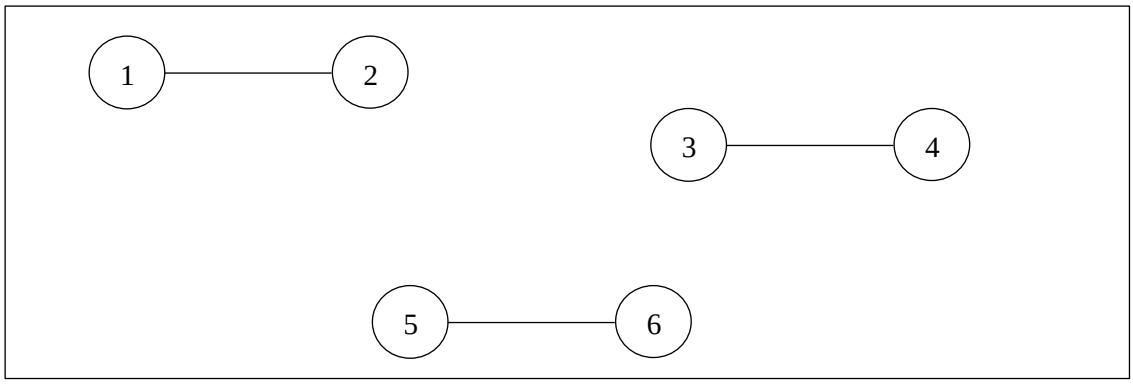
[34] Co to znaczy „nie mogą", „nie powinny", a także „wysoko skorelowane" – było wyjaśniane przy okazji omawiania doboru cech w oparciu o metodę grafu.
[35] Związki korelacyjne mierzone będą współczynnikiem korelacji prostej.
[36] Może się zdarzyć, że grupa będzie jednoelementowa. Wówczas nie ma żadnego problemu wyboru. Cechą spełniającą warunek jest ta, na którą wskazuje dany (pojedynczy) wierzchołek.

Punkt pierwszy z powyższego algorytmu został już wyjaśniony w poprzednim podrozdziale.

Ad. 2) Punkt ten sprowadza się do tego, że w każdym wierszu macierzy korelacji (również w każdej kolumnie, gdyż jest to macierz symetryczna) szukamy wartości, co do modułu najwyższej. Interpretując nieco inaczej, oznacza to, że dla każdej cechy (kolejne wiersze lub kolumny macierzy korelacji) szukamy takiej cechy, z którą jest ona w najwyższym stopniu skorelowana.

Ad. 3) Znajdując taką cechę, reprezentujące je wierzchołki łączymy wiązadłami. Postępujemy tak kolejno w stosunku do wszystkich wierszy (kolumn) macierzy korelacji. W efekcie otrzymujemy sieć powiązań będących najczęściej dendrytem niespójnym[37], tzn. taką jego postacią, że pojawiają się części izolowane (bez połączeń z innymi częściami).

Ad. 4) Następnym krokiem (punkt 4) jest „uspójnienie" dendrytu, tj. w oparciu o zasadę najwyższego skorelowania poszukiwać należy sposobu połączenia wiązadłem danego izolowanego fragmentu dendrytu z innym fragmentem na zasadzie najwyższego skorelowania. Zobrazujmy to prostym przykładem. Dla uproszczenia rozważań przyjmijmy, że otrzymaliśmy trzy izolowane dwuelementowe części dendrytu o postaci przedstawione na rysunku 4.



Rysunek 4. Części niespójnego, przykładowego dendrytu.
Źródło: opracowanie własne.

Załóżmy, że rozważamy podłączenie/uspójnienie pierwszej izolowanej części dendrytu (wierzchołki 1 i 2). Należy zauważyć, że może ona być „podłączona" do któregoś wierzchołka pozostałych części, wiązadłem wychodzącym albo z wierzchołka 1, albo z wierzchołka 2. Szukamy zatem najwyższego skorelowania cechy reprezentowanej przez wierzchołek 1 z którąkolwiek z pozostałych cech (oprócz cechy 2, gdyż z tą jest już połączona). Następnie szukamy najwyższego skorelowania cechy reprezentowanej przez wierzchołek 2 z którąkolwiek z pozostałych cech (oprócz cechy 1). Z dwóch ustalonych wariantów wybieramy ten, który legitymuje się wyższym poziomem skorelowania. Identycznie postępujemy w odniesieniu do pozostałych – izolowanych – część dendrytu, aż otrzymamy w pełni spójny jego obraz. W tym miejscu warto przedstawić praktyczną podpowiedź, a mianowicie – pamiętać należy, że uspójniniane warto rozpocząć od części najmniej licznej w wierzchołki. Pozwoli to zmniejszyć pracochłonność realizacji zadania.

[37] W przeciwieństwem do grafu, otrzymanie dendrytu spójnego byłoby ze wszech miar pożądane w omawianej metodzie wyboru cech.

Ad. 5) Kwestia ustalania wartości krytycznej współczynnika korelacji jest identyczna jak w przypadku metody grafu, a więc albo poprzez z góry przyjętą wartość krytyczną tego współczynnika (r_k), albo też poprzez odpowiednie wyliczenia bazujące na idei weryfikacji jego istotności (było to przedmiotem szerokich wyjaśnień we wcześniejszym podrozdziale).

Ad. 6), Ad. 7) i Ad. 8) Mając sporządzony dendryt spójny, dokonujemy jego „rozspójnienia" (w żadnym wypadku nie jest to proste odwrócenie wcześniej sygnalizowanego „uspójninia"). Rezygnujemy mianowicie z wiązadeł reprezentujących powiązania korelacyjne poniżej wartości krytycznej, a więc dla takich wartości r, które należą do przedziału $[-r_k; r_k]$). Otrzymamy w ten sposób części dendrytu, których liczba wskazuje na liczbę wyselekcjonowanych cech. Z każdej części dendrytu wybieramy jedną cechę-reprezentanta w oparciu o zasady analogiczne jak w odniesieniu do omówionych już w przypadku metody grafu. Pozwala nam to na ustalenie ostatecznej listy cech będących podstawą przeprowadzanej dalej oceny jednostek terytorialnych.

Celem przećwiczenia omawianej metody selekcji cech, prześledzimy konkretny przykład. W założeniu jest on identyczny do omawianego już przykładu ilustrującego metodę grafu. Naszym zadaniem jest zatem dokonanie oceny poziomu społeczno-gospodarczego rozwoju województw. Cała faktografia związana z tym zadaniem została ukazana w podrozdziałach poprzednich. Warto jedynie przypomnieć zestaw cech wyjściowych reprezentujących kryterium oceny (poziom rozwoju województw):

1) PKB *per capita*.
2) Stopa bezrobocia.
3) Wskaźnik zatrudnienia (pracujący do ludności w wieku 15 lat i więcej).
4) Odsetek pracujących w usługach rynkowych.
5) Nakłady na B+R na mieszkańca.
6) Zatrudnienie w B+R na 1 000 osób aktywnych zawodowo.
7) Saldo migracji na 1 000 ludności (krajowe i zagraniczne).
8) Szkoły wyższe ogółem.
9) Studenci ogółem na 10 000 ludności.
10) Nauczyciele akademiccy na 10 000 ludności.
11) Liczba studentów przypadająca na jednego profesora.
12) Teatry i instytucje muzyczne.
13) Plony zbóż.
14) Miejsca noclegowe ogółem na 100 tys. ludności.
15) Korzystający z noclegów na 1 000 ludności.
16) Turyści zagraniczni korzystający z noclegów na 1 000 ludności.
17) Odsetek gospodarstw domowych posiadających więcej niż jeden samochód osobowy.
18) Odsetek osób korzystających z Internetu z dostępem przez stałe łącze szerokopasmowe.
19) Odsetek osób korzystających z usług administracji publicznej za pomocą Internetu.
20) Przeciętny miesięczny dochód rozporządzalny na 1 osobę w wieloosobowych gospodarstwach domowych.

Zacznijmy więc od punkt 1 powyższego algorytmu budowy dendrytu. Macierz korelacji 20-stu cech ujmuje tabela 9[38].

W macierzy tej w każdym jej wierszu, zgodnie z punktem 2 algorytmu metody dendrytu, zaznaczono poprzez kolorowanie najwyższe – co do modułu – współczynniki korelacji (wartości głównej przekątne są ignorowane). Uwzględniając kolejne wiersze macierzy korelacji, na tej podstawie zbudowany został niespójny dendryt, ilustrowany rysunkiem 5.

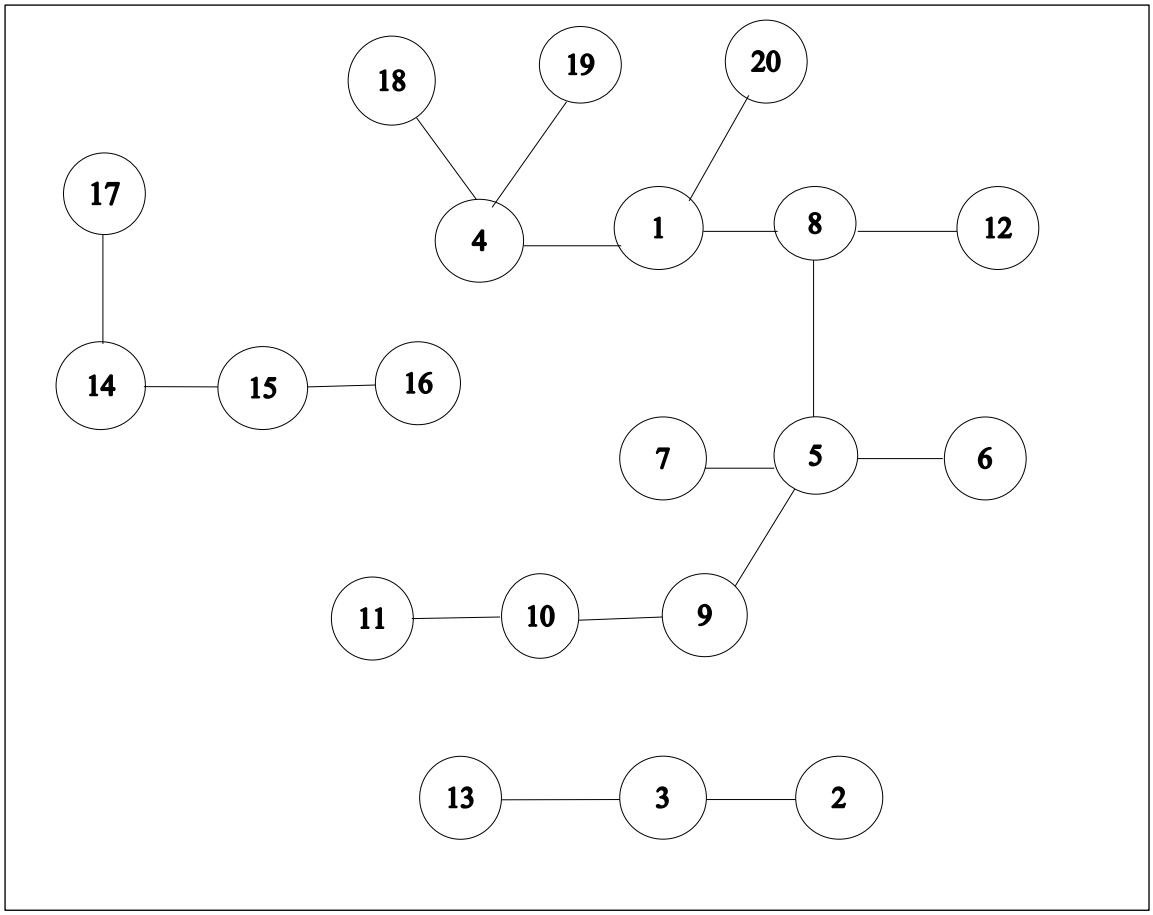
[38] Zdecydowano ponownie zamieścić tę macierzy z uwagi na potrzebę zaznaczenia w każdym jej wierszu (kolumnie) wartości ukazującej najwyższy poziom skorelowania.

Tabela 9

Macierz korelacji cech charakteryzujących rozwój społeczno-gospodarczy województwa – z zaznaczeniem wartości ekstremalnych

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	1,00	-0,39	0,03	0,82	0,79	0,78	0,71	**0,89**	0,62	0,40	-0,34	0,86	0,08	-0,03	0,15	0,24	-0,05	0,63	0,58	0,85
2	-0,39	1,00	**-0,69**	-0,03	-0,58	-0,56	-0,48	-0,57	-0,49	-0,56	0,40	-0,54	0,04	0,49	0,24	0,13	-0,46	-0,17	-0,46	-0,39
3	0,03	**-0,69**	1,00	-0,33	0,42	0,38	0,35	0,30	0,32	0,41	-0,38	0,09	-0,33	-0,50	-0,35	-0,20	0,56	-0,10	0,02	0,03
4	**0,82**	-0,03	-0,33	1,00	0,57	0,63	0,62	0,61	0,43	0,36	-0,17	0,74	0,25	0,43	0,60	0,62	-0,39	0,66	0,71	0,76
5	0,79	-0,58	0,42	0,57	1,00	**0,97**	0,85	0,94	0,81	0,69	-0,53	0,86	-0,19	-0,14	0,17	0,33	0,10	0,48	0,58	0,73
6	0,78	-0,56	0,38	0,63	**0,97**	1,00	0,84	0,90	0,79	0,74	-0,57	0,88	-0,11	-0,01	0,28	0,42	-0,03	0,55	0,62	0,77
7	0,71	-0,48	0,35	0,62	**0,85**	0,84	1,00	0,78	0,72	0,64	-0,42	0,71	-0,23	0,08	0,38	0,45	0,07	0,53	0,59	0,53
8	0,89	-0,57	0,30	0,61	**0,94**	0,90	0,78	1,00	0,79	0,59	-0,42	0,91	-0,18	-0,16	0,09	0,21	0,05	0,47	0,55	0,75
9	0,62	-0,49	0,32	0,43	**0,81**	0,79	0,72	0,79	1,00	0,79	-0,50	0,76	-0,14	0,02	0,32	0,44	-0,08	0,38	0,28	0,55
10	0,40	-0,56	0,41	0,36	0,69	0,74	0,64	0,59	**0,79**	1,00	-0,72	0,69	-0,06	0,12	0,44	0,52	-0,15	0,37	0,45	0,51
11	-0,34	0,40	-0,38	-0,17	-0,53	-0,57	-0,42	-0,42	-0,50	**-0,72**	1,00	-0,43	0,00	0,04	-0,08	-0,15	-0,03	-0,49	-0,12	-0,48
12	0,86	-0,54	0,09	0,74	0,86	0,88	0,71	**0,91**	0,76	0,69	-0,43	1,00	-0,03	0,01	0,30	0,39	-0,14	0,58	0,70	0,80
13	0,08	0,04	**-0,33**	0,25	-0,19	-0,11	-0,23	-0,18	-0,14	-0,06	0,00	-0,03	1,00	0,08	0,11	0,10	-0,05	0,17	0,16	0,23
14	-0,03	0,49	-0,50	0,43	-0,14	-0,01	0,08	-0,16	0,02	0,12	0,04	0,01	0,08	1,00	**0,83**	0,72	-0,62	0,27	0,11	0,10
15	0,15	0,24	-0,35	0,60	0,17	0,28	0,38	0,09	0,32	0,44	-0,08	0,30	0,11	0,83	1,00	**0,96**	-0,57	0,34	0,45	0,22
16	0,24	0,13	-0,20	0,62	0,33	0,42	0,45	0,21	0,44	0,52	-0,15	0,39	0,10	0,72	**0,96**	1,00	-0,46	0,34	0,51	0,33
17	-0,05	-0,46	0,56	-0,39	0,10	-0,03	0,07	0,05	-0,08	-0,15	-0,03	-0,14	-0,05	**-0,62**	-0,57	-0,46	1,00	0,09	-0,06	-0,24
18	0,63	-0,17	-0,10	**0,66**	0,48	0,55	0,53	0,47	0,38	0,37	-0,49	0,58	0,17	0,27	0,34	0,34	0,09	1,00	0,39	0,49
19	0,58	-0,46	0,02	**0,71**	0,58	0,62	0,59	0,55	0,28	0,45	-0,12	0,70	0,16	0,11	0,45	0,51	-0,06	0,39	1,00	0,58
20	**0,85**	-0,39	0,03	0,76	0,73	0,77	0,53	0,75	0,55	0,51	-0,48	0,80	0,23	0,10	0,22	0,33	-0,24	0,49	0,58	1,00

Źródło: obliczenia własne.

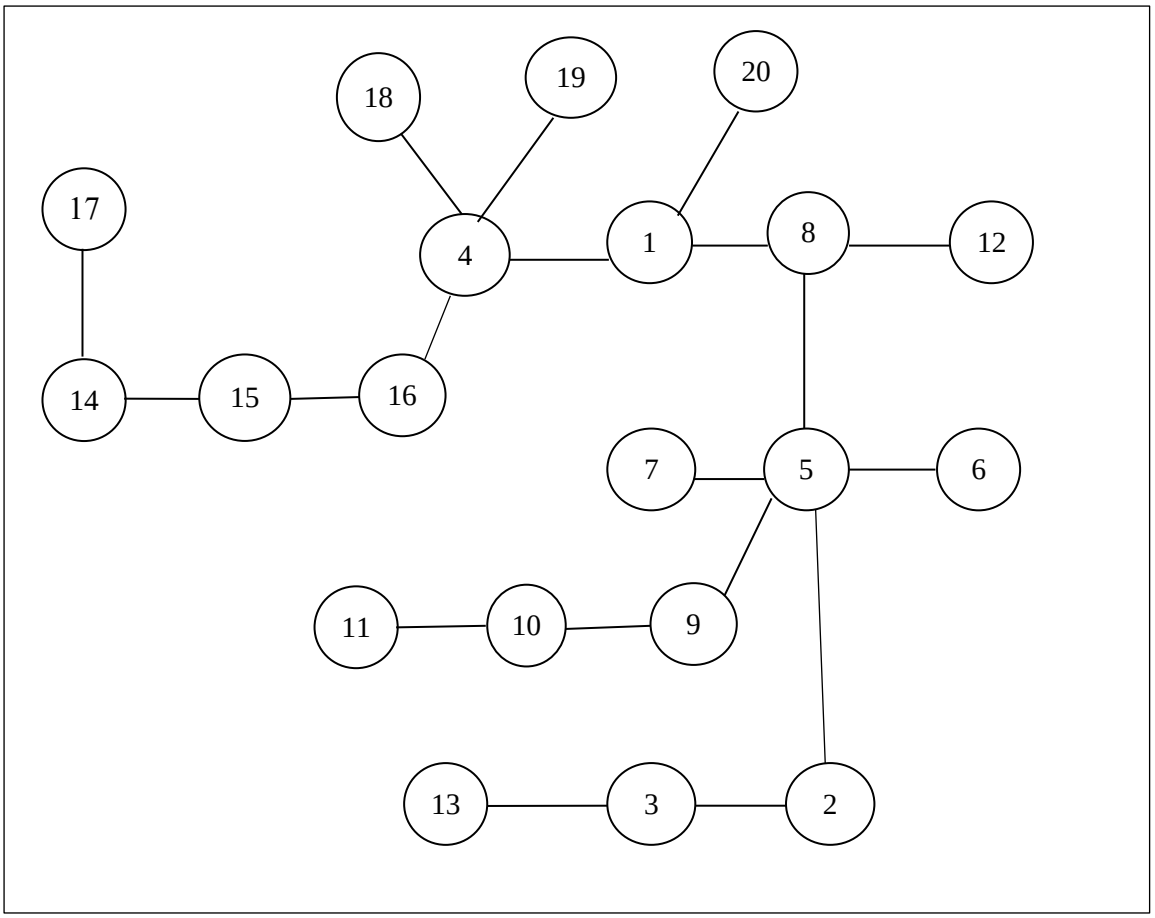


Rysunek 5. Dendryt niespójny.
Źródło: opracowanie własne.

Nietrudno zauważyć, że dendryt, po wykonaniu czynności wskazywanych punktem 3 algorytmu, składa się z trzech izolowanych części. Uspójnianie zacznijmy – zgodnie ze wskazanymi już wcześniej rekomendacjami – od najmniej licznej w wierzchołki części. Posiłkując się informacjami zawartymi w macierzy korelacji, zauważamy, że:

- ✓ cecha reprezentowana przez wierzchołek nr 2 najwyższe skorelowanie wykazuje z cechą nr 5 (r = -0,58);
- ✓ cecha reprezentowana przez wierzchołek nr 3 najwyższe skorelowanie wykazuje z cechą nr 17 (r = 0,56);
- ✓ cecha reprezentowana przez wierzchołek nr 13 najwyższe skorelowanie wykazuje z cechą nr 4 (r = 025).

Ostatecznie więc – na zasadzie najwyższego skorelowania – łączymy wierzchołek 2 z wierzchołkiem 5. Analogiczne postępowanie w odniesieniu do kolejnej części prowadzi do ustalenia, że wierzchołek 16 łączy się z wierzchołkiem nr 4. W efekcie otrzymujemy dendryt spójny w postaci przedstawionej na rysunku 6:



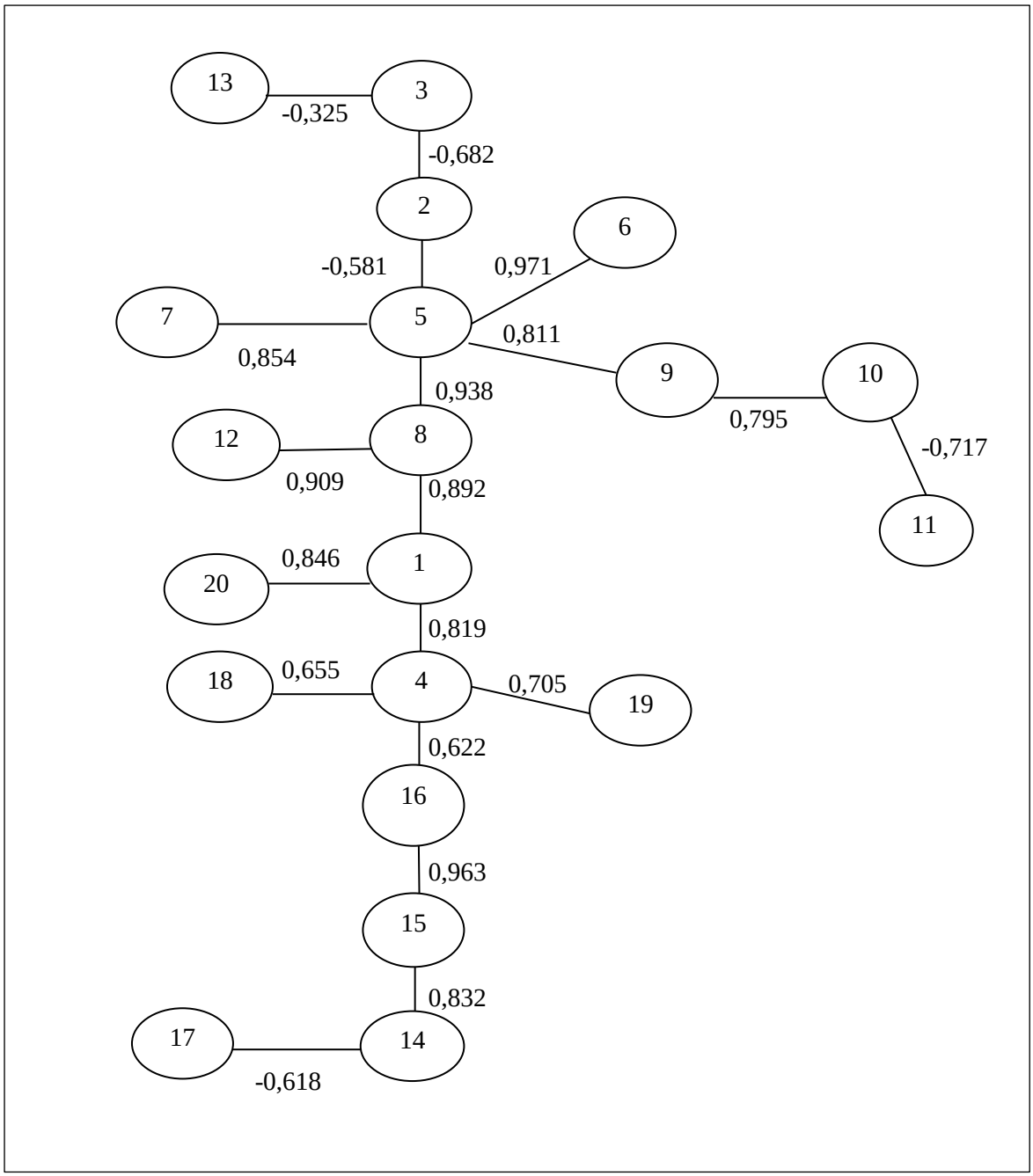
Rysunek 6. Dendryt uspójniony.
Źródło: opracowanie własne.

Na rysunku 7 prezentowany jest dendryt (w zmodyfikowanym nieco kształcie graficznym), w którym przy wiązadłach wpisano wartości odpowiadających im współczynników korelacji. Po usunięciu wiązadeł reprezentujących współczynniki korelacji o wartości niższej od krytycznej (przyjmujemy wartość krytyczną dla $\alpha = 0{,}01$; $\mathbf{r_k = 0{,}623}$) otrzymujemy dendryt postaci przedstawionej na rysunku 8.

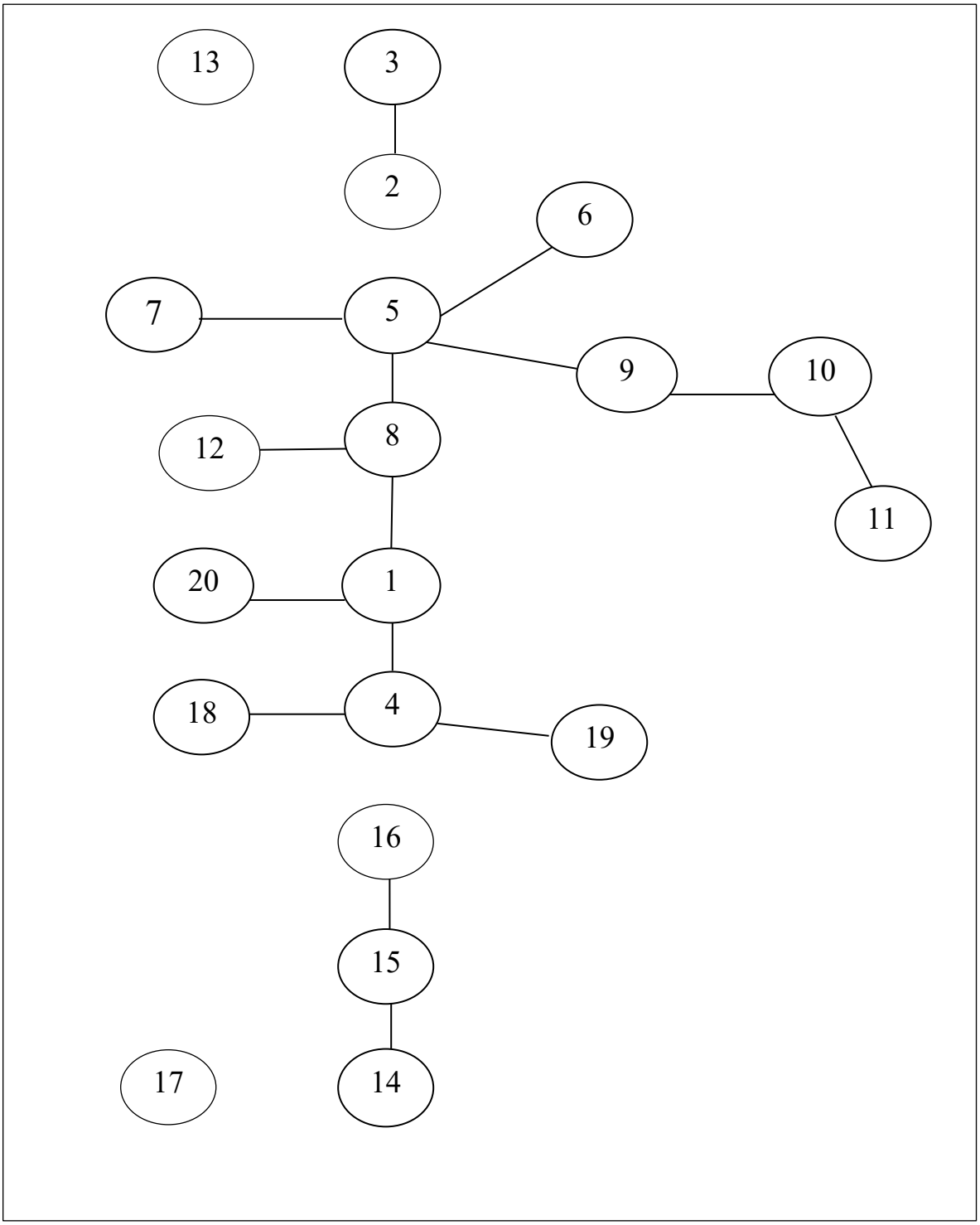
Otrzymany dendryt ukazuje pięć części, w ramach których dokonać należy wyboru cech-reprezentantów. Przypomnijmy omówione przy okazji prezentowania metody grafu zasady/kryteria wyboru cech reprezentantów.

a) najwyższa liczba wiązadeł

W oparciu o tą zasadę (kryterium wyboru cechy reprezentanta) dokonujemy wyboru tej cechy, dla której reprezentujący ją wierzchołek w danej części dendrytu cechuje się największą liczbą krawędzi połączeń.



Rysunek 7. Dendryt z uwzględnieniem współczynników skorelowania.
Źródło: opracowanie własne.



Rysunek 8. Dendryt po usunięciu wiązadeł reprezentujących współczynniki korelacji o wartości niższej od krytycznej.
Źródło: opracowanie własne.

b) najwyższe wartości skorelowania

Ta zasada wymaga ustalenia stopnia skorelowania każdej cechy danej części dendrytu z wszystkimi pozostałymi cechami tej części. Dla ustalenia tego należałoby więc na podstawie wyznaczonej macierzy R dokonać zestawienia współczynników korelacji cech dla tej części dendrytu; czyli sporządzić/zestawić dla każdego części dendrytu oddzielną macierz korelacji odpowiednio „wykrojoną" z pełnej macierzy korelacji R. Pewną pomocą w rozstrzygnięciu zagadnienia wyboru cechy-reprezentanta może okazać się zsumowanie bezwzględnych wartość korelacji każdej kolumny (lub wiersza) tak ustalonych macierzy. Suma o najwyższej wartości może sugerować wybór.

c) w wyniku oceny merytorycznej przydatności cech

Ta zasada (kryterium wyboru) sprowadza się do z góry przyjętego rozstrzygnięcia, którą cechę uważamy za ważną z merytorycznego punktu widzenia; „merytorycznego", tzn. jej przydatności (znaczenia) w charakteryzowaniu kryterium dokonywanej oceny jednostek terytorialnych. Warto zauważyć, że w pewnych przypadkach dwie pierwsze zasady wyboru cechy-reprezentanta zawodzą. Dotyczy to m.in. sytuacji, gdy dana część dendrytu składa się z dwóch lub trzech wierzchołków.

Wracając do naszego przykładu, zauważmy, że:

- ✓ nie ma żadnych wątpliwości, że do dalszego postępowania brane będą pod uwagę cechy nr 13 oraz 17. Są to jednoelementowe części dendrytu, a więc nic tu nie jest przedmiotem wyboru.
- ✓ Z części dwuelementowej wybierzemy cechę 2 lub 3. Podstawą jest jedynie kryterium „c". Z listy prezentowanych wcześniej cech wynika, że są to:
    - stopa bezrobocia,
    - wskaźnik zatrudnienia (pracujący do ludności w wieku 15 lat i więcej).

  Z punktu widzenia kryterium oceny cecha nr 2 wydaje się ważniejsza i tą wybieramy.
- ✓ W oparciu o to samo kryterium (merytorycznej wartości cechy) decydujemy, że część trzyelementową reprezentować będzie cech nr 16.
- ✓ Pozostała jeszcze do rozstrzygnięcia część najbardziej liczna w wierzchołki (reprezentuje 13 cech). W tym przypadku do zastosowania są wszystkie trzy powyższe kryteria wyboru. Gdyby wybór oprzeć na liczbie łuków, wówczas cechą reprezentującą byłaby cecha nr 5 (cztery wchodzące/wychodzące krawędzie). Gdyby z kolei zastosować kryterium siły skorelowania cech tworzących omawianą część dendrytu, musielibyśmy sporządzić macierz korelacji. Korzystając z wyliczonych współczynników korelacji ujętych w tabeli 9, macierz korelacji cech reprezentowanych przez wierzchołki rozważanej części dendrytu jest ujęta[39] w tabeli 10.

[39] Należy pamiętać, że:

– Po pierwsze, prezentowana macierz ujmuje wartości bezwzględne wyliczonych współczynników korelacji (w przypadku wartości ujemnych przyjęto ich moduły), gdyż wysoką korelację obrazować mogą zarówno wartości dodatnie, jak i ujemne.
– Po drugie, w wyliczaniu sumy pominięto wartość głównej przekątnej.
– Po trzecie, wyliczone sumy współczynników korelacji nie mają żadnej interpretacji merytorycznej. Mogą jedynie służyć do podpowiedzi, jaką cechę-reprezentanta wybrać.

Tabela 10

Moduły współczynników korelacji cech reprezentowanych przez wierzchołki wybranej części dendrytu

	1	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**18**	**19**	**20**	**Σ**
1	1	0,82	0,79	0,78	0,71	0,89	0,62	0,4	0,34	0,86	0,63	0,58	0,85	8,27
4	0,82	1	0,57	0,63	0,62	0,61	0,43	0,36	0,17	0,74	0,66	0,71	0,76	7,08
5	0,79	0,57	1	0,97	0,85	0,94	0,81	0,69	0,53	0,86	0,48	0,58	0,73	8,8
6	0,78	0,63	0,97	1	0,84	0,9	0,79	0,74	0,57	0,88	0,55	0,62	0,77	9,04
7	0,71	0,62	0,85	0,84	1	0,78	0,72	0,64	0,42	0,71	0,53	0,59	0,53	7,94
8	0,89	0,61	0,94	0,9	0,78	1	0,79	0,59	0,42	0,91	0,47	0,55	0,75	8,6
9	0,62	0,43	0,81	0,79	0,72	0,79	1	0,79	0,5	0,76	0,38	0,28	0,55	7,42
10	0,4	0,36	0,69	0,74	0,64	0,59	0,79	1	0,72	0,69	0,37	0,45	0,51	6,95
11	0,34	0,17	0,53	0,57	0,42	0,42	0,5	0,72	1	0,43	0,49	0,12	0,48	5,19
12	0,86	0,74	0,86	0,88	0,71	0,91	0,76	0,69	0,43	1	0,58	0,7	0,8	8,92
18	0,63	0,66	0,48	0,55	0,53	0,47	0,38	0,37	0,49	0,58	1	0,39	0,49	6,02
19	0,58	0,71	0,58	0,62	0,59	0,55	0,28	0,45	0,12	0,7	0,39	1	0,58	6,15
20	0,85	0,76	0,73	0,77	0,53	0,75	0,55	0,51	0,48	0,8	0,49	0,58	1	7,8
Σ	8,27	7,08	8,8	9,04	7,94	8,6	7,42	6,95	5,19	8,92	6,02	6,15	7,8	

Jak zauważamy, najwyższą siłą skorelowania legitymuje się cecha nr 6. Opierając się na omawianym kryterium, tę cechę należałoby wybrać jako najlepiej reprezentującą wszystkie cechy pozostałe tej części dendrytu. Nieco tylko niższą siłę skorelowania posiada cecha nr 12.

Umówmy się, że ostatecznie wybór cechy-reprezentanta oprzemy jednak na kryterium trzecim[40], tj. na ocenie merytorycznej przydatności danej cechy w przeprowadzanej ocenie jednostek terytorialnych (województw). Jako przeprowadzający przedmiotową ocenę, podjąłem więc decyzję, że cechą tą będzie cecha nr 12 (teatry i instytucje muzyczne). Wybór ten argumentuję tym, że reprezentuje ona ważną składową ogólnego poziomu rozwoju województw (pamiętamy, że w naszym przykładzie poziom ten jest kryterium przeprowadzanej oceny). Nie bez znaczenia jest też relatywnie wysoka siła skorelowania tej cechy z cechami, które ma reprezentować.

Na rysunku 9 prezentowany jest dendryt z zaznaczeniem wybranych cech.

Ostatecznie, ze zbioru 20 cech wyjściowych wyselekcjonowanych zostało 5 cech legitymujących się dwoma ważnymi właściwościami, tj.:

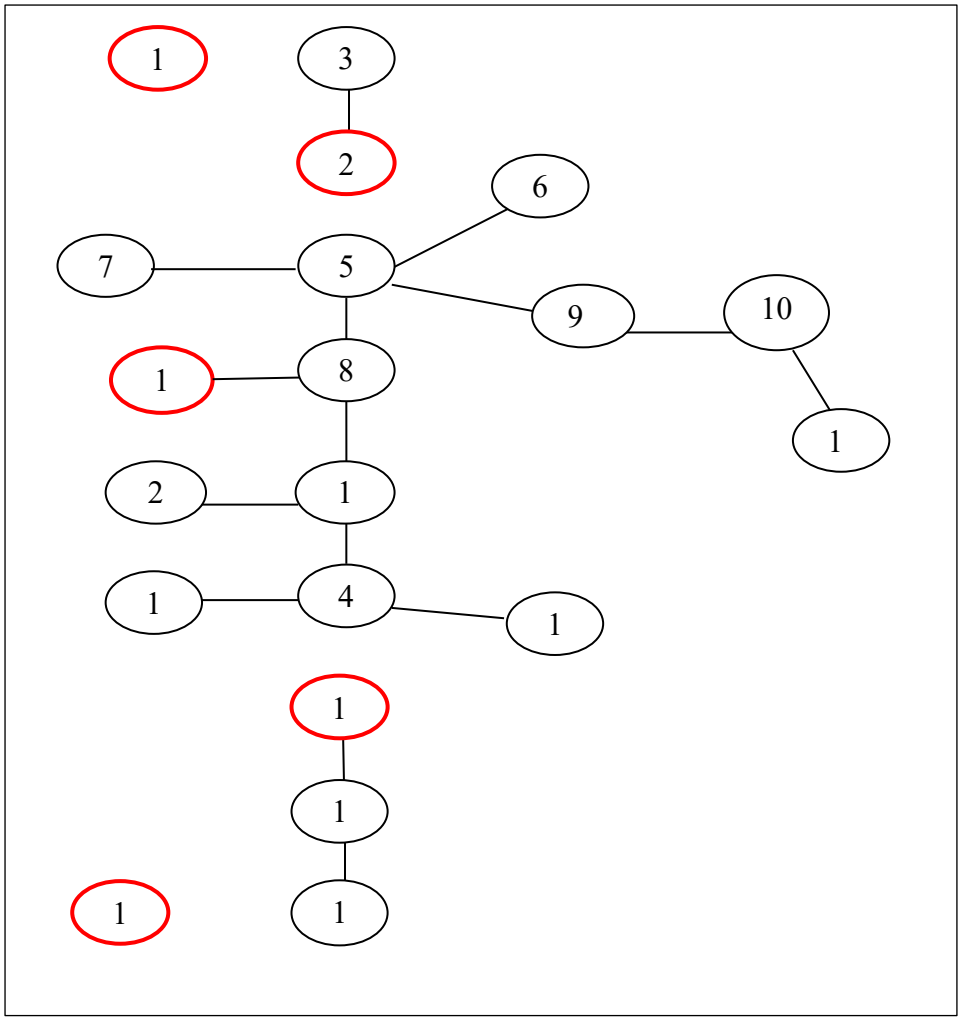
- ✓ po pierwsze, nie są ze sobą wysoko skorelowane ($|r| \leq 0{,}623$),
- ✓ po drugie, dobrze reprezentują pozostałe, nie uwzględnione cechy (są z nimi wysoko skorelowane).

Lista wybranych cech przedstawia się następująco (zachowano numerację pierwotną z listy całego zestawu cech):

2. Stopa bezrobocia
12. Teatry i instytucje muzyczne
13. Plony zbóż
16. Turyści zagraniczni korzystający z noclegów na 1000 ludności
17. Odsetek gospodarstw domowych posiadających więcej niż jeden samochód osobowy.

[40] Raz jeszcze warto przypomnieć, że żadna ze sformułowanych zasad (kryterium) wyboru cech-reprezentantów nie ma przewagi nad pozostałymi. Każdorazowo analityk musi decydować, na którym kryterium wybór opiera.

Te właśnie cechy będą podstawą dalszego postępowania związanego z oceną poziomu rozwoju województw. Będzie to przedmiotem rozważań w dalszych podrozdziałach podręcznika.



Rysunek 9. Dendryt z zaznaczeniem wybranych cech.
Źródło: opracowanie własne.

Pytania/zadania kontrolne

1. Które etapy algorytmu związanego z zastosowaniem metody dendrytu do wyboru cech diagnostycznych mają charakter, w części przynajmniej, wyboru uznaniowego? Tzn. jakie konkretne rozstrzygnięcia zależą od twojej (badacza) decyzji?
2. Czy możliwa jest sytuacja, w której w wyniku zrealizowania etapu oznaczonego nr 3 w algorytmie otrzymujemy dendryt spójny?
3. Obniżenie wartości |rk| prowadzić będzie do zwiększenia czy zmniejszenia liczby wybranych cech diagnostycznych?

2.4. Metody syntetycznej oceny jednostek terytorialnych – standaryzacja cech

Zgodnie z procedurą wyznaczania wskaźnika syntetycznej oceny (patrz tabela 1), omówienia wymagają trzy sposoby (metody) standaryzacji cech:

a) *Zero-jedynkowa*
b) Uproszczona
c) *Min-max*.

Zanim jednak przejdziemy do kolejnego omawiania tych metod, wcześniej wyjaśnieniom należy poddać samą ideę oraz cel standaryzacji.

W etapie wyboru cech diagnostycznych za pomocą omówionych powyżej metod (grafu i dendrytu) dokonano ostatecznej selekcji cech, które mają być podstawą budowy wskaźnika syntetycznej oceny jednostek terytorialnych. Jak pamiętamy, w naszym przykładzie, który dotyczył oceny województw z punktu widzenia osiągniętego poziomu rozwoju, z 20 cech ostatecznie wybranych zostało 5 z nich, a więc te, które spełniają wymóg: *nie są ze sobą wysoko skorelowane*. Przypomnijmy te cechy:

1) Stopa bezrobocia.
2) Teatry i instytucje muzyczne.
3) Plony zbóż.
4) Turyści zagraniczni korzystający z noclegów na 1000 ludności.
5) Odsetek gospodarstw domowych posiadających więcej niż jeden samochód osobowy.

Zauważmy, że cechy te wyrażane są w różnych jednostkach, co czyni, iż ich wartości liczbowe są nieporównywalne ze sobą. Dokonajmy kolejnego przypomnienia, a mianowicie, że naszym zadaniem finalnym są oceny względne, tzn. takie, na podstawie których możliwe jest wnioskowanie, na ile dana jednostka terytorialna jest lepsza lub gorsza od każdej innej w rozważanym ich zbiorze.

W tym miejscu sformułowane zostanie zadanie niniejszego podrozdziału: **dokonać należy takiego przekształcenia ciągu liczb każdej cechy z osobna, aby w efekcie doprowadzić je do wzajemnej porównywalności, z jednoczesnym zachowaniem pełnej, pierwotnej (czyli sprzed przekształcenia) informacji o ocenianych obiektach**. Podręczniki statystyki dostarczają całej gamy formuł służących takiemu przekształceniu. Tego rodzaju zabieg rachunkowy będziemy nazywać standaryzacją cech[41]. Rozważania ograniczymy do trzech, wymienionych wyżej, formuł standaryzacji.

1) Metoda *zero-jedynkowa*:

$$t_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad (7)$$

x_{ij} – wyjściowa wartość j-tej cechy dla i-tej jednostki terytorialnej (patrz oznaczenia przyjęte w tabeli 2, w części poświęconej wymogom stawianym wskaźnikom oceny)

t_{ij} – standaryzowana wartość j-tej cechy dla i-tego obiektu (jednostki terytorialnej)

$\bar{x}_j$ – średnia arytmetyczna cechy j-tej

s_j – odchylenie standardowe cechy j-tej.

[41] Dla uproszczenia rezygnujemy z rozróżniania „standaryzacja", „normalizacja", „unitaryzacja" cech, co przy przyjętym poziomie ogólności dociekań dotyczących prezentowanych metod nie będzie miało wpływu na poprawność dalszych rozważań. Trzeba jednak zauważyć, że w bardziej statystycznie zaawansowanych rozważaniach argumentowana jest zasadność rozróżniania tych pojęć. Por. np. Kukuła K. (2000); także Malchar J., Zielińska-Sitkiewicz M. (2017).

Należy zauważyć, że parametrem „standaryzującym" jest w tej metodzie średnia arytmetyczna oraz odchylenie standardowe.

Zanim przejdziemy do omówienia właściwości cechy, która zestandaryzowana została metodą *zero-jedynkową*, warto zwrócić uwagę na jej nazwę. Nie jest to nazwa przypadkowa; jest ona stosowana powszechnie w literaturze dotyczącej statystyki. Pierwsza część nazwy (*zero*) bierze się stąd, że średnia arytmetyczna zmiennej (cechy) standaryzowanej tą metodą wynosi właśnie „0". Właściwość tę prawie wprost uwidacznia formuła (7). Gdybyśmy prawą stronę tego równania zsumowali po „i", nietrudno zauważyć, że suma licznika zawsze wynosiłaby 0 (a więc średnia arytmetyczna również wynosiłaby 0). **Przy okazji warto zauważyć, że zmienna standaryzowana tą metodą przyjmuje zawsze wartości zarówno dodatnie, jak i ujemne**.

Druga część nazwy (*jedynkow*a) informuje z kolei, że odchylenie standardowe zmiennej standaryzowanej wynosi „1". Jest to bardzo ważne ustalenie, na które zwrócimy uwagę w dalszej części informacji o właściwościach cech standaryzowanych *zero-jedynkowo*. Jak pamiętamy ze statystyki, średnia arytmetyczna oraz odchylenie standardowe są dwoma podstawowymi parametrami rozkładu cechy.

Przyjmijmy założenie, że nasze cechy mają rozkład normalny (a bardziej realistycznie: rozkład zbliżony do normalnego[42]). Możemy powiedzieć wtedy, że najbardziej prawdopodobna, dowolnie wskazana wartość zmiennej będzie bliska jej średniej arytmetycznej; albo inaczej, bardziej obrazowo – największe zagęszczenie obserwacji (u nas liczba obserwacji wynosi „m") będzie w pasie bliskim wartości średniej; im dalej od punktu wartości średniej (na lewo i na prawo w wykresie rozkładu normalnego) tym rzadsze będą realizacje zmiennej i w dodatku spadek ten będzie symetryczny względem średniej (zob. rysunek 10).

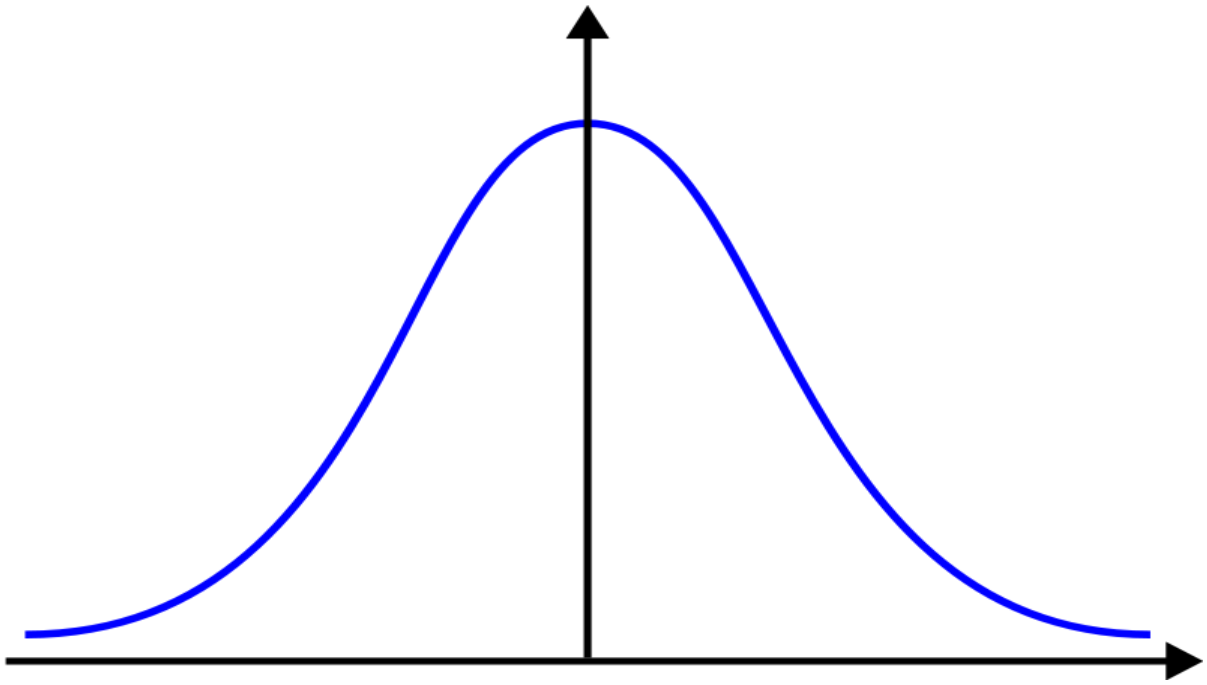
Z rozkładem normalnym wiąże się powszechnie znana w statystyce reguła trzech sigm (nazwa związana z rozkładem teoretycznym populacji generalnej) lub też trzech odchyleń standardowych (nazwa związana z rozkładem próby). Jeżeli dla danej cechy o rozkładzie normalnym znamy jej wartość średniej arytmetycznej oraz wartość odchylenia standardowego, wówczas znamy również procentowy udział trzech przedziałów liczbowych w ogólnej liczbie obserwacji danej cechy. Są to następujące przedziały:

W przedziale: $[\bar{x} - s;\ \ \bar{x} + s]$ mieści się 68,2% obserwacji.

W przedziale: $[\bar{x} - 2s;\ \ \bar{x} + 2s]$ mieści się 95,4% obserwacji.

W przedziale: $[\bar{x} - 3s;\ \ \bar{x} + 3s]$ mieści się 99,7% obserwacji.

[42] Nie jest to zbyt wyidealizowane założenie względem cech charakteryzujących procesy społeczno-ekonomiczne.



Rysunek 10. Rozkład normalny.
Źródło: opracowanie własne.

Korzystając z przedstawionego wyżej ustalenia, wskazującego, że dla zmiennej standaryzowanej formułą *zero-jedynkową*, średnia arytmetyczna wynosi 0, zaś odchylenie standardowe 1, przedziały liczbowe wynikające z omawianej reguły dla tej zmiennej (cechy) przyjmują postać (patrz również rysunek 11):

W przedziale: [-1; 1] mieści się 68,2% obserwacji.

W przedziale: [-2; 2] mieści się 95,4% obserwacji.

W przedziale: [-3; 3] mieści się 99,7% obserwacji.

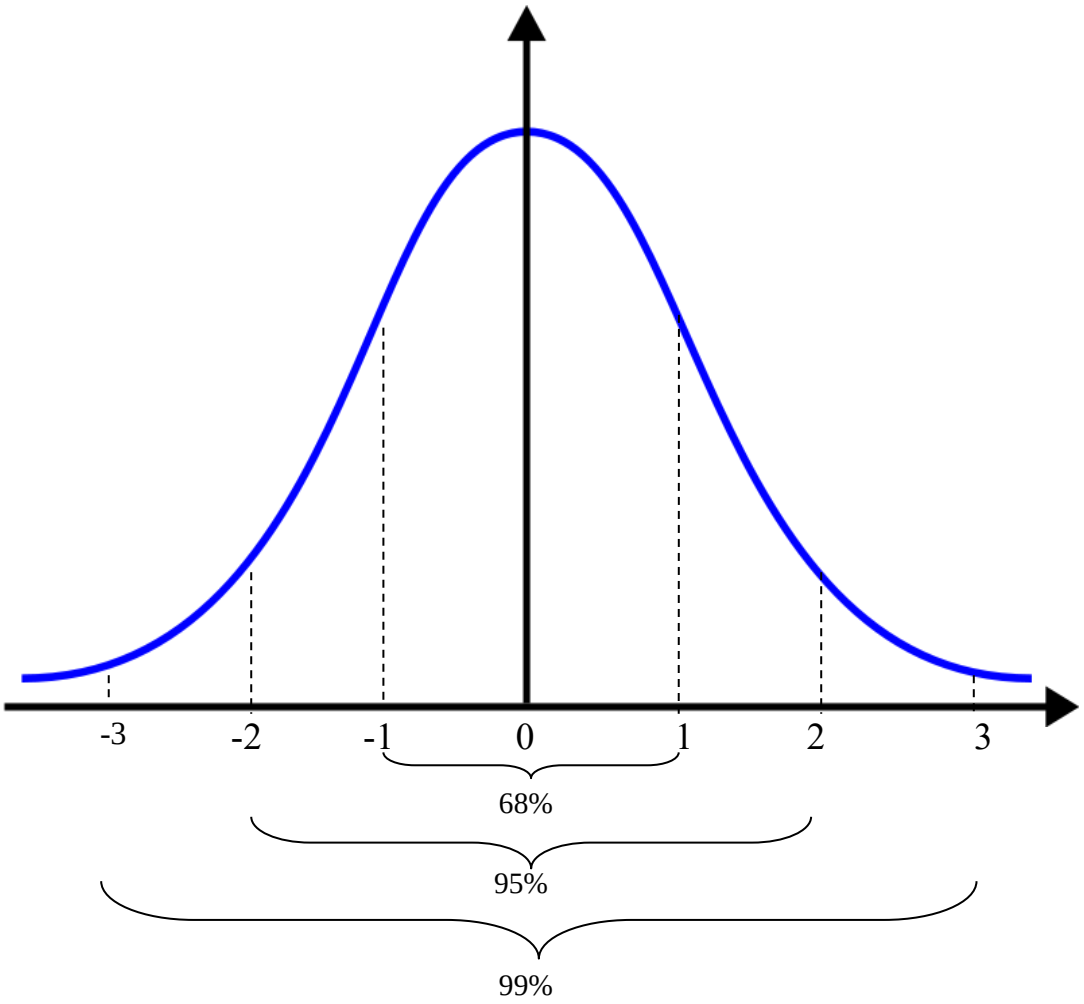
Z rozkładu procentowego możemy łatwo przejść na rozkład prawdopodobieństwa. Możemy bowiem w odniesieniu do zmiennej standaryzowanej *zero-jedynkowo* powiedzieć, że:

Z prawdopodobieństwem 0,682 jej wartości zamykać się będą w przedziale [-1; 1].

Z prawdopodobieństwem 0,954 jej wartości zamykać się będą w przedziale [-2; 2].

Z prawdopodobieństwem 0,997 jej wartości zamykać się będą w przedziale [-3; 3].

Konkluzja nasuwająca się z powyższych rozważań jest następująca: **zmienna (cecha) standaryzowana *zero-jedynkowo* przyjmuje zawsze relatywnie nieduże wartości**; albo inaczej: **niewielkie jest prawdopodobieństwo tego, że wartości tak standaryzowanej zmiennej będą spoza przedziału, np. [-3; 3]**. Dla przeprowadzających analizy terytorialnego systemu społeczno-gospodarczego jest to ważna informacja. Jeżeli bowiem dla jakiejś cechy pojawi się wartość spoza tego przedziału (dotyczyć to będzie jakiejś jednostki terytorialnej), powinno to być dla nich sygnałem zachęcającym do zgłębienia przyczyn zaistnienia w tej jednostce takiej właśnie wartości (ekstremalnej).



Rysunek 11. Reguła trzech sigm cechy standaryzowanej *zero-jedynkowo*.
Źródło: opracowanie własne.

2) Metoda uproszczona:

$$t_{ij} = \frac{x_{ij}}{s_j} \qquad t_{ij} = \frac{x_{ij}}{\bar{x}_j} \qquad (8)$$

Na użytek niniejszego podręcznika przyjęto powyższą nazwę z uwagi na to, że jest ona pewnym uproszczeniem formuły poprzedniej. W liczniku nie występuje średnia arytmetyczna. Tak uproszczona formuła może przybierać dwie postacie: w pierwszej parametrem standaryzującym jest odchylenie standardowe cechy, a w drugiej średnia arytmetyczna. Podobnie jak w przypadku standaryzacji *zero-jedynkowej*, otrzymywane wartości zmiennej standaryzowanej (t_{ij}) spełniają zakładany cel, którym jest doprowadzenie do porównywalności z wartościami innych cech standaryzowanych tą samą postacią formuły, a jednocześnie zachowują pierwotny ładunek informacji o ocenianych jednostkach terytorialnych.

Stosując którąś postać powyższej formuły, pamiętać trzeba o dwóch istotnych kwestiach dotyczących poprawności otrzymywanych wyników:

a) Wartości cechy standaryzowanej formułą, w której parametrem standaryzującym jest odchylenie standardowe, zależą od stopnia wewnętrznego zróżnicowania tej cechy. Jeżeli jest ono wyższe, otrzymujemy mniejsze wartości t_{ij}, i odwrotnie przy zróżnicowaniu mniejszym. Wynika to z tego, że wartość odchylenia standardowego, będąca w mianowniku formuły, wprost zależy od stopnia tego zróżnicowania (odchylenie standardowe jest miarą stopnia zróżnicowania cechy). Dokonując porównywania wartości wielu cech standaryzowanych taką formułą, zauważamy wtedy, że są one poniekąd sztucznie ważone wartością s_j, czyli stopniem wewnętrznego zróżnicowania cech.

b) Standaryzacja cech drugą postacią tej formuły mankamentu tego nie posiada. Nie może być jednak stosowana w sytuacji, gdy któraś z cech diagnostycznych przyjmuje ujemne wartości (np. saldo migracji), a bardziej konkretnie – jeżeli średnia arytmetyczna tej cechy jest wartością ujemną.

3) Metoda *min-max*:

$$t_{ij} = \begin{cases} \dfrac{x_{ij} - x_{min}}{x_{max} - x_{min}} & \text{dla stymulant} \\ \dfrac{x_{max} - x_{ij}}{x_{max} - x_{min}} & \text{dla destymulant} \end{cases} \qquad (9)$$

Metoda *min-max* jest często stosowaną formułą standaryzacji cech, głównie z uwagi na interesujące właściwości tak otrzymywanych wartości standaryzowanych. Parametrami standaryzującymi daną cechę jest jej wartość maksymalna (max) oraz wartość minimalna (min). Zauważyć trzeba, że nieco inna jest formuła standaryzacji dla cech stymulant oraz cech destymulant. Formuły różnią się licznikami ilorazu. Jest tak dlatego, że przy standaryzacji *min-max* wszystkie zmienne będące destymulantami nie tylko są standaryzowane, ale jednocześnie są przekształcane w stymulanty. Wystarczy zwrócić uwagę na licznik. Jeśli od wartości maksymalnej w danej cesze odejmowane są wartości bieżące należące do danej (i-tej, i=1, 2, 3…m) jednostki terytorialnej, to reszty, które zostają z odejmowania, oznaczają, że tam gdzie była niska pierwotna wartość cechy (x_{ij}), będzie wysoka jej wartość po standaryzacji, zaś tam, gdzie była wysoka, po odjęciu będzie niska.

O ile poprzednie dwie metody standaryzowania były „neutralne" względem podziału cech na stymulanty i destymulanty, tzn. nie wymagały przed standaryzacją rozpoznania ich w tym względzie, o tyle metoda *min-max* takiego rozpoznania wymaga. Ale też standaryzacja poprzednimi metodami oznaczała, że również po standaryzacji cechy utrzymywały swój status podziału na stymulanty i destymulanty.

Wspominano wyżej o interesujących właściwościach cech standaryzowanych omawianą metodą. Przyglądnijmy się więc bliżej formułom (9):

✓ Po pierwsze, proszę zwrócić uwagę, że mianowniki są zawsze dodatnie. Ktoś może powiedzieć, że teoretycznie może się zdarzyć, że wartość maksymalna będzie równa wartości minimalnej, wtedy pojawiałby się kłopot z mianownikiem (jego wartość wynosiłaby 0). W istocie, teoretycznie tak mogłoby być, ale w naszym przypadku jest to wykluczone.

„Wykluczenie" to nastąpiło na etapie *dyskusja cech – wybór mierników szczegółowych* (punkt 2 ogólnego algorytmu), kiedy to postawiliśmy warunek/wymóg, że cechy *muszą wykazywać dostateczne wewnętrzne zróżnicowanie*[43]. Wartości maksymalna i minimalna będą sobie równie tylko wtedy, gdy dana cecha przyjmuje te same wartości dla wszystkich rozważanych obiektów (jednostek terytorialnych), a takie cechy nie mają żadnej wartości ocennej.

- ✓ Po drugie, w obydwóch ilorazach licznik jest mniejszy od mianownika, ale zawsze z jednym wyjątkiem. Wyjątkiem tym jest to, że – w przypadku stymulant – dla jednostki, która legitymuje się wartością maksymalną danej cechy, licznik przyjmie postać/wartość mianownika; zaś w przypadku destymulant to samo pojawi się dla jednostki, która osiągnęła wartość minimalną. Wynika z tego, że górną granicą cechy standaryzowanej *min-max* jest 1.
- ✓ Po trzecie, zauważmy również, że najmniejszą wartością, jaką może przybierać licznik jest 0. Dotyczy to obiektu, który przyjmuje wartość minimalną dla stymulant oraz wartość maksymalną dla destymulat. Oznacza to, że dolną granicą wartości standaryzowanej metodą *min-max* jest 0.

Konkluzja końcowa dotycząca właściwości cech standaryzowanych jest następująca: **Wartości standaryzowane dowolnej cechy metodą *min-max* zamykają się w granicach [0; 1]; ponadto, zawsze przynajmniej raz[44] pojawi się wśród wartości standaryzowanych 0, jak również przynajmniej raz 1.**

Na zakończenie prezentujemy cztery tabele (nr 11-14), ujmujące wartości standaryzowane (t_{ij}) cech diagnostycznych, które w naszym przykładzie wybrane zostały spośród wstępnie przyjętego szerszego ich zbioru, z wykorzystaniem zasady: *nie są ze sobą wysoko skorelowane*. Celem stworzenia możliwości przećwiczenia odpowiednich wyliczeń, w tabelach zamieszczono również wartości wyjściowe cech (x_{ij}) oraz wyliczone wartości parametrów standaryzujących ($\bar{x}_j$, s_j, max, min).

[43] Podrozdział 2.1

[44] „Przynajmniej raz" dlatego, że nie można wykluczać sytuacji, że takie same wartości może osiągnąć więcej niż jeden obiekt (jednostka terytorialna) i że te właśnie wartości mogą okazać się wartościami maksymalnymi lub minimalnymi.

Tabela 11
Wyniki standaryzacji zero-jedynkowej

Województwa	Dane wyjściowe					Dane standaryzowane				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
dolnośląskie	17,3	17	43,9	163	41	0,258	0,630	1,820	0,662	-1,443
kujawsko-pomorskie	19,4	7	32,4	42	46	0,861	-0,478	-0,055	-0,780	-0,253
lubelskie	15,5	6	29,6	42	50	-0,258	-0,589	-0,511	-0,780	0,699
lubuskie	20	3	32,9	152	47	1,034	-0,921	0,026	0,531	-0,015
łódzkie	15,3	14	27,2	37	47	-0,316	0,298	-0,903	-0,840	-0,015
małopolskie	11,6	20	33,3	273	49	-1,378	0,963	0,092	1,974	0,461
mazowieckie	12,2	34	27	155	48	-1,206	2,514	-0,935	0,567	0,223
opolskie	16,6	3	48,6	31	50	0,057	-0,921	2,586	-0,911	0,699
podkarpackie	16,4	3	29,7	30	53	0,000	-0,921	-0,495	-0,923	1,413
podlaskie	13,4	5	26,8	72	44	-0,861	-0,700	-0,968	-0,422	-0,729
pomorskie	16	15	32,3	134	42	-0,115	0,409	-0,071	0,317	-1,205
śląskie	13,5	24	34,7	56	47	-0,833	1,406	0,320	-0,613	-0,015
świętokrzyskie	18	3	27,2	19	48	0,459	-0,921	-0,903	-1,054	0,223
warmińsko-mazurskie	23,6	4	27,9	153	45	2,067	-0,810	-0,789	0,543	-0,491
wielkopolskie	12,1	11	35,4	79	56	-1,235	-0,035	0,434	-0,339	2,127
zachodniopomorskie	21,5	12	34,9	281	40	1,464	0,076	0,353	2,069	-1,680
Średnia arytm.	16,4000	11,3125	32,7375	107,4375	47,0625	-0,002	0,000	0,001	0,001	-0,001
Odch. stand.	3,4831	9,0238	6,1344	83,8840	4,2027	1000	1000	1000	1000	1000

Źródło: obliczenia własne.
Należy wyjaśnić, że niewielkie odchylenia od zera średniej arytmetycznej cech standaryzowanych wynikają z zaokrągleń.

Tabela 12
Wyniki standaryzacji uproszczonej (x/odchylenie standardowe)

Województwa	Dane wyjściowe					Dane standaryzowane				
	1	2	3	4	5	1	2	3	4	5
dolnośląskie	17,3	17	43,9	163	41	4,967	1,884	7,156	1,943	9,756
kujawsko-pomorskie	19,4	7	32,4	42	46	5,570	0,776	5,282	0,501	10,945
lubelskie	15,5	6	29,6	42	50	4,450	0,665	4,825	0,501	11,897
lubuskie	20	3	32,9	152	47	5,742	0,332	5,363	1,812	11,183
łódzkie	15,3	14	27,2	37	47	4,393	1,551	4,434	0,441	11,183
małopolskie	11,6	20	33,3	273	49	3,330	2,216	5,428	3,254	11,659
mazowieckie	12,2	34	27	155	48	3,503	3,768	4,401	1,848	11,421
opolskie	16,6	3	48,6	31	50	4,766	0,332	7,923	0,370	11,897
podkarpackie	16,4	3	29,7	30	53	4,708	0,332	4,842	0,358	12,611
podlaskie	13,4	5	26,8	72	44	3,847	0,554	4,369	0,858	10,470
pomorskie	16	15	32,3	134	42	4,594	1,662	5,265	1,597	9,994
śląskie	13,5	24	34,7	56	47	3,876	2,660	5,657	0,668	11,183
świętokrzyskie	18	3	27,2	19	48	5,168	0,332	4,434	0,227	11,421
warmińsko-mazurskie	23,6	4	27,9	153	45	6,776	0,443	4,548	1,824	10,707
wielkopolskie	12,1	11	35,4	79	56	3,474	1,219	5,771	0,942	13,325
zachodniopomorskie	21,5	12	34,9	281	40	6,173	1,330	5,689	3,350	9,518
Średnia arytm.	16,4000	11,3125	32,7375	107,4375	47,0625					
Odch. stand.	3,4831	9,0238	6,1344	83,8840	4,2027					

Źródło: obliczenia własne.

Tabela 13
Wyniki standaryzacji uproszczonej (x/(średnia arytmetyczna)

Województwa	Dane wyjściowe					Dane standaryzowane				
	1	2	3	4	5	1	2	3	4	5
dolnośląskie	17,3	17	43,9	163	41	1,055	1,503	1,341	1,517	0,871
kujawsko-pomorskie	19,4	7	32,4	42	46	1,183	0,619	0,990	0,391	0,977
lubelskie	15,5	6	29,6	42	50	0,945	0,530	0,904	0,391	1,062
lubuskie	20	3	32,9	152	47	1,220	0,265	1,005	1,415	0,999
łódzkie	15,3	14	27,2	37	47	0,933	1,238	0,831	0,344	0,999
małopolskie	11,6	20	33,3	273	49	0,707	1,768	1,017	2,541	1,041
mazowieckie	12,2	34	27	155	48	0,744	3,006	0,825	1,443	1,020
opolskie	16,6	3	48,6	31	50	1,012	0,265	1,485	0,289	1,062
podkarpackie	16,4	3	29,7	30	53	1,000	0,265	0,907	0,279	1,126
podlaskie	13,4	5	26,8	72	44	0,817	0,442	0,819	0,670	0,935
pomorskie	16	15	32,3	134	42	0,976	1,326	0,987	1,247	0,892
śląskie	13,5	24	34,7	56	47	0,823	2,122	1,060	0,521	0,999
świętokrzyskie	18	3	27,2	19	48	1,098	0,265	0,831	0,177	1,020
warmińsko-mazurskie	23,6	4	27,9	153	45	1,439	0,354	0,852	1,424	0,956
wielkopolskie	12,1	11	35,4	79	56	0,738	0,972	1,081	0,735	1,190
zachodniopomorskie	21,5	12	34,9	281	40	1,311	1,061	1,066	2,615	0,850
Średnia arytm.	16,4000	11,3125	32,7375	107,4375	47,0625					
Odch. stand.	3,4831	9,0238	6,1344	83,8840	4,2027					

Źródło: obliczenia własne.

Tabela 14
Wyniki standaryzacji min-max

Województwa	Dane wyjściowe					Dane standaryzowane				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
dolnośląskie	17,3	17	43,9	163	41	0,525	0,452	0,784	0,550	0,063
kujawsko-pomorskie	19,4	7	32,4	42	46	0,350	0,129	0,257	0,088	0,375
lubelskie	15,5	6	29,6	42	50	0,675	0,097	0,128	0,088	0,625
lubuskie	20	3	32,9	152	47	0,300	0,000	0,280	0,508	0,438
łódzkie	15,3	14	27,2	37	47	0,692	0,355	0,018	0,069	0,438
małopolskie	11,6	20	33,3	273	49	1,000	0,548	0,298	0,969	0,563
mazowieckie	12,2	34	27	155	48	0,950	1,000	0,009	0,519	0,500
opolskie	16,6	3	48,6	31	50	0,583	0,000	1,000	0,046	0,625
podkarpackie	16,4	3	29,7	30	53	0,600	0,000	0,133	0,042	0,813
podlaskie	13,4	5	26,8	72	44	0,850	0,065	0,000	0,202	0,250
pomorskie	16	15	32,3	134	42	0,633	0,387	0,252	0,439	0,125
śląskie	13,5	24	34,7	56	47	0,842	0,677	0,362	0,141	0,438
świętokrzyskie	18	3	27,2	19	48	0,467	0,000	0,018	0,000	0,500
warmińsko-mazurskie	23,6	4	27,9	153	45	0,000	0,032	0,050	0,511	0,313
wielkopolskie	12,1	11	35,4	79	56	0,958	0,258	0,394	0,229	1,000
zachodniopomorskie	21,5	12	34,9	281	40	0,175	0,290	0,372	1,000	0,000
Max	23,6	34	48,6	281	56					
Min	11,6	3	26,8	19	40					

Źródło: obliczenia własne.

Pytania/zadania kontrolne

1. Podejmij próbę udokumentowania, że odchylenie standardowe zmiennej standaryzowanej formułą *zero-jedynkową* wynosi 1.
2. Wyjaśnij słabą stronę standaryzacji cech formułą uproszczoną postaci: $t_{ij} = \frac{x_{ij}}{s_j}$.
3. Dlaczego w prezentacji formuły standaryzacji *zero-jedynkowej* oraz *min-max* (wzór 7 i 9) możliwe było pominięcie założenia, że mianownik musi być różny (większy) od 0?
4. Przemyśl, jak na podstawie danych standaryzowanych wnioskować można o formule tej standaryzacji.
5. Czytelnikowi przeglądającemu powyższe tabele ze zmiennymi standaryzowanymi proponujemy – dla sprawdzenia wiedzy dotyczącej omawianej wcześniej standaryzacji – zastanowienie się nad kilkoma kwestiami:
   a) Warto podjąć próbę wyjaśnienia, dlaczego zmienne standaryzowane *zero-jedynkowo* (tabela 11) dla niektórych jednostek terytorialnych przyjmują wartości wykraczające poza przedział [-2; 2], a więc o relatywnie niskim prawdopodobieństwie pojawienia się. Uzasadnienia należy szukać w treści merytorycznej cech.
   b) Jakie mogą być przyczyny tego, że w tabeli 12 rząd wartości zmiennej standaryzowanej o nr 5 jest znacząco wyższy w porównaniu z wartościami standaryzowanymi pozostałych cech?
   c) Jakimi właściwościami, z punktu widzenia danej cechy, legitymują się jednostki terytorialne (województwa), dla których zmienna standaryzowana *min-max* przyjmuje wartość „0”?; to samo rozważyć należy dla przypadku wartości standaryzowanych wynoszących „1”.
   d) Porównać „dane standaryzowane” z „danymi wyjściowymi” (odpowiednio prawy oraz lewy zestaw kolumn w powyższych tabelach), po to, aby ustalić:
      – zastosowaną formułę standaryzacji,
      – w przypadku standaryzacji *min-max*, które cechy są stymulantami, a które destymulantami.

2.5. Metody syntetycznej oceny jednostek terytorialnych – agregacja cech metodą sumy cech standaryzowanych

W poprzednim podrozdziale omówiono kwestie związane ze standaryzacją cech, a więc takim ich przekształceniem, które doprowadza, że ich wartości dla danej jednostki terytorialnej stają się porównywalne ze sobą.

Przedmiotem niniejszych rozważań jest ostatni etap całej procedury, a więc wyznaczenie wskaźnika syntetycznego oceny badanych jednostek terytorialnych. Jest to etap stawiający przysłowiową „kropkę nad i” w omawianej metodzie, wprost odzwierciedlający jej istotę. Etapy poprzednie służyły przygotowaniu cech szczegółowych do ich agregowania w jeden wskaźnik uogólniający rozważaną oceną.

Spośród kilku znanych w literaturze możliwości metodycznych tego etapu, wybieramy dwie z nich:

1) Metoda sumy cech standaryzowanych.
2) Metoda wzorca rozwoju (modelową).

Pierwsza z tych metod występuje niekiedy pod nazwą metody Perkala, a druga pod nazwą metody Hellwiga (od nazwisk ich autorów). W tym podrozdziale omówiona zostanie metoda sumy cech standaryzowanych. Jest ona koncepcyjnie i rachunkowo bardzo prosta. Być może dlatego często jest stosowana w praktyce. Zestawienie (10) prezentuje macierz wartości standaryzowanych cech diagnostycznych. Każdy i-ty wiersz tej macierzy jest wektorem wartości oceniających i-tą jednostkę terytorialną z punktu widzenia zadanego kryterium. Ponieważ wartości te są porównywalne[45], **istota metody sumy cech standaryzowanych, stosownie do jej nazwy, polega na zsumowaniu tych wartości**. Ilustrują to formuły 11 i 12.

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & \ldots t_{1n} \\ t_{21} & t_{22} & t_{23} & t_{24} & \ldots t_{2n} \\ t_{31} & t_{32} & t_{33} & t_{34} & \ldots t_{3n} \\ t_{41} & t_{42} & t_{43} & t_{44} & \ldots t_{4n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ t_{m1} & t_{m2} & t_{m3} & t_{m4} & \ldots t_{nn} \end{bmatrix} \qquad (10)$$

$$W_i = \sum_{j=1}^{n} kt_{ij} \qquad (11)$$

lub

$$W_i = \frac{1}{n}\sum_{j=1}^{n} kt_{ij} \qquad (12)$$

$i = 1,2, \ldots, m$

$k = 1$ dla stymulant i -1 dla destymulant (należy pamiętać, że wszystkie zmienne po standaryzacji metodą *min-max* zawsze są stymulantami).

Parametr „k" służy przekształcaniu destymulant w stymulanty[46], co wydaje się oczywistą koniecznością. Możemy więc powiedzieć, że wyliczanie wskaźnika syntetycznego polega na dodawaniu wartości cech stymulant i odejmowaniu wartości cech destymulant. Warto przypomnieć, że jeżeli standaryzacja była przeprowadzona metodą *min-max*, wówczas parametr „k" można ignorować, gdyż – jak pamiętamy – wszystkie cechy po standaryzacji tą metodą są już symulantami.

[45] Jak wiemy, o tym przesądziła standaryzacja cech.

[46] Nieco inny sposób przekształcania destymulant w stymulanty zaprezentowany był przy okazji omawiania formuły standaryzacji *min-max*, w poprzednim podrozdziale.

Wyjaśnienia wymaga sens przytaczania dwóch, nieco tylko odmiennych formuł. Jedna z nich jest prostą sumą wartości cech, druga zaś ich średnią arytmetyczną. Przede wszystkim należy zwrócić uwagę, że z punktu widzenia problemu, do rozwiązania którego służy omawiana metoda, obydwie formuły prowadzą do identycznego wyniku oceny. Różnią się oczywiście liczbami, ale wnioski dotyczące oceny względnej powinny być takie same. Sądzę, że nikt nie ma wątpliwości co do pełnej zbieżności porządkowania monotonicznego ocenianych jednostek terytorialnych. Druga część zadania[47], mówiąca, że oceny powinny ustalać dystans dzielący daną jednostkę od wszystkich pozostałych, też jest spełniona. Te dwa ciągi liczb W_i oczywiście różnią się wartościami, ale proporcje między nimi pozostają takie same.

Po tym wyjaśnieniu zauważamy, że operowanie obydwoma formułami w pojedynczym przypadku oceny zbioru jednostek terytorialnych traci sens. Można się umówić, że w przypadku takim będzie stosowana formuła (11), jako nieco prostsza. Formuła ta jednak zawodzi w sytuacji, gdy chcielibyśmy niejako „po drodze" wyliczać cząstkowe wskaźniki syntetycznej oceny. Wyobraźmy sobie, podobne jak w naszym przykładzie, zadanie: dokonać oceny województw z punktu widzenia osiągniętego poziomu rozwoju. W naszym przyjętym sposobie postępowania dążymy do wyznaczenia wskaźnika syntetycznej oceny dla całego zjawiska (poziomu rozwoju). Ale moglibyśmy nieco zmodyfikować nasze zadanie. Mianowicie, ponieważ kategoria „rozwój jednostki terytorialnej" jest bardzo złożona, zasadna może być modyfikacja, aby dla celów pogłębionej analizy, najpierw dokonać oceny składowych tego rozwoju, np. rozwój społeczny, rozwój gospodarczy, rozwój przestrzenny, rozwój ekologiczny. Każda z tych składowych byłaby opisywana pewnym pakietem cech szczegółowych, na bazie których wyliczane byłyby cząstkowe wskaźniki syntetyczne, według omawianej procedury. Dopiero w następnym etapie wyznaczany byłby wskaźnik uogólniony. Ponieważ każda ze składowych rozwoju mogłaby być opisywana przez różną liczbę cech szczegółowych, zsumowanie ich wartości zestandaryzowanych prowadziłoby do wyników nieporównywalnych pomiędzy nimi. Aby uniknąć wspominanej nieporównywalności, wystarczy posłużyć się formułą (12), czyli średnią arytmetyczną wartości standaryzowanych. W takich zatem sytuacjach formuła ta jest stosowana.

Posługiwanie się w wyliczaniu syntetycznego wskaźnika oceny metodą sum standaryzowanych prowadzi do wartości ocen o właściwościach zależnych od zastosowanej metody standaryzacji:

1) W przypadku standaryzacji *zero-jedynkowej* suma wartości W_i (1=1, 2, 3,…, m) zawsze powinna wynosić zero (przy obydwóch formułach: 11 oraz 12).
2) W przypadku standaryzacji *min-max* i zastosowaniu do wyliczania wskaźnika syntetycznej oceny powyższej formuły (11), wartości ocen zawierać się będą w przedziale [0; n] (n – liczba cech), zaś przy zastosowaniu formuły (12) [0; 1].

Agregacja cech prowadząca do wyznaczenia syntetycznego wskaźnika oceny jest końcowym etapem omawianych metod. Nawiązując do naszego przykładu, w którym dążymy do dokonania oceny poziomu rozwoju województw, w etapie niniejszym (końcowym) należałoby na podstawie wartości standaryzowanych (tabele z ich wartościami ujmuje poprzedni podrozdział) wyliczyć syntetyczny wskaźnik oceny województw. Wyniki tych wyliczeń prezentują tabele 15-17.

[47] Przypomnijmy, że całe zadanie, do rozwiązania którego służą metody omawiane w tym rozdziale, jest następujące: należy skonstruować syntetyczny wskaźnik/ocenę, który pozwala monotonicznie uporządkować badane obiekty (u nas jednostki terytorialne) i jednocześnie mierzyć dystans dzielący dany obiekt od wszystkich pozostałych pod względem zadanego kryterium oceny.

Pytania/zadania kontrolne (dotyczą części empirycznej (wyliczeniowej) podrozdziału)

1. Jak można byłoby scharakteryzować właściwości ocen wyznaczonych w oparciu o poszczególne formuły standaryzacji?
2. W jakim zakresie możliwe jest ustalenie zastosowanej formuły standaryzacji na podstawie analizy wartości ocen syntetycznych?
3. Jak wiadomo, wskaźniki syntetyczne pozwalają na monotoniczne uporządkowanie ocenianych województw oraz na ustalanie dzielącego je dystansu z punktu widzenia zadanego kryterium oceny; w naszym przykładzie, z punktu widzenia osiągniętego poziomu rozwoju. Warto natomiast dodatkowo zastanowić się, na ile prawdziwe byłoby twierdzenie formułowane na podstawie otrzymanych wyników, że województwa znajdujące się na najwyższych pozycjach ocen są rzeczywiście zawsze wysoko rozwinięte.

Tabela 15

Wskaźniki sumaryczne ocen z wykorzystaniem standaryzacji zero-jedynkowej i min-max

Województwa	**Standaryzacja *zero-jedynkowa***					**Wsk. synt.**	**Standaryzacja *min-max***					**Wsk. synt.**
	1	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**	
dolnośląskie	0,26	0,63	1,82	0,66	-1,44	**0,282**	0,53	0,45	0,78	0,55	0,06	**0,475**
kujawsko-pomorskie	0,86	-0,48	-0,06	-0,78	-0,25	**-0,485**	0,35	0,13	0,26	0,09	0,38	**0,240**
lubelskie	-0,26	-0,59	-0,51	-0,78	0,70	**-0,185**	0,68	0,10	0,13	0,09	0,63	**0,323**
lubuskie	1,03	-0,92	0,03	0,53	-0,01	**-0,282**	0,30	0,00	0,28	0,51	0,44	**0,305**
łódzkie	-0,32	0,30	-0,90	-0,84	-0,01	**-0,229**	0,69	0,35	0,02	0,07	0,44	**0,314**
małopolskie	-1,38	0,96	0,09	1,97	0,46	**0,973**	1,00	0,55	0,30	0,97	0,56	**0,676**
mazowieckie	-1,21	2,51	-0,94	0,57	0,22	**0,715**	0,95	1,00	0,01	0,52	0,50	**0,596**
opolskie	0,06	-0,92	2,59	-0,91	0,70	**0,279**	0,58	0,00	1,00	0,05	0,63	**0,451**
podkarpackie	0,00	-0,92	-0,50	-0,92	1,41	**-0,185**	0,60	0,00	0,13	0,04	0,81	**0,318**
podlaskie	-0,86	-0,70	-0,97	-0,42	-0,73	**-0,391**	0,85	0,06	0,00	0,20	0,25	**0,273**
pomorskie	-0,11	0,41	-0,07	0,32	-1,20	**-0,087**	0,63	0,39	0,25	0,44	0,13	**0,367**
śląskie	-0,83	1,41	0,32	-0,61	-0,01	**0,386**	0,84	0,68	0,36	0,14	0,44	**0,492**
świętokrzyskie	0,46	-0,92	-0,90	-1,05	0,22	**-0,623**	0,47	0,00	0,02	0,00	0,50	**0,197**
warmińsko-mazurskie	2,07	-0,81	-0,79	0,54	-0,49	**-0,723**	0,00	0,03	0,05	0,51	0,31	**0,181**
wielkopolskie	-1,23	-0,03	0,43	-0,34	2,13	**0,684**	0,96	0,26	0,39	0,23	1,00	**0,568**
zachodniopomorskie	1,46	0,08	0,35	2,07	-1,68	**-0,129**	0,18	0,29	0,37	1,00	0,00	**0,367**

Źródło: obliczenia własne.

Tabela 16

Wskaźniki sumaryczne ocen z wykorzystaniem standaryzacji uproszczonej

Województwa	Standaryzacja uproszczona (x/s)					Wsk. synt.	Standaryzacja uproszczona (x/śred.)					Wsk. synt.
	1	2	3	4	5		1	2	3	4	5	
dolnośląskie	4,97	1,88	7,16	1,94	9,76	**3,154**	1,05	1,50	1,34	1,52	0,87	**4,177**
kujawsko-pomorskie	5,57	0,78	5,28	0,50	10,95	**2,387**	1,18	0,62	0,99	0,39	0,98	**1,794**
lubelskie	4,45	0,66	4,83	0,50	11,90	**2,688**	0,95	0,53	0,90	0,39	1,06	**1,943**
lubuskie	5,74	0,33	5,36	1,81	11,18	**2,590**	1,22	0,27	1,00	1,41	1,00	**2,464**
łódzkie	4,39	1,55	4,43	0,44	11,18	**2,643**	0,93	1,24	0,83	0,34	1,00	**2,479**
małopolskie	3,33	2,22	5,43	3,25	11,66	**3,846**	0,71	1,77	1,02	2,54	1,04	**5,660**
mazowieckie	3,50	3,77	4,40	1,85	11,42	**3,587**	0,74	3,01	0,82	1,44	1,02	**5,549**
opolskie	4,77	0,33	7,92	0,37	11,90	**3,151**	1,01	0,27	1,48	0,29	1,06	**2,088**
podkarpackie	4,71	0,33	4,84	0,36	12,61	**2,687**	1,00	0,27	0,91	0,28	1,13	**1,578**
podlaskie	3,85	0,55	4,37	0,86	10,47	**2,481**	0,82	0,44	0,82	0,67	0,93	**2,049**
pomorskie	4,59	1,66	5,27	1,60	9,99	**2,785**	0,98	1,33	0,99	1,25	0,89	**3,477**
śląskie	3,88	2,66	5,66	0,67	11,18	**3,258**	0,82	2,12	1,06	0,52	1,00	**3,878**
świętokrzyskie	5,17	0,33	4,43	0,23	11,42	**2,249**	1,10	0,27	0,83	0,18	1,02	**1,195**
warmińsko-mazurskie	6,78	0,44	4,55	1,82	10,71	**2,149**	1,44	0,35	0,85	1,42	0,96	**2,147**
wielkopolskie	3,47	1,22	5,77	0,94	13,32	**3,556**	0,74	0,97	1,08	0,74	1,19	**3,241**
zachodniopomorskie	6,17	1,33	5,69	3,35	9,52	**2,743**	1,31	1,06	1,07	2,62	0,85	**4,281**

Źródło: obliczenia własne.

Tabela 17

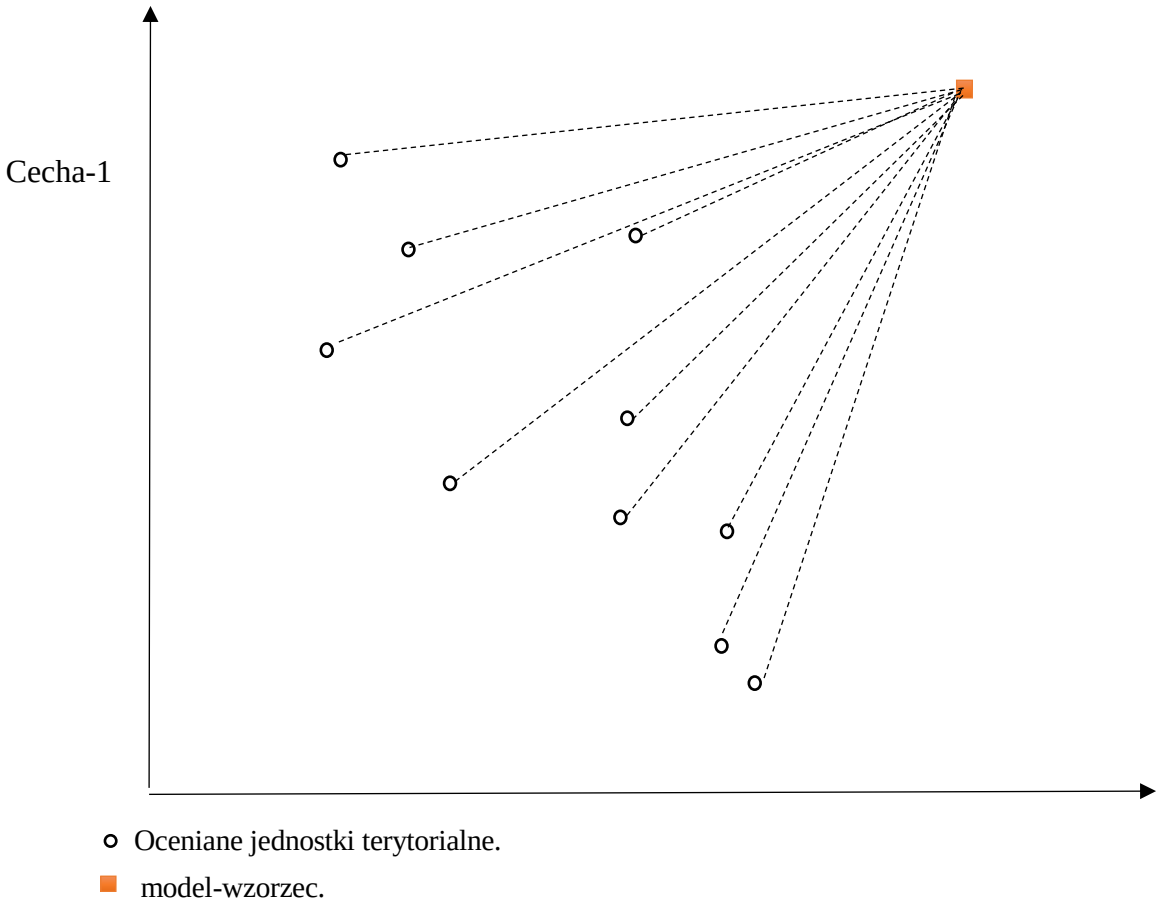
Zestawienie wskaźników syntetycznej oceny wyznaczanych z wykorzystaniem różnych formuł standaryzacji

Województwa	**0-1**	**Województwa**	**x/s**	**Województwa**	**x/śred**	**Województwa**	***min-max***
małopolskie	0,973	małopolskie	3,846	małopolskie	5,660	małopolskie	0,676
mazowieckie	0,715	mazowieckie	3,587	mazowieckie	5,549	mazowieckie	0,596
wielkopolskie	0,684	wielkopolskie	3,556	zachodniopomorskie	4,281	wielkopolskie	0,568
śląskie	0,386	śląskie	3,258	dolnośląskie	4,177	śląskie	0,492
dolnośląskie	0,282	dolnośląskie	3,154	śląskie	3,878	dolnośląskie	0,475
opolskie	0,279	opolskie	3,151	pomorskie	3,477	opolskie	0,451
pomorskie	-0,087	pomorskie	2,785	wielkopolskie	3,241	zachodniopomorskie	0,367
zachodniopomorskie	-0,129	zachodniopomorskie	2,743	łódzkie	2,479	pomorskie	0,367
lubelskie	-0,185	lubelskie	2,688	lubuskie	2,464	lubelskie	0,323
podkarpackie	-0,185	podkarpackie	2,687	warmińsko-mazurskie	2,147	podkarpackie	0,318
łódzkie	-0,229	łódzkie	2,643	opolskie	2,088	łódzkie	0,314
lubuskie	-0,282	lubuskie	2,590	podlaskie	2,049	lubuskie	0,305
podlaskie	-0,391	podlaskie	2,481	lubelskie	1,943	podlaskie	0,273
kujawsko-pomorskie	-0,485	kujawsko-pomorskie	2,387	kujawsko-pomorskie	1,794	kujawsko-pomorskie	0,240
świętokrzyskie	-0,623	świętokrzyskie	2,249	podkarpackie	1,578	świętokrzyskie	0,197
warmińsko-mazurskie	-0,723	warmińsko-mazurskie	2,149	świętokrzyskie	1,195	warmińsko-mazurskie	0,181

Źródło: obliczenia własne.

2.6. Metody syntetycznej oceny jednostek terytorialnych – agregacja cech metodą wzorca rozwoju

Na zupełnie innym podejściu metodycznym opiera się metoda wzorca rozwoju, zwana inaczej metodą modelową. Istota tej metody sprowadza się do budowy jednostki modelowej, z którą porównywane będą oceniane obiekty (w naszym przykładzie województwa). Budowa modelu polegać będzie na ustaleniu wartości modelowych (czyli w jakimś sensie najlepszych) dla każdej z cech diagnostycznych. W ten sposób nasz zbiór rozważanych jednostek terytorialnych, wraz z jednostką modelową, wynosił będzie m+1. Jednostce modelowej nadawać będziemy numer „0". Zauważmy, że każda jednostka terytorialna opisywana jest wektorem „n" liczb (n – liczba cech). W interpretacji geometrycznej jednostki te możemy traktować jako współrzędne punktu w przestrzeni n-wymiarowej. Spostrzeżenie to prowadzi nas do wniosku, że w budowie wskaźnika syntetycznego w oparciu o omawianą metodę wykorzystywana jest odległość punktu reprezentującego daną jednostkę terytorialną od punktu modelu-wzorca. Obrazowo ujmuje to rysunek 12, dla przykładu, w którym są tylko dwie cechy diagnostyczne (co pozawala zilustrować przykład na płaszczyźnie).



Rysunek 12. Przykładowa ilustracja relacji: jednostki terytorialne – model.
Źródło: opracowanie własne.

Przyjmujemy założenie, że odległość punktów wyznaczać będziemy w oparciu o standardowy, znany z geometrii klasycznej, wzór na odległość w przestrzeni fizycznej[48]. W naszym przypadku wzór ten mierzyć więc będzie odległość punktu reprezentującego i-tą jednostkę terytorialną od modelu wzorca. Przybiera on postać:

$$c_{i0} = \sqrt{\sum_{j=1}^{n}(t_{0j} - t_{ij})^2} \qquad (13)$$

gdzie:

c_{i0} – odległość punktów w przestrzeni „n" wymiarowej

t_{0j}, t_{ij} – wartości standaryzowane cech.

Z rysunku 12 wynika, że wzór (13) mierzy podległość punktów w układzie współrzędnych kartezjańskich, tzn. prostych wzajemnie prostopadłych. Możemy to uogólnić i powiedzieć, że wzór jest prawdziwy dla tzw. n-wymiarowych przestrzeni ortogonalnych, a więc takich, gdzie zmienne wyznaczające osie są względem siebie niezależne. Ta właściwość była właśnie podstawą sformułowanego wcześniej wymogu (patrz podrozdział 2.2): ***cechy nie mogą być ze sobą wysoko skorelowane*** (skorelowanie na istotnym statystycznie poziomie oznacza, że cechy nie są niezależne względem siebie).

Zauważmy, że wartość c_{i0} w interpretacji geometrycznej, mierząca odległość i-tego punktu od punktu modelowego, jest już wskaźnikiem mogącym uogólniać ocenę i-tej jednostki terytorialnej. Jest jednak pewna niedogodność tej miary, gdyż nie jest ona unormowana, tzn. że może przyjmować dowolnie duże wartości dodatnie (także 0). Dolną granicą c_{i0} jest 0, górna granica natomiast nie istnieje, teoretycznie może nią być $+\infty$[49]. Ponadto, jak łatwo zauważyć, c_{i0} ma właściwości destymulanty; im większa odległość od modelu-wzorca, tym gorsza jest ocena, i odwrotnie. To właśnie jest powodem, że miara ta zostanie nieco przekształcona, według formuły[50]:

$$D_i = 1 - \frac{c_{i0}}{c_{i0\{max\}}} \qquad (14)$$

gdzie:

D_i – miernik syntetycznej oceny i-tej jednostki terytorialnej,
$c_{i0\{max\}}$ – maksymalna wartość c_{i0} (w zbiorze dla i=1, 2, …m).

Wyliczane ze wzoru (14) wartości D_i są unormowane w przedziale [0-1]. D_i=0 oznacza najniższą z możliwych ocen, a D_i=1 dotyczy obiektu (obiektów) tożsamych z modelem-wzorcem. Oznacza to jednocześnie, że miara D_i ma właściwości stymulanty.

---

[48] Geometria klasyczna zwana jest również geometrią euklidesową, podobnie jak definiowana na jej gruncie przestrzeń (przestrzeń euklidesowa).

[49] Nieskończoność górnej granicy jest założeniem teoretycznym; tzn. wtedy, gdy $n \to \infty$.

[50] Warto poinformować, że w praktyce przyjmowany jest często nieco innych wzór, a mianowicie: $D_i = 1 - \frac{c_{i0}}{d}$, gdzie:
$d = \bar{c} + 2s_c$;
$\bar{c}$ – średnia arytmetyczna wartości c_{i0};
s_c – odchylenie standardowe c_{i0}.
Rezygnujemy z niego, gdyż wartości D_i nie są wtedy w pełni unormowane (m.in. D_i może przyjmować wartości ujemne).

Do wyjaśnienia została jeszcze budowa modelu-wzorca. Powinien on cechować się najlepszymi (optymalnymi) parametrami w stosunku do ocenianych jednostek terytorialnych. Powszechnie przyjmuje się jeden z najprostszych sposobów wyznaczania parametrów modelu, a mianowicie tzw. zasadę *mini-max*. Ze zbioru cech diagnostycznych dla stymulant przyjmuje się wartość maksymalną, zaś dla cech destymulant – wartość minimalną.

Podsumowując, cała procedura wyznaczania wskaźnika syntetycznej oceny D_i przebiega według następującej kolejności:

1) budowa modelu,

2) standaryzacja cech,

3) wyznaczanie wartości c_{i0},

4) obliczanie wskaźnika syntetycznego D_i.

Zaznaczyć wyraźnie należy, że wartości dla modelu wybieramy (zasada *mini-max*) na etapie cech diagnostycznych jeszcze nie standaryzowanych. Do standaryzacji każdej cechy mamy zatem nie „m", lecz „m+1" obiektów. Oznacza to, że wartości parametrów standaryzujących (średnia, odchylenie standardowe), wyliczone dla metody poprzednio omówionej (suma cech standaryzowanych), stają się nieprzydatne.

Korzystając z naszego dotychczasowego przykładu, w tabeli 18 zamieszczono wartości standaryzowane formułą *zero-jedynkową*, a w tabeli 19 zestawione zostały wyliczone wartości wskaźników syntetycznych (D_i).

Pytania/zadania kontrolne

1. Gdyby w naszym przykładzie cechy zostały wystandaryzowane metodą *min-max*, zastanów się, jaka mogłaby być maksymalna wartość wskaźnika c_{i0} (wzór 13)?
2. Co oznaczałby przypadek, w którym przy standaryzacji *min-max* wszystkie wartość modelu wynoszą 1?
3. Załóżmy, że cechy były standaryzowane formułą *zero-jedynkową*. Czy w tej sytuacji jest możliwe, że wartość modelowa którejś cechy wynosić będzie 0?
4. Nawiązując do przedstawionych wyżej wyliczeń (tabela 18 i 19), proponujemy rozważyć:
   a) Dlaczego wartości standaryzowane różnią się od analogicznych zamieszczonych w podrozdziale 2.4 (tabela 11)?
   b) Gdyby w tabeli 18 pominięto jej tytuł, czy możliwe byłoby ustalenie, które cechy są stymulantami, a które destymulantami? (należy skorzystać z porównania danych standaryzowanych z danymi wyjściowymi). Czy taka możliwość istniałaby również w przypadku dwóch innych formuł standaryzacji?
   c) Uzasadnić, że na podstawie tak wyznaczonych wskaźników syntetycznej oceny możliwe jest rozstrzygnięcie dylematu: jednostki zajmujące najwyższe pozycje w przeprowadzonej ocenie legitymują się czy też nie legitymują się rzeczywiście wysokim poziomem rozwoju?

Tabela 18

Cechy standaryzowane zero-jedynkowo wraz z modelem

Województwa	Dane wyjściowe					Dane standaryzowane				
	1	2	3	4	5	1	2	3	4	5
MODEL	11,6	34	48,6	281	56	-1,266	2,068	2,110	1,786	1,824
dolnośląskie	17,3	17	43,9	163	41	0,331	0,422	1,446	0,496	-1,429
kujawsko-pomorskie	19,4	7	32,4	42	46	0,920	-0,547	-0,180	-0,827	-0,344
lubelskie	15,5	6	29,6	42	50	-0,173	-0,644	-0,575	-0,827	0,523
lubuskie	20	3	32,9	152	47	1,088	-0,934	-0,109	0,376	-0,128
łódzkie	15,3	14	27,2	37	47	-0,229	0,131	-0,914	-0,882	-0,128
małopolskie	11,6	20	33,3	273	49	-1,266	0,712	-0,052	1,698	0,306
mazowieckie	12,2	34	27	155	48	-1,098	2,068	-0,943	0,408	0,089
opolskie	16,6	3	48,6	31	50	0,135	-0,934	2,110	-0,947	0,523
podkarpackie	16,4	3	29,7	30	53	0,079	-0,934	-0,561	-0,958	1,174
podlaskie	13,4	5	26,8	72	44	-0,762	-0,741	-0,971	-0,499	-0,778
pomorskie	16	15	32,3	134	42	-0,033	0,228	-0,194	0,179	-1,212
śląskie	13,5	24	34,7	56	47	-0,734	1,099	0,145	-0,674	-0,128
świętokrzyskie	18	3	27,2	19	48	0,528	-0,934	-0,914	-1,078	0,089
warmińsko-mazurskie	23,6	4	27,9	153	45	2,097	-0,837	-0,815	0,386	-0,561
wielkopolskie	12,1	11	35,4	79	56	-1,126	-0,160	0,244	-0,422	1,824
zachodniopomorskie	21,5	12	34,9	281	40	1,509	-0,063	0,174	1,786	-1,646
Średnia arytm.	16,1176	12,6471	33,6706	117,6471	47,5882	0,000	0,000	0,000	0,000	0,000
Odch. stand.	3,5678	10,3256	7,0767	91,4808	4,6106					
Wart. max.	23,6	34	48,6	281	56					
Wart. min.	11,6	3	26,8	19	40					

Źródło: obliczenia własne.

Tabela 19
Syntetyczne wskaźniki rozwoju województw (metoda wzorca rozwoju)

Układ alfabetyczny		Układ uporządkowany monotonicznie	
Województwa	**Wsk. D_i**	**Województwa**	**Wsk. D_i**
dolnośląskie	0,335	1. małopolskie	0,534
kujawsko-pomorskie	0,164	2. wielkopolskie	0,427
lubelskie	0,227	3. mazowieckie	0,408
lubuskie	0,211	4. śląskie	0,394
łódzkie	0,217	5. dolnośląskie	0,335
małopolskie	0,534	6. opolskie	0,296
mazowieckie	0,408	7. pomorskie	0,264
opolskie	0,296	8. lubelskie	0,227
podkarpackie	0,201	9. łódzkie	0,217
podlaskie	0,146	10. lubuskie	0,211
pomorskie	0,264	11. podkarpackie	0,201
śląskie	0,394	12. zachodniopomorskie	0,169
świętokrzyskie	0,104	13. kujawsko-pomorskie	0,164
warmińsko-mazurskie	0,059	14. podlaskie	0,146
wielkopolskie	0,427	15. świętokrzyskie	0,104
zachodniopomorskie	0,169	16. warmińsko-mazurskie	0,059

Źródło: obliczenia własne.

2.7. Metody syntetycznej oceny z wykorzystaniem obiektookresów

W niniejszym podrozdziale przedstawione zostanie rozwinięcie omawianych wcześniej metod syntetycznej oceny w kierunku ich zastosowań do analiz dynamicznych. Warto przypomnieć istotę poznanych już metod. Posłużmy się przykładem problemu, do rozwiązania którego służą. Otóż, mamy zdefiniowany terytorialny system społeczno-gospodarczy, w ramach którego funkcjonuje „m" jednostek terytorialnych. Zadaniem stojącym przed analitykiem jest dokonanie oceny tych jednostek z punktu widzenia zadanego kryterium. Dodatkowym założeniem jest, że odzwierciedlenie (opisanie) tego kryterium w oparciu o jeden wskaźnik nie jest możliwe; dla kompleksowości oceny, konieczne jest użycie szerszego zestawu cech szczegółowych. Naszym przykładowym, uprzednio rozwiązywanym, problemem była ocena województw w Polsce pod względem osiągniętego poziomu rozwoju. Terytorialnym systemem społeczno-gospodarczym w tym przykładzie jest Polska (cały kraj), jednostkami podlegającymi ocenie są województwa, zaś kryterium oceny jest poziom rozwoju.

Dla rozwiązania tego problemu służył algorytm postępowania, ujęty w podrozdziale 2.1, tabela 1. Każdy z punktów tego algorytmu był szczegółowo rozważany we wcześniejszych podrozdziałach, w których omawiane były kolejne punkty wspominanego algorytmu. W tym fragmencie zmodyfikujemy nieco problem rozwiązywany omawianymi już metodami syntetycznej oceny. Dotychczas znana ich wersja odnosiła się do problemu o charakterze statycznym. To znaczy kryterium oceny dotyczyło konkretnego okresu. Była to więc ocena na dany moment. Jakże często potrzeby praktyki dotyczą jednak sytuacji, w której należy dokonać przedmiotowej oceny dla różnych okresów, aby na tej podstawie stworzyć możliwość dokonywania porównań w czasie dla uchwycenie kierunku i siły dynamiki zmian rozważanego zjawiska (u nas jest nim kryterium wyznaczanej oceny). Biorąc pod uwagę nasz przykład, dokonanie oceny poziomu rozwoju województw np. w trzech różnych okresach[51] (a, b, c) istotnie rozszerzałoby wartości poznawcze podjętej analizy. Stworzylibyśmy wtedy podstawy nie tylko do oceny województw, w pewnym sensie oceny wyrywkowej, ale również podstawy do porównań w czasie zachodzących zmian w pozycjach (rankingu) województw[52], a przede wszystkim do wnioskowania indywidualnie dla każdego województwa o kierunku dynamiki (wzrost, stabilność czy spadek).

Możliwa, nasuwająca się podpowiedź jest następująca: dokonajmy oceny wedle powyższego algorytmu oddzielnie dla każdego z okresów a, b, oraz c. Otrzymamy w ten sposób trzy szeregi ocen. Oczywiście taka możliwość istnieje, jednak wynik takiego postępowania miałby bardzo ograniczone pole wnioskowania. Dlaczego? Odpowiedź jest dosyć prosta – moglibyśmy wtedy jedynie poznać miejsce w rankingu każdego z województw wraz z oceną dystansu dzielącego go od innych jednostek, ale oddzielnie dla każdego okresu. Możliwe byłoby również wnioskowanie o zmianach pozycji w rankingu w kolejnych okresach. Przy takim podejściu utracilibyśmy jednak możliwość ustalania dynamiki zamian, a przecież w założeniu przyjętej powyżej modyfikacji sformułowanego problemu o dynamikę właśnie chodziło. Musimy zauważyć, że w omawianym podejściu każde z województw legitymowałoby się trzema ocenami syntetycznymi, dla okresu a, b oraz c. Rzecz w tym, że otrzymane wskaźniki byłyby nieporównywalne między sobą. Inaczej ujmując, nie mielibyśmy prawa twierdzić na ich podstawie, że np. wyższa wartość wskaźnika dla okresu „b" w stosunku do wartości wskaźnika z okresu „a" oznacza wzrost zjawiska, które przyjęte zostało jako kryterium oceny. Nieporównywalność jest związana z czynnościami mającymi miejsce w punkcie 4 algorytmu, czyli dotyczącymi standaryzacji cech. Opisywane powyżej podejście oznacza, że cechy byłyby standaryzowane oddzielnie dla każdego przyjętego okresu. Nieco głębsza refleksja prowadzi do wniosku, że porównywalność w takim przypadku zaistniałaby jedynie w sytuacji, gdyby wartości parametrów standaryzujących były identyczne dla każdego z trzech okresów[53]. Nietrudno zauważyć, że byłoby to zbyt idealizujące założenie, w zasadzie nigdy niezachodzące w konkretnym przypadku praktyki. Weźmy pod uwagę standaryzację *zero-jedynkową*. Jeżeli np. w okresie „b" w stosunku do okresu „a" nastąpi istotne zmniejszenie

[51] Odwołanie się do trzech okresów należy traktować przykładowo. Równie dobrze mogłyby to być dwa okresy, albo cztery lub więcej. Liczba momentów czasu dla dokonywanych ocen zależy od przyjętych na wstępie każdego badana założeń wynikających z jego potrzeb.

[52] Koniecznym oczywiście jest założenie, że kryterium oceny jest opisywane w każdym z przyjętych okresów przez ten sam zestaw cech wyjściowych.

[53] Przypominamy: parametrami standaryzującymi dla kolejnych trzech formuł są:
a) *Zero-jedynkowa*: średnia arytmetyczna i odchylenie standardowe
b) Uproszczona: odchylenie standardowe lub średnia arytmetyczna
c) *Min-max*: wartość maksymalna i wartość minimalna.

różnic między województwami (zmniejszy się dyspersja pod względem danej cechy opisującej kryterium oceny), wówczas wartość mianownika również istotnie się zmniejszy. Oznacza to, że wartości standaryzowane zostaną także istotnie powiększone, co ostatecznie przeniesie się na powiększoną wartość wskaźnika syntetycznego dla każdego z województw w okresie „b". Wyższa wartość wskaźnika syntetycznego dla danego województwa w roku „b" wynikać będzie zatem z przyczyn czysto statystycznych, a nie z rzeczywiście zachodzących zmian. Podobne nieporównywalności wynikać mogą z opisywanych rozbieżności pozostałych parametrów standaryzujących między poszczególnymi latami (np. *max* oraz *min*).

Rozwiązaniem umożliwiającym porównywalność w czasie jest posługiwanie się obiektookresami. Obiektookresem nazywać będziemy daną jednostkę terytorialną z wartościami cech przypisanymi dla danego okresu. Jeżeli zatem decydujemy się na ocenę poziomu rozwoju województw np. w trzech okresach czasu, wówczas każde województwo występuje pod postacią trzech obiektookresów, a zatem w naszym przykładzie mamy 48 obiektookresów. Standaryzacji cech dokonujemy na jednym ich zbiorze (liczącym m=48), a nie oddzielnie na trzech zbiorach cech (każdy po m=16). W tej sytuacji całkowicie znika problem nieporównywalności wynikającej z różnic międzyokresowych w wartościach parametrów standaryzujących.

Wprowadzenie do naszych rozważań obiektookresów w zasadzie nie komplikuje procedury postępowania w stosowaniu omawianych ocen. Zachodzi jedynie modyfikacja kilku etapów (punktów algorytmu) – o czym będzie mowa poniżej. Macierz wartości wyjściowych cech (wskaźników szczegółowych oceny), spełniających stawiane wymogi (patrz punkt 2 algorytmu), w ogólnym zapisie symbolicznym, ujmuje tabela 20. W tabeli występuje k * m obiektookresów (k jest przyjętą liczbą okresów, dla których wyznaczane są wskaźniki syntetycznej oceny). Przyjęte oznaczenia:

$x_{ij}^{(t)}$ – wartość j-tej cechy w regionie i-tym, dla roku t

i = 1, 2, … m (m - liczba jednostek terytorialnych); j = 1, 2, … n (n- liczba cech)

t = 1, 2, … k (liczba okresów, np. lat).

Tabela 20

Zestaw wskaźników szczegółowych do oceny jednostek terytorialnych z punktu widzenia zadanego kryterium, według obiektookresów (cechy niestandaryzowane)

Lata (t)	**Region**	**Cecha** / **Obiekto--okresy**	**1**	**2**	**3**	**….**	**n**
1	1	1	$x_{11}^{(1)}$	$x_{12}^{(1)}$	$x_{13}^{(1)}$	….	$x_{1n}^{(1)}$
	2	2	$x_{21}^{(1)}$	$x_{22}^{(1)}$	$x_{23}^{(1)}$	….	$x_{2n}^{(1)}$
	3	3	$x_{31}^{(1)}$	$x_{32}^{(1)}$	$x_{33}^{(1)}$	….	$x_{3n}^{(1)}$
	….	….	….	….	….	….	….
	m	m	$x_{m1}^{(1)}$	$x_{m2}^{(1)}$	$x_{m3}^{(1)}$	….	$x_{mn}^{(1)}$
2	1	m+1	$x_{11}^{(2)}$	$x_{12}^{(2)}$	$x_{13}^{(2)}$	….	$x_{1n}^{(2)}$
	2	m+2	$x_{21}^{(2)}$	$x_{22}^{(2)}$	$x_{23}^{(2)}$	….	$x_{2n}^{(2)}$
	3	m+3	$x_{31}^{(2)}$	$x_{32}^{(2)}$	$x_{33}^{(2)}$	….	$x_{3n}^{(2)}$
	….	….	….	….	….	….	….
	m	m+m	$x_{m1}^{(2)}$	$x_{m2}^{(2)}$	$x_{m3}^{(2)}$	….	$x_{mn}^{(2)}$
3	1	2m+1	$x_{11}^{(3)}$	$x_{12}^{(3)}$	$x_{13}^{(3)}$	….	$x_{1n}^{(3)}$
	2	2m+2	$x_{21}^{(3)}$	$x_{22}^{(3)}$	$x_{23}^{(3)}$	….	$x_{2n}^{(3)}$
	3	2m+3	$x_{31}^{(3)}$	$x_{32}^{(3)}$	$x_{33}^{(3)}$	….	$x_{3n}^{(3)}$
	….	….	….	….	….	….	….
	m	2m+m	$x_{m1}^{(3)}$	$x_{m2}^{(3)}$	$x_{m3}^{(3)}$	….	$x_{mn}^{(3)}$
….	….		….	….	….	….	….
k	1	(k-1)*m+1	$x_{11}^{(k)}$	$x_{12}^{(k)}$	$x_{13}^{(k)}$	….	$x_{1n}^{(k)}$
	2	(k-1)*m+2	$x_{21}^{(k)}$	$x_{22}^{(k)}$	$x_{23}^{(k)}$	….	$x_{2n}^{(k)}$
	3	(k-1)*m+3	$x_{31}^{(k)}$	$x_{32}^{(k)}$	$x_{33}^{(k)}$	….	$x_{3n}^{(k)}$
	….	….	….	….	….	….	….
	m	(k-1)*m+m	$x_{m1}^{(k)}$	$x_{m2}^{(k)}$	$x_{m3}^{(k)}$	….	$x_{mn}^{(k)}$

Źródło: opracowanie własne.

Należy zauważyć, że wskazany w ostatnim wierszu numer obiektookresu, jako (k-1)*m+m, po wykonaniu działań jest równy k * m („*” jest w tym zapisie znakiem mnożenia).

W następnym etapie dokonujemy wyboru cech diagnostycznych według reguł opisanych w punkcie 3 algorytmu. Wyliczana jest m.in. macierz korelacji, będąca podstawą selekcji cech spełniających warunek: *nie są ze sobą wysoko skorelowane*. Współczynniki korelacji wyliczane są na podstawie wszystkich obiektookresów, a więc w naszym przykładzie, przy założeniu trzech okresów, liczebność cech nie wynosiłaby 16, a 48. Wybór cech diagnostycznych dokonywany jest według znanych już procedur metody grafu lub dendrytu.

Standaryzacja cech wyselekcjonowanych ze zbioru cech wyjściowych jest kolejnym etapem postępowania. Jak już wcześniej wyjaśniano, właśnie standaryzacja na obiektookresach zapewnia porównywalność w czasie wyznaczanych wskaźników ocen syntetycznych. W ogólnym ujęciu, formuły standaryzacji pozostają te same. Korekcie podlegają jedynie zapisy wzorów uwzględniające obiektookresy.

Standaryzacja *zero-jedynkowa*

$$g_{ij}^{(t)} = \frac{x_{ij}^{(t)} - \bar{x}_j}{s_j}$$

Standaryzacja uproszczona

$$g_{ij}^{(t)} = \frac{x_{ij}^{(t)}}{s_j} \qquad\qquad g_{ij}^{(t)} = \frac{x_{ij}^{(t)}}{\bar{x}_j}$$

Standaryzacja *min-max*

$$g_{ii}^{(t)} = \begin{cases} \dfrac{x_{ij}^{(t)} - x_{min}}{x_{max} - x_{min}} & \text{dla stymulant} \\ \dfrac{x_{max} - x_{ij}^{(t)}}{x_{max} - x_{min}} & \text{dla destymulant} \end{cases}$$

Wyjaśnienia wymagają przyjęte oznaczenia. Dla uniknięcia niejednoznaczności symbol zmiennej standaryzowanej zmieniono z „*t*" na „*g*", przyjmując:

$g_{ij}^{(t)}$ standaryzowana wartość j-tej cechy w roku t, dla i-tej jednostki terytorialnej

Pozostałe oznaczenia pozostają bez zmian.

Celem niewprowadzania dodatkowych oznaczeń przyjmijmy, że nadal przez „n" oznaczać będziemy liczbę wyselekcjonowanych cech. Macierz cech standaryzowanych ujmuje tabela 21. W ostatniej kolumnie tej tabeli wpisano wartości wskaźnika syntetycznego. Przykładowo przyjęto, że wskaźnik syntetyczny liczony był według metody modelowej (oznaczenie wskaźnika przez „D"). Analogicznie zapis wyglądałby w przypadku przyjęcia metody sumy cech standaryzowanych (w miejsce „D" pojawiłby się symbol „W").

Tabela 21

Zestaw wskaźników szczegółowych do oceny jednostek terytorialnych z punktu widzenia zadanego kryterium, według obiektookresów (cechy standaryzowane)

Lata (t)	**Cecha / Region**	**Obiekto--okresy**	**1**	**2**	**3**	**….**	**n**	**Wsk. synt.**
1	1	1	$g_{11}^{(1)}$	$g_{12}^{(1)}$	$g_{13}^{(1)}$	….	$g_{1n}^{(1)}$	$\boldsymbol{D_1^{(1)}}$
	2	2	$g_{21}^{(1)}$	$g_{22}^{(1)}$	$g_{23}^{(1)}$	….	$g_{2n}^{(1)}$	$\boldsymbol{D_2^{(1)}}$
	3	3	$g_{31}^{(1)}$	$g_{32}^{(1)}$	$g_{33}^{(1)}$	….	$g_{3n}^{(1)}$	$\boldsymbol{D_3^{(1)}}$
	….	….	….	….	….	….	….	**….**
	m	m	$g_{m1}^{(1)}$	$g_{m2}^{(1)}$	$g_{m3}^{(1)}$	….	$g_{mn}^{(1)}$	$\boldsymbol{D_m^{(1)}}$
2	1	m+1	$g_{11}^{(2)}$	$g_{12}^{(2)}$	$g_{13}^{(2)}$	….	$g_{1n}^{(2)}$	$\boldsymbol{D_1^{(2)}}$
	2	m+2	$g_{21}^{(2)}$	$g_{22}^{(2)}$	$g_{23}^{(2)}$	….	$g_{2n}^{(2)}$	$\boldsymbol{D_2^{(2)}}$
	3	m+3	$g_{31}^{(2)}$	$g_{32}^{(2)}$	$g_{33}^{(2)}$	….	$g_{3n}^{(2)}$	$\boldsymbol{D_3^{(2)}}$
	….	….	….	….	….	….	….	**….**
	m	m+m	$g_{m1}^{(2)}$	$g_{m2}^{(2)}$	$g_{m3}^{(2)}$	….	$g_{mn}^{(2)}$	$\boldsymbol{D_m^{(2)}}$
3	1	2m+1	$g_{11}^{(3)}$	$g_{12}^{(3)}$	$g_{13}^{(3)}$	….	$g_{1n}^{(3)}$	$\boldsymbol{D_1^{(3)}}$
	2	2m+2	$g_{21}^{(3)}$	$g_{22}^{(3)}$	$g_{23}^{(3)}$	….	$g_{2n}^{(3)}$	$\boldsymbol{D_2^{(3)}}$
	3	2m+3	$g_{31}^{(3)}$	$g_{32}^{(3)}$	$g_{33}^{(3)}$	….	$g_{3n}^{(3)}$	$\boldsymbol{D_3^{(3)}}$
	….	….	….	….	….	….	….	**….**
	m	2m+m	$g_{m1}^{(3)}$	$g_{m2}^{(3)}$	$g_{m3}^{(3)}$	….	$g_{mn}^{(3)}$	$\boldsymbol{D_m^{(3)}}$
….	….		….	….	….	….	….	**….**
k	1	(k-1)*m+1	$g_{11}^{(k)}$	$g_{12}^{(k)}$	$g_{13}^{(k)}$	….	$g_{1n}^{(k)}$	$\boldsymbol{D_1^{(k)}}$
	2	(k-1)*m+2	$g_{21}^{(k)}$	$g_{22}^{(k)}$	$g_{23}^{(k)}$	….	$g_{2n}^{(k)}$	$\boldsymbol{D_2^{(k)}}$
	3	(k-1)*m+3	$g_{31}^{(k)}$	$g_{32}^{(k)}$	$g_{33}^{(k)}$	….	$g_{3n}^{(k)}$	$\boldsymbol{D_3^{(k)}}$
	….	….	….	….	….	….	….	**….**
	m	(k-1)*m+m	$g_{m1}^{(k)}$	$g_{m2}^{(k)}$	$g_{m3}^{(k)}$	….	$g_{mn}^{(k)}$	$\boldsymbol{D_m^{(k)}}$

Źródło: opracowanie własne.

Po wyliczeniu wskaźników syntetycznych na obiektookresach dokonujemy ich zestawienia poszczególnymi okresami, co ujmuje tabela 22. Otrzymujemy w ten sposób wskaźniki oceny, których wartości są porównywalne w czasie.

Tabela 22
Wskaźniki syntetycznej oceny wyznaczane na obiektookresach

lata / region	1	2	3	….	k
1	$D_1^{(1)}$	$D_1^{(2)}$	$D_1^{(3)}$	….	$D_1^{(k)}$
2	$D_2^{(1)}$	$D_2^{(2)}$	$D_2^{(3)}$	….	$D_2^{(k)}$
3	$D_3^{(1)}$	$D_3^{(2)}$	$D_3^{(3)}$	….	$D_3^{(k)}$
….	….	….	….	….	….
m	$D_m^{(1)}$	$D_m^{(2)}$	$D_m^{(3)}$	….	$D_m^{(k)}$

Źródło: opracowanie własne.

Poniżej prześledzimy konkretny przykład, w którym:

1) Przedmiotem oceny jest siedem największych polskich miast:
   - ✓ Gdańsk
   - ✓ Katowice
   - ✓ Kraków
   - ✓ Łódź
   - ✓ Poznań
   - ✓ Warszawa
   - ✓ Wrocław.
2) Kryterium oceny jest poziom społeczno-gospodarczego rozwoju, opisywany czterema cechami diagnostycznymi[54]:
   - ✓ Udział pracujących w sektorze usług
   - ✓ Techniczne uzbrojenie pracy (wartość środków trwałych przypadająca na pracującego
   - ✓ Udział wynagrodzenia w średniej w kraju
   - ✓ Wskaźnik salda migracji (saldo migracji na 1 000 ludności).
3) Ocena dotyczy lat:
   - ✓ 2011
   - ✓ 2015
   - ✓ 2019.

Przyjmujemy założenie, że:

1) Standaryzacja dokonana zostanie formułą *zero-jedynkową*.
2) Wskaźnik syntetyczny wyliczony zostanie za pomocą metody modelowej.

Wartości liczbowe cech, wraz z wartościami dla modelu, prezentuje tabela 23.

[54] Przyjmujemy dla uproszczenia postępowania założenie, że wymienione cechy są już po selekcji wskazywanej punktem 3 algorytmu (nie są ze sobą wysoko skorelowane). Zauważamy dodatkowo, że wszystkie cechy są stymulantami.

Tabela 23
Wskaźniki poziomu rozwoju miast

Nr obiektookresu	Obiektookres*)	Udział pracujących w sektorze usług	Techniczne uzbrojenie pracy	Udział wynagrodzenia w średniej w kraju	Wskaźnik salda migracji
0	model	86,8	364	156,8	6,2
1	Gdańsk-11	73,9	210	119,6	-0,2
2	Katowice-11	75,1	152	132,9	-3,5
3	Kraków-11	75,7	156	103,2	0,5
4	Łódź-11	71,8	124	94,4	-2,3
5	Poznań-11	75,6	198	111,0	-5,4
6	Warszawa-11	85,3	295	136,7	2,3
7	Wrocław-11	77,9	157	107,0	1,1
8	Gdańsk-15	76,9	279	140,1	1,9
9	Katowice-15	77,0	201	151,4	-3,3
10	Kraków-15	78,7	170	120,9	2,0
11	Łódź-l5	73,4	150	111,7	-2,0
12	Poznań-15	78,3	225	126,8	-3,4
13	Warszawa-15	86,8	320	156,8	5,2
14	Wrocław-15	80,0	199	126,3	3,0
15	Gdańsk-19	73,9	266	116,7	4,2
16	Katowice-19	75,1	194	117,9	-2,5
17	Kraków-19	75,7	171	111,0	6,1
18	Łódź-19	71,8	165	98,9	-1,7
19	Poznań-19	75,6	234	110,8	-3,5
20	Warszawa-19	85,3	364	133,1	6,2
21	Wrocław-19	77,9	208	110,4	2,3
Średnia arytmetyczna		77,66	218,27	122,47	0,60
Odchylenie standardowe		4,558	68,933	17,785	3,613
Wartość maksymalna		86,8	364	156,8	6,2

Źródło: Dane GUS.
*) liczba dodawana do nazwy miasta wskazuje na dwie ostatnie cyfry danego roku.

W tabeli 24 zamieszczone zostały wartości standaryzowane cech wraz z wartościami wskaźnika syntetycznego.

Tabela 24

Wartości standaryzowane cech metodą zero-jedynkową

Nr obiektookresu	Obiektookres*)	Cecha-1	Cecha-2	Cecha-3	Cecha-4	C_{i0}	D_i
0	model	2,005	2,114	1,930	1,550	model	model
1	Gdańsk-11	-0,825	-0,120	-0,162	-0,221	4,529	0,291
2	Katowice-11	-0,561	-0,961	0,586	-1,135	5,006	0,216
3	Kraków-11	-0,430	-0,903	-1,084	-0,028	5,158	0,192
4	Łódź-11	-1,285	-1,368	-1,578	-0,803	6,387	0,000
5	Poznań-11	-0,452	-0,294	-0,645	-1,661	5,365	0,160
6	Warszawa-11	1,676	1,113	0,800	0,471	1,885	0,705
7	Wrocław-11	0,053	-0,889	-0,870	0,138	4,761	0,255
8	Gdańsk-15	-0,167	0,881	0,991	0,360	2,922	0,543
9	Katowice-15	-0,145	-0,251	1,626	-1,080	4,150	0,350
10	Kraków-15	0,228	-0,700	-0,088	0,388	4,063	0,364
11	Łódź-15	-0,934	-0,990	-0,606	-0,720	5,465	0,144
12	Poznań-15	0,141	0,098	0,243	-1,107	4,177	0,346
13	Warszawa-15	2,005	1,476	1,930	1,273	0,696	0,891
14	Wrocław-15	0,514	-0,280	0,215	0,664	3,418	0,465
15	Gdańsk-19	-0,825	0,692	-0,325	0,996	3,927	0,385
16	Katowice-19	-0,561	-0,352	-0,257	-0,858	4,822	0,245
17	Kraków-19	-0,430	-0,686	-0,645	1,522	4,517	0,293
18	Łódź-19	-1,285	-0,773	-1,325	-0,637	5,877	0,080
19	Poznań-19	-0,452	0,228	-0,656	-1,135	4,847	0,241
20	Warszawa-19	1,676	2,114	0,598	1,550	1,373	0,785
21	Wrocław-19	0,053	-0,149	-0,679	0,471	4,112	0,356

Źródło: obliczenia własne.

*) liczba dodawana do nazwy miasta wskazuje na dwie ostatnie cyfry danego roku.

Tabela 25 ujmuje zestawienie wartości wskaźnika syntetycznej oceny poziomu rozwoju dużych miast dla lat 2011, 2015, 2019.

Tabela 25
Wskaźniki syntetycznej oceny poziomu rozwoju miast w latach 2010, 2014, 2018

Miasto	2011	2015	2019
Gdańsk	0,291	0,543	0,385
Katowice	0,216	0,350	0,245
Kraków	0,192	0,364	0,293
Łódź	0	0,144	0,080
Poznań	0,160	0,346	0,241
Warszawa	0,705	0,891	0,785
Wrocław	0,255	0,465	0,356

Źródło: obliczenia własne.

Tabela 26
Wskaźniki syntetycznej oceny poziomu rozwoju miast w latach 2010, 2014, 2018 (uporządkowane monotonicznie względem 2010 roku)

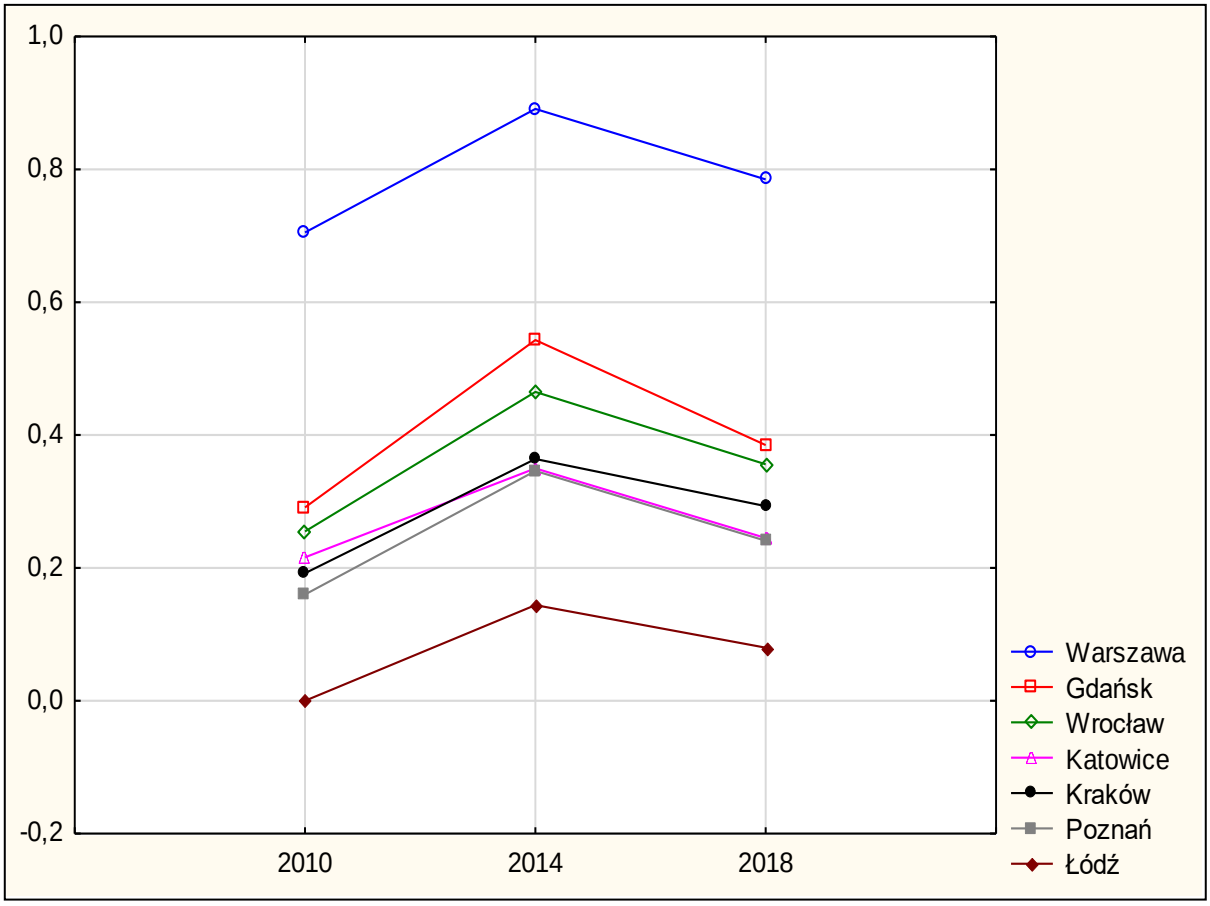
Miasto	2011	2015	2019
Warszawa	0,705	0,891	0,785
Gdańsk	0,291	0,543	0,385
Wrocław	0,255	0,465	0,356
Katowice	0,216	0,350	0,245
Kraków	0,192	0,364	0,293
Poznań	0,160	0,346	0,241
Łódź	0	0,144	0,080

Źródło: obliczenia własne.

Rysunek 13 stanowi graficzne uzupełnienie informacji prezentowanych w tabelach 25 oraz 26.

Jak pamiętamy, wyliczone na obiektookresach wskaźniki syntetycznej oceny poziomu rozwoju miast stwarzają nam możliwość dokonywania oceny zmian w czasie. Otrzymane wyniki pozwalają m.in. wysunąć następujące kluczowe wnioski:

1) Wśród branych pod uwagę miast, Warszawa okazuje się miastem zdecydowanie wyprzedzającym pozostałe jednostki miejskie – co zresztą nie może stanowić zaskoczenia.
2) Wyraźnie odstającym in minus, we wszystkich przyjętych latach, jest miasto Łódź.
3) Prawidłowością rozwoju wszystkich miast jest relatywnie wysokie tempo wzrostu w latach 2011-2015 oraz spadek poziomu rozwoju w kolejnym okresie (2015-2019).
4) We wszystkich jednak miastach poziom rozwoju w 2019 roku był wyższy od poziomu w roku 2011.



Rysunek 13. Dynamika rozwoju dużych miast w latach 2010, 2014, 2018.
Źródło: opracowanie własne.

Pytania/zadania kontrolne

1. Przemyśl odpowiedź na pytanie: Jak sformułowany musi być problem dotyczący oceny jednostek terytorialnych, żeby w jego rozwiązaniu istniała konieczność operowania obiektookresami?
2. Celem przekonania się o słuszności posługiwania się obiektookresami w ocenie jednostek terytorialnych w różnych okresach czasu proponuję w powyższym przykładzie dokonać ponownych obliczeń, jednak z zastosowaniem odrębnej standaryzacji cech dla każdego z trzech przyjętych okresów.
3. Proszę o przemyślenie następującego zagadnienia: jeżeli w procedurze oceny dynamicznej pominięto zastosowanie obiektookresów, to zatem która z poznanych standaryzacji wydaje się bardziej wrażliwa na zniekształcenie ocen dla porównań międzyokresowych?

3. Część trzecia: Metody taksonomiczne

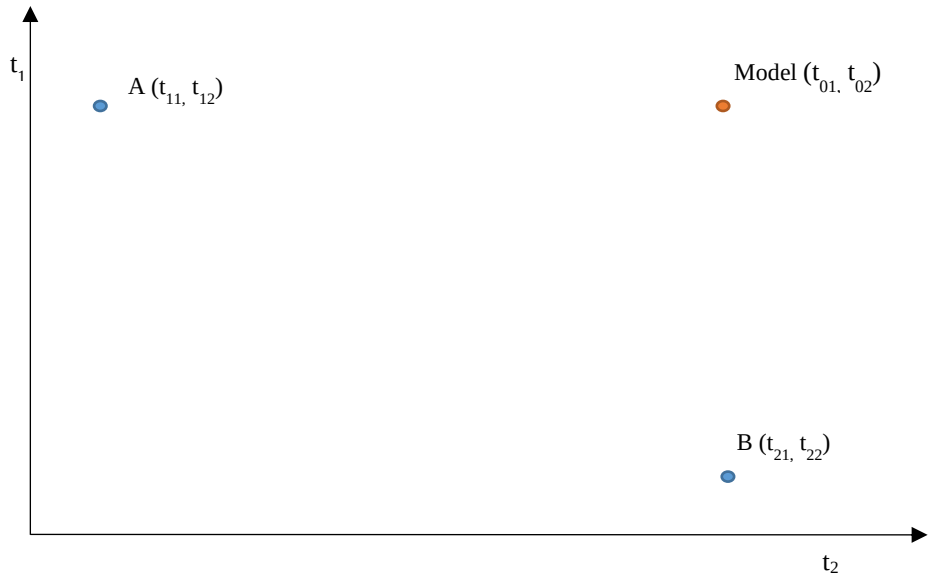
3.1. Metody taksonomiczne – istota metod i macierz odległości taksonomicznych

Metody taksonomiczne – podobnie jak metody syntetycznej oceny – należą do tej samej grupy metod określanych nazwą: wielowymiarowa analiza porównawcza. Są to metody klasyfikacji zbioru danych obiektów z punktu widzenia określonego kryterium. Metody syntetycznej oceny, oprócz wcześniej omawianych ich zastosowań, są metodami klasyfikacji liniowej, zaś metody taksonomiczne, których charakterystykę rozpoczynamy, są metodami klasyfikacji wielowymiarowej. Najpierw wyjaśniona więc zostanie istota jednego i drugiego rodzaju klasyfikacji.

Klasyfikacja oparta na metodach syntetycznej oceny **jest klasyfikacją jednowymiarową (liniową)**. Należy zauważyć, że ich istota sprowadza się do tego, iż na podstawie wielu cech szczegółowych charakteryzujących kryterium oceny konstruowany jest jeden syntetyczny wskaźnik; czyli, że w pewnym sensie wielowymiarowość (wiele cech szczegółowych) redukowana jest do jednego wymiaru (wskaźnik syntetyczny). Albo nieco inaczej: w punkcie wyjścia, każda jednostka terytorialna opisywana jest kilkoma liczbami (wartościami różnych cech) po to, aby następnie na ich podstawie przypisać każdej jednostce jeden wskaźnik (odpowiednio syntetyzujący te liczby). W efekcie tego jesteśmy w stanie dokonać jednoznacznego uszeregowania ocenianych jednostek, od najlepszej do najgorszej w świetle zadanego kryterium, lub odwrotnie (od najgorszej do najlepszej). Uszeregowanie to jest już rodzajem klasyfikacji. Ponadto, stwarza możliwość dodatkowej, bardziej uogólnionej klasyfikacji, a mianowicie poprzez przyjęcie przedziałów liczbowych dla otrzymanych wartości wskaźnika syntetycznego możemy ustalić kilka klas ocenianych jednostek terytorialnych; np. najwyżej rozwinięte, średnio rozwinięte, najniżej rozwinięte[55]. Tak więc, jak już zauważono, omawiana procedura prowadzi do klasyfikacji liniowej, gdyż wielowymiarowość w punkcie wyjścia sprowadzana jest do jednego wymiaru (wskaźnika syntetyzującego wartości cech szczegółowych). Jak pamiętamy, jest to konieczne dla stworzenia możliwości jednoznacznej oceny jednostek terytorialnych.

Koniecznego podkreślenia wymaga bardzo ważna właściwość tak ustalonej klasyfikacji. Opierając się na znanej już procedurze budowy wskaźnika syntetycznej oceny, możemy powiedzieć, że wskaźnik ten jest pewnego rodzaju uśrednieniem wartości cech szczegółowych i dotyczy to obydwóch metod agregacji. Oznacza to, że formułowana ocena każdej jednostki terytorialnej jest wypadkową jej mocnych i słabszych stron opisywanych pakietem *wskaźników szczegółowych; inaczej jeszcze ujmując, jest wynikiem „bilansowania"* tego, co ją pozytywnie wyróżnia z negatywnymi jej właściwościami, w relacji do innych, ocenianych jednostek terytorialnych. W sposób obrazowy ilustruje to rysunek 14.

[55] Oczywiście, że liczba przedziałów może być znacznie większa niż tu wymieniana.



Rysunek 14. Ilustracja klasyfikacji liniowej.
Źródło: opracowanie własne.

Dla konstrukcji tego rysunku i ułatwienia przeprowadzanego rozumowania przyjmijmy następujące założenia, które w niczym nie obniżają ogólności rozważań, a przede wszystkim nie wypaczają sensu konkluzji końcowej:

1) Mamy dwie jednostki terytorialne (A i B) oceniane z punktu widzenia jakiegoś kryterium[56].
2) Zakładamy, że kryterium oceny jest opisywane dwoma cechami, co pozwala na ilustrowanie przykładu w układzie współrzędnych prostokątnych na płaszczyźnie. Po standaryzacji cech, każda jednostka jest reprezentowana przez punkt na płaszczyźnie o dwóch współrzędnych A $(t_{11,}\ t_{12})$; B $(t_{21,}\ t_{22})$.
3) Przyjmujemy, że wskaźnik syntetyczny jest wyliczany w oparciu o metodę wzorca rozwoju. Punkt modelowy ma współrzędne: model $(t_{01,}\ t_{02})$.
4) Oznaczmy przez C_A odległość punku A od modelu, zaś przez C_B odległość punktu B od modelu. Jak pamiętamy, odległości te mają już właściwości wskaźnika syntetycznego.
5) Przyjmijmy dodatkowo[57], że (patrz rysunek 14):
   a) $t_{01} = t_{02}$
   b) $t_{01} = t_{21}$ oraz $t_{02} = t_{12}$
   c) $t_{11} = t_{22}$ oraz $t_{12} = t_{21}$ wtedy również: $t_{12} = t_{21} = t_{01} = t_{02}$.

[56] Konkretna postać tego kryterium nie jest ważna dla dalszych rozważań. Może to być dowolne kryterium oceny, mające jakiś sens merytoryczny.
[57] Założenia te ułatwiają przeprowadzenie rozumowania i – jak już wskazywano – nie wypaczają sensu konkluzji końcowej, do której będziemy zmierzać.

Przechodząc do analizy rysunku 14, przy przyjętych założeniach zwróćmy uwagę na rozłożenie punktów (reprezentują jednostkę A oraz B) w układzie współrzędnych. Wynika z tego, **że jednostki te diametralnie różnią się między sobą.** Dla wyeksponowania prezentowanych spostrzeżeń operować będziemy skrajnymi określeniami. Jednostka „A" cechuje się ekstremalnie wysoką wartością cechy pierwszej i ekstremalnie niską wartością cechy drugiej. Przypadek jednostki „B" jest dokładnie odwrotny: ekstremalnie niska wartość cechy pierwszej i bardzo wysoka wartość cechy drugiej. To właśnie jest podstawą twierdzenia, że oceniane jednostki diametralnie różnią się między sobą.

Opierając się na przyjętych oznaczeniach i pamiętając o założeniach, przeprowadźmy odpowiednie wyliczenia. Najpierw jednak przypomnijmy z wcześniejszych rozważań wzór na odległość punktów reprezentujących oceniane jednostki terytorialne:

$$c_i = \sqrt{\sum_{j=1}^{n}(t_{0j} - t_{ij})^2} \qquad \text{„i" przybiera wartość A oraz B}$$

$C_A = ((t_{01}\text{-}t_{11})^2 + (t_{02}\text{-}t_{12})^2)^{1/2} = ((t_{01}\text{-}t_{11})^2)^{1/2}$

Pamiętając z założenia, że (patrz również rysunek): $t_{02} = t_{12}$, otrzymujemy:

$C_A = ((t_{01}\text{-}t_{11})^2)^{1/2}$

$C_B = ((t_{01}\text{-}t_{21})^2 + (t_{02}\text{-}t_{22})^2)^{1/2} = ((t_{02}\text{-}t_{22})^2)^{1/2}$

Pamiętając z założenia (patrz również rysunek), że: $t_{01} = t_{21}$, otrzymujemy:

$C_B = ((t_{02}\text{-}t_{22})^2)^{1/2}$

Pamiętając z kolei z założenia (patrz również rysunek), że $t_{11} = t_{22}$ oraz $t_{12} = t_{21}$ (wtedy również: $t_{12} = t_{21} = t_{01} = t_{02}$), otrzymujemy ostatecznie:

$C_A = C_B$

Wykorzystując przyjęte wyżej założenia, proponujemy przeprowadzenie analogicznego rozumowania w odniesieniu do drugiej metody wyznaczania wskaźnika syntetycznego (metody sumy cech standaryzowanych). Wzór ten, jak pamiętamy, ma postać[58]:

$$W_i = \sum_{j=1}^{n} kt_{ij} \quad \text{„i" przebiega od A do B}$$

W finalnym wyniku powinniśmy otrzymać: $W_A = W_B$

Ostateczna konkluzja, do której zmierzaliśmy jest następująca: pomimo, że oceniane jednostki znacząco różnią się pozycją odzwierciedlaną wartościami cech szczegółowych, ich ocena w oparciu o procedurę klasyfikacji liniowej jest identyczna.

[58] Dla przećwiczenia proponujemy wykorzystanie metody sum standaryzowanych do pełnego prześledzenia wyniku prezentowanego rozumowania (wynik musi być identyczny, tzn.: $W_A = W_B$).

Klasyfikacja wielowymiarowa, do której służą omawiane dalej metody taksonomiczne, polega na pozycjonowaniu badanych jednostek terytorialnych w pełnym wymiarze stanowionym przez liczbę cech szczegółowych. W odniesieniu do naszego powyższego przykładu wynik klasyfikacji wielowymiarowej byłby następujący: jednostki A i B diametralnie różnią się między sobą; są do siebie skrajnie niepodobne – stwierdzenie to będzie w pełni udokumentowane dalszymi rozważaniami.

Przejdźmy zatem do zaprezentowania istoty metod taksonomicznych.

Istotą omawianych dalej metod **jest podział niejednorodnego zbioru obiektów (u nas jednostek terytorialnych) w podzbiory (grupy) bardziej do siebie podobne**. Grupowanie to dokonywane jest zawsze z punktu widzenia określonego kryterium. Posłużmy się przykładem. Załóżmy, że rozważamy powiaty w Polsce (jest ich łącznie 380) pod względem warunków dla rozwoju turystyki. Nietrudno zauważyć, że w zbiorze tym są powiaty mające wybitne warunki dla rozwoju turystyki oraz takie, w których turystyka raczej nie zaistnieje. Ponadto, te pierwsze również wykazują duże odmienności, gdyż przykładowo warunki dla turystyki nadmorskiej znacząco różnią się od warunków dla turystyki górskiej. Między wskazanymi skrajnościami lokuje się duża liczba powiatów o średnim poziomie warunków dla rozwoju turystyki. Logika podpowiada, że jeżeli chcielibyśmy poszukiwać jakichś związków korelacyjnych w zjawiskach związanych z rozwojem turystyki w zbiorze polskich powiatów, to najpierw należałoby je pogrupować w podzbiory o podobnych warunkach, a następnie dopiero związków tych poszukiwać w podzbiorach powiatów względnie do siebie podobnych. Do takiego podziału (grupowania) służą metody taksonomiczne.

Zauważmy, że z zaprezentowanej istoty metod taksonomicznych wynika, że w zadaniu, do rozwiązania którego służą, wyróżnić można trzy komponenty, bardzo podobne do składowych wskazywanych w omawianych wcześniej metodach syntetycznej oceny jednostek terytorialnych, są to:

1) Terytorialny system społeczno-gospodarczy (TSSG), z którego wynikają jednostki terytorialne podlegające klasyfikacji (grupowaniu); w naszym przykładzie Polska.
2) Jednostki tworzące TSSG, podlegające grupowaniu; w naszym przykładzie powiaty.
3) Kryterium badania podobieństwa; w naszym przykładzie warunki dla rozwoju turystyki.

Bardzo ważna różnica w stosunku do poprzedniej grupy metod dotyczy trzeciej składowej. O ile metody syntetycznej oceny z definicji służą do oceniania, metody taksonomiczne służą do ustalania podobieństwa. Są to dwa ważne słowa kluczowe odróżniające istotę tych metod.

Spośród wielu metod taksonomicznych do dalszego omówienia zdecydowano wybrać cztery następujące:

- **Dendryt wrocławski**
- **Diagram Czekanowskiego**
- **Metody aglomeracyjne**
- **Metoda k-średnich**.

Wybór tych metod nie jest przypadkowy. Podyktowany jest po pierwsze, częstością ich zastosowań w praktyce (dendryt wrocławski metody aglomeracyjne oraz metoda k-średnich); a po drugie, wartościami poznawczymi problemu grupowania podobnych do siebie obiektów (metoda Czekanowskiego).

Poniżej (tabela 27) prezentowana jest procedura/algorytm postępowania związany z klasyfikacją zbioru jednostek terytorialnych (ich grupowaniem) w oparciu o metody taksonomiczne. Łatwo zauważyć, że punkty 1-4 tego algorytmu są sformułowane identycznie jak w przypadku metod syntetycznej oceny. Różnice zaczynają się dopiero od punktu 5. Zauważana zbieżność nie jest przypadkowa.

Tabela 27
Metody taksonomiczne w zastosowaniu do klasyfikacji jednostek terytorialnych – główne etapy procedury

1. **Sformułowanie problemu (określenie kryterium klasyfikacji):**
    a) Terytorialny system społeczno-gospodarczy [TSSG] (np. kraj, region)
    b) Jednostki tworzące ten system
    c) Kryterium oceny (zjawisko podlegające ocenie w ramach TSSG).
2. **Dyskusja cech – mierników szczegółowych grupowania.**
3. **Wybór cech diagnostycznych – metody:**
    a) metoda grafu
    b) metoda dendrytowa.
4. **Standaryzacja cech diagnostycznych; metody:**
    a) *zero-jedynkowa*
    b) uproszczona
    c) *min-max*.
5. **Zdefiniowanie kryteriów porządkowania obiektów.**
6. **Porządkowanie obiektów w oparciu o przyjęte kryterium.**
7. **Interpretacja wyników.**

Źródło: opracowanie własne.

Obydwie grupy metod (oceny i taksonomiczne) są metodami klasyfikacji, co było już wcześniej wyjaśniane. Różnią się zasadniczo między sobą, co do celu finalnego (jedna grupa służy ocenianiu, druga ustalaniu podobieństwa), ale wspólnym ich wyróżnikiem jest badanie sytuowania się danej jednostki terytorialnej względem jednostek pozostałych. W obydwóch przypadkach mamy zbiór jednostek terytorialnych opisywanych zespołem cech szczegółowych pod względem zadanego kryterium (dotyczącym oceniania lub badania podobieństwa).

Pomimo, że brzmienie punktów 1-4 jest w pełni zbieżne z algorytmem metod wcześniej omawianych, to jednak wskazać należy na pewne ważne wyjątki, odróżniające jedną grupę od drugiej.

Po pierwsze, inny jest problem, do rozwiązania którego służą metody (punkt 1). Było to już wyżej omówione.

Po drugie, wśród wymagań stawianych cechom szczegółowym (punkt 2 algorytmu) w metodach syntetycznej oceny znalazł się następujący wymóg: *Cechy przydatne do oceny (stymulanty, destymulanty)*. W metodach tych nie miało bowiem sensu uwzględnianie cech, na podstawie których nie można było jednoznacznie wskazać, czy więcej to lepiej, czy też

gorzej w świetle przyjętego kryterium. Cechy takie nazwaliśmy umownie nominantami. W przypadku metod taksonomicznych nominanty są również uwzględniane, gdyż one w pełni nadają się do ustalania podobieństwa branych pod uwagę jednostek terytorialnych. Wymóg ten (tylko stymulanty lub destymulanty) adresowany do cech szczegółowych zostaje zatem wykreślony w procedowaniu metod taksonomicznych.

Po trzecie, w metodach syntetycznej oceny wybór cech diagnostycznych (punkt 3) oparto na kryterium: *nie mogą lub nie powinny być ze sobą wysoko skorelowane*. W przypadku metod taksonomicznych obowiązuje tylko kryterium mocne: *nie mogą być ze sobą wysoko skorelowane*, co zostanie wyjaśnione w dalszej części opracowania.

Po czwarte, w zaprezentowanym algorytmie powtórzono wszystkie trzy sposoby standaryzacji cech szczegółowych (punkt 4). Prawdą bowiem jest, że w metodach tych, podobnie jak w grupie poprzedniej, standaryzację oprzeć można na dowolnych formułach byle tylko doprowadzały do porównywalności w wartościach cech oraz zachowywały pierwotne (czyli sprzed standaryzacji) proporcje między analizowanymi jednostkami terytorialnymi. Trzeba jednak zauważyć, że w praktycznych zastosowaniach metod taksonomicznych zdecydowanie najczęściej stosowana jest metoda *zero-jedynkowa*.

Trzy pierwsze metody (dendryt wrocławski, diagram Czekanowskiego oraz metody aglomeracyjne) mają dodatkowy etap wspólny. Ilustruje to zestawienie przedstawione w tabeli 28. Tym wspólnym etapem jest *macierz odległości taksonomicznych*. Od tego etapu rozpoczynać będziemy omawianie kolejno trzech wymienionych metod.

Tabela 28
Metody taksonomiczne w zastosowaniu do klasyfikacji jednostek terytorialnych – główne etapy procedury dendrytu wrocławskiego, diagramu Czekanowskiego oraz metod aglomeracyjnych

1. **Sformułowanie problemu (określenie kryterium klasyfikacji):**
   a) Terytorialny system społeczno-gospodarczy [TSSG] (np. kraj, region)
   b) Jednostki tworzące ten system
   c) Kryterium oceny (zjawisko podlegające ocenie w ramach TSSG).
2. **Dyskusja cech – mierników szczegółowych grupowania.**
3. **Wybór cech diagnostycznych – metody:**
   a) metoda grafu
   b) metoda dendrytowa.
4. **Standaryzacja cech diagnostycznych; metody:**
   a) *zero-jedynkowa*
   b) uproszczona
   c) *min-max*.
5. **Wyznaczanie macierzy odległości taksonomicznych.**
6. **Porządkowanie obiektów na podstawie macierzy odległości taksonomicznych.**
7. **Interpretacja wyników.**

Źródło: opracowanie własne.

Przejdźmy zatem do punktu 5 algorytmu (tabela 28): ***wyznaczanie macierzy odległości taksonomicznych***. Rozróżnimy dwie interpretacje *odległości taksonomicznych*. Jedną nazwiemy interpretacją merytoryczną i oznacza ona, że odległość taksonomiczna mierzy stopień podobieństwa między dwoma obiektami (jednostkami terytorialnymi) „i" oraz „k" pod względem zadanego kryterium, np. osiągniętego poziomu rozwoju, warunków dla rozwoju turystyki itd. (i=1, 2, …m[59]; oraz k=1,2, …m). Drugą jest interpretacja geometryczna. Oznacza ona odległość punktów w przestrzeni n-wymiarowej[60]. Współrzędnymi tych punktów są wartości cech (po standaryzacji) przypisane i-temu oraz k-temu obiektowi.

Ważną kwestią nasuwającą się z powyżej zarysowanych interpretacji odległości taksonomicznych jest pomiar tych odległości, czyli zaproponowanie formuły matematycznej mierzącej podobieństwo i-tej i k-tej jednostki terytorialnej. W literaturze spotkać można wiele propozycji w tym względzie. Nas interesować będą metryki, które spełniają następujące warunki[61]:

Przyjmijmy założenie, że: a, b, c – obiekty należące do zbioru Ω

$$\bigwedge_{a,b} a = b \leftrightarrow d(a,b) = 0 \qquad (15)$$

Odległość taksonomiczna[62] dwóch identycznych obiektów wynosi 0.

$$\bigwedge_{a,b} d(a,a) = d(b,b) \leq d(a,b) \qquad (16)$$

Odległość taksonomiczna danego obiektu do siebie wynosi tyle, co odległość każdego innego obiektu do siebie i jednocześnie odległość ta jest mniejsza lub co najwyżej równa odległości tego obiektu od każdego innego.

$$\bigwedge_{a,b} d(a,b) = d(b,a) \qquad (17)$$

warunek symetrii

Odległość taksonomiczna obiektu „a" od obiektu „b" jest taka sama jak obiektu „b" od obiektu „a".

$$\bigwedge_{a,b,c} d(a,c) \leq d(a,b) + d(b,c) \qquad (18)$$

warunek trójkąta

Odległość taksonomiczna obiektu „a" od obiektu „c" jest mniejsza lub co najwyżej równa sumie odległości obiektu „a" od obiektu „b" oraz odległości obiektu „b" od obiektu „c".

Sygnalizowane powyżej wymogi dotyczące ustalania metryki odległości taksonomicznej (miary podobieństwa) spełnia wiele formuł. Najczęściej stosowaną w praktyce jest formuła zaczerpnięta wprost z geometrii analitycznej, mierząca odległość punktów w n-wymiarowej przestrzeni euklidesowej:

$$d_{ik} = \sqrt{\sum_{j=1}^{n} (t_{ij} - t_{kj})^2} \qquad (19)$$

[59] Pamiętamy, że przez „m" oznaczyliśmy liczbę obiektów (jednostek terytorialnych).
[60] Pamiętamy, że przez „n" oznaczyliśmy liczbę cech opisujących kryterium klasyfikacji.
[61] Przez „d" oznaczono odległość.
[62] Zamiast określenia „odległość taksonomiczna" możemy operować terminem „podobieństwo".

gdzie:

d_{ik} – odległości punktu "i" od punktu "k"; d_{ik} nazywać będziemy też wskaźnikiem podobieństwa,

inaczej: podobieństwo i-tej jednostki terytorialnej do k-tej jednostki terytorialnej

t_{ij}, t_{kj} – standaryzowane wartości cech.

Wzór powyższy (19) przyjmuje niekiedy postać nieco zmodyfikowaną:

$$d_{ik} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(t_{ij} - t_{kj})^2} \qquad (20)$$

Warto jeszcze wspomnieć o tzw. *odległości miejskiej*, która również często w praktyce jest stosowana:

$$d_{ik}=\sum_{j=1}^{n}|t_{ij} - t_{kj}| \qquad (21)$$

W dalszych rozważaniach posługiwać się będziemy wzorem na odległość, znanym z klasycznej geometrii (występującym również pod nazwą: wzór na odległość euklidesową).

Zauważmy, że jeżeli zbiór obiektów (jednostek terytorialnych) liczy „m", to łącznie wyznaczyć możemy m * m wskaźników podobieństwa d_{ik}, tworzących macierz odległości postaci:

$$\mathbf{d} = \begin{bmatrix} d_{11} & d_{12} & d_{13} & .... & d_{1m} \\ d_{21} & d_{22} & d_{23} & .... & d_{2m} \\ d_{31} & d_{32} & d_{33} & .... & d_{3m} \\ ..... & & & & \\ d_{m1} & d_{m2} & d_{m3} & .... & d_{mm} \end{bmatrix} \qquad (22)$$

Uwzględniając jednak relację symetrii (17), a także zerową wartość wskaźnika podobieństwa danej jednostki z nią samą – patrz relacja (15) – liczba różnych do wyliczenia wskaźników wynosi:

$$\frac{m*(m-1)}{2} \qquad (23)$$

Macierz (22) ma na głównej przekątnej zera oraz $d_{ik} = d_{ki}$ dla i, k = 1, 2, …m; czyli jest postaci:

$$\mathbf{d} = \begin{bmatrix} 0 & d_{12} & d_{13} & .... & d_{1m} \\ d_{21} & 0 & d_{23} & .... & d_{2m} \\ d_{31} & d_{32} & 0 & .... & d_{3m} \\ ..... & & & & \\ d_{m1} & d_{m2} & d_{m3} & .... & 0 \end{bmatrix} \qquad (24)$$

Należy zauważyć, że macierz ta ujmuje wszystkie możliwe pary podobieństw między m-licznym zbiorem analizowanych obiektów (jednostek terytorialnych). Jest to więc pełna informacja w „przedmiotowej sprawie". Choć ujmuje wszystkie możliwe informacje dotyczące podobieństwa, wprost nie nadaje się do wydzielania na jej podstawie podzbiorów obiektów podobnych do siebie pod względem zadanego kryterium. Wynika to chociażby z dużej na ogół liczby informacji ujmowanej przez macierz. Jeżeli, przykładowo, przedmiotem analizy (grupowania) byłyby powiaty w Polsce, wówczas wymiar tej macierzy wynosiłby 380x380, czyli różnych w wartościach wskaźników podobieństwa byłoby: (380x379)/2 = 72010.

Chcąc ustalić grupy jednostek podobnych do siebie, trzeba użyć odpowiedniej procedury uogólniania informacji zawartych w omawianej macierzy. Zagadnienie to będzie przedmiotem dalszej części podręcznika.

Pytania/zadania kontrolne

1. Jakie są główne różnice w klasyfikacji jednostek terytorialnych, będące wynikiem zastosowania wskaźników oceny syntetycznej i opartej na podejściu wielowymiarowym?
2. Proszę o przemyślenie prawdziwości stwierdzenia: wyniki badania podobieństwa obiektów pod względem określonego kryterium nie mogą być bezpośrednią podstawą dokonywania ocen tych obiektów.
3. Co to znaczy, że macierz odległości taksonomicznych dostarcza pełnej informacji o podobieństwie badanych obiektów (jednostek terytorialnych)?
4. Czy w świetle poznanych zasad badania podobieństwa określonego zbioru obiektów, np. miast, wynika, że jeżeli dowolne miasto A jest najbardziej podobne do miasta B, to wówczas zawsze miasto B musi być najbardziej podobne do miasta A?
5. Czy w świetle poznanych zasad badania podobieństwa określonego zbioru obiektów, np. miast, wynika, że jeżeli dowolne miasto A jest najbardziej podobne do miasta B, zaś miasto B jest najbardziej podobne do miasta C, to wówczas zawsze również miasto A musi być najbardziej podobne do miasta C?

Przyjmijmy założenie, że badanie podobieństwa dotyczy „m" obiektów pod względem kryterium charakteryzowanym „n" cechami. Wskaźniki podobieństwa ustalone zostały według formuły odległości euklidesowej (wzór 20). Jakie maksymalne wartości przyjmować może macierz odległości taksonomicznych w przypadku, gdy cechy standaryzowane były formułą *min-max*?; czy możliwe jest przypuszczenie, jak te wartości kształtować się mogą w przypadku standaryzacji *zero-jedynkowej*?

3.2. Metody taksonomiczne – dendryt wrocławski

W poprzednim podrozdziale omówione zostały punkty 1-5 algorytmu postępowania związanego z procedurą zastosowania metod taksonomicznych do grupowania podobnych do siebie jednostek terytorialnych. Warto zwrócić uwagę na ostatni z omawianych punktów algorytmu, a więc punkt 5, który dotyczy macierzy odległości taksonomicznych postaci (24). Przedstawiona tam konkluzja wskazuje, że choć macierz ta ujmuje wszystkie możliwe informacja o podobieństwie badanym obiektów, wprost nie nadaje się do wydzielania na jej podstawie podzbiorów obiektów podobnych do siebie, m.in. ze względu na ogrom zawartych w niej informacji. Konieczne jest więc użycie odpowiedniej procedury uogólniania zawartych tam informacji. Procedurom tym poświęcona jest dalsza część niniejszego podręcznika (podrozdziały 3.2-3.4).

Jak to zostało sformułowane powyżej, z uwagi na ogrom informacji zawartych w macierzy odległości taksonomicznych konieczne jest użycie procedury odpowiedniego ich uogólniania. Literatura proponuje cały szereg metod taksonomicznych, w oparciu o które możliwe jest grupowanie obiektów w podzbiory cechujące się podobieństwem pod względem określonego kryterium. W nawiązaniu do powyżej zaprezentowanego algorytmu, wspólną cechą większości metod taksonomicznych jest zbieżność ich procedury wyrażanej w punktach 1-5[63], czyli do macierzy odległości taksonomicznych włącznie[64]. Od tego dopiero punktu rozpoczynają się między nimi rozbieżności. Nieco metaforycznie ujmując, można więc powiedzieć, że każda z metod wykorzystujących macierzy odległości taksonomicznych podpowiada swoisty dla siebie sposób uogólniania zawartych w niej informacji. Należy koniecznie zauważyć, że procesowi uogólniania zawsze towarzyszy upraszczanie, czyli rezygnacja z jakiejś części pełnych informacji zawartych w macierzy odległości. Będzie to eksponowane przy okazji omawiania każdej z trzech pierwszych, poniżej omawianych metod (oprócz niniejszego, podrozdziały 3.3 oraz 3.4).

Dendryt wrocławski

W nazwie metody pojawia się termin już nam znany („dendryt"). Dendryt wykorzystywany był w poprzednio omawianych metodach do doboru cech diagnostycznych spełniających warunek, że nie są ze sobą wysoko skorelowane. Ogólna idea metody pozostaje taka sama. Z tym, że dendryt wrocławski budowany jest na podstawie macierzy odległości taksonomicznych. Do wyboru cech diagnostycznych dendryt budowany był natomiast na podstawie macierzy korelacji cech. W dendrycie wrocławskim, jako metodzie taksonomicznej, wierzchołkami (węzłami) dendrytu są obiekty (jednostki terytorialne), zaś wiązadłami (łukami) wskaźniki podobieństwa. Budowa dendrytu wrocławskiego przebiega według następującego algorytmu:

[63] W odniesieniu do metod omawianych w niniejszym podręczniku dotyczy to: dendrytu wrocławskiego, diagramu Czekanowskiego oraz metod aglomeracyjnych.

[64] Nie znaczy to oczywiście, że każdy punkt algorytmu jest identycznie rozstrzygany przez każdą z metod taksonomicznych. Z naszych opisów wynika, że przykładowo sposoby standaryzacji czy również sposoby wyznaczania odległości taksonomicznych mogą opierać się na różnych formułach. Wspominaną zbieżność procedury należy rozumieć jako jedynie pojawianie się w kolejnych jej etapach (1-5) tych samych zagadnień wymagających rozstrzygnięcia.

1) Wyznaczenie macierzy odległości taksonomicznych.
2) Ustalenie najmniejszych wartości wskaźników podobieństwa (odległości taksonomicznych) w ramach poszczególnych obiektów.
3) Interpretując każdy obiekt jako wierzchołek dendrytu, zaś wskaźnik podobieństwa jako jego wiązadło (łuk), połączyć wierzchołki na podstawie najmniejszych wartości wskaźników podobieństwa.
4) W przypadku, gdy otrzymany dendryt nie jest spójny należy postępowanie powtórzyć w ten sposób, by dla danego, izolowanego fragmentu dendrytu wyszukać jego połączenie łukiem z innym fragmentem na zasadzie najmniejszych wartości wskaźników podobieństwa.
5) Ustalamy wartość krytyczną wskaźników podobieństwa (d_k).
6) Wszystkie wiązadła o wartości równej lub wyższej od krytycznej zostają usunięte.
7) Powstałe w ten sposób części dendrytu reprezentują względnie jednorodne grupy obiektów.

Pierwszy punkt powyższego algorytmu (macierzy odległości taksonomicznych) został już wcześniej omówiony.

Punkt 2 informuje, że dla każdego obiektu należy znaleźć inny, najbardziej do niego podobny. Czyli w każdym wierszu lub kolumnie macierzy (jest to macierz symetryczna) szukamy wskaźnika podobieństwa o wartości najmniejszej[65] (ignorujemy oczywiście zera na głównej przekątnej). Następne z kolei zadanie (punkt 3) polega na tym, że kółeczko z odpowiednim numerem (wierzchołek) reprezentujące daną jednostkę terytorialną łączymy łukiem z wierzchołkiem reprezentującym jednostkę najbardziej do niej podobną. Przechodząc wszystkie kolejne wiersze (kolumny) i budując połączenia między odpowiadającymi im wierzchołkami, w finale otrzymujemy dendryt, najczęściej niespójny. Istnieje wtedy konieczność jego uspójnienia (punkt 4). Postępowanie w tym względzie jest analogiczne jak w niespójnym dendrycie budowanym dla wyboru cech diagnostycznych, z jednym wyjątkiem, bardzo ważnym i wartym zapamiętania. Wyjątkiem tym jest to, że poszukujemy połączenia danego fragmentu[66] dendrytu, z którymś innym na zasadzie największego podobieństwa, czyli szukając najmniejszej wartości odległości taksonomicznej. W poprzednim wykorzystaniu metody dendrytu, szukaliśmy wartości największej (największego skorelowania).

Osobnego i nieco szerszego omówienia wymaga punkt 5 algorytmu: *ustalamy wartość krytyczną wskaźników podobieństwa*. Wartość krytyczna wskaźnika podobieństwa przesądza o tym, do jakiego stopnia różnic jesteśmy skłonni twierdzić, że dane jednostki są do siebie podobne. Jak pamiętamy, istotą omawianych metod jest podział niejednorodnego zbioru obiektów w podzbiory obiektów do siebie podobnych, pod względem zadanego kryterium. Ustalona wartość krytyczna podobieństwa decyduje zatem o tym, czy rozważany obiekt (jednostka terytorialna) będzie mógł należeć do danego ich podzbioru, czy też nie będzie mógł należeć. W dendrycie służącym do wyboru cech diagnostycznych wartość krytyczna[67] ustalana była w oparciu o dwa

[65] Warto przypomnieć, że budując dendryt służący wyborowi cech diagnostycznych, w macierzy korelacji szukaliśmy wartości największej (największego skorelowania danej cech z jakąś inną, spośród rozważanych).
[66] Przypominamy, że dla zmniejszenia pracochłonności uspójnianie warto rozpocząć od elementu najmniej licznego w wierzchołki.
[67] Wartość krytyczna dotyczyła wtedy współczynnika korelacji.

możliwe podejścia: z góry przyjęta wartość; wartość ustalana z wykorzystaniem procedury badania istotności współczynnika korelacji. W przypadku wartości krytycznej dotyczącej podobieństwa grupowanych obiektów, do dyspozycji jest tylko jeden sposób, a mianowicie, wartość z góry zakładana. Czyli inaczej, wartość, na którą będzie musiał zdecydować się badacz dążący do podziału niejednorodnego zbioru obiektów w podzbiory względnie do siebie podobne. W pierwszym odczuciu może nasuwać się obawa o zbyt duży ładunek subiektywizmu, skutkujący dużą jednocześnie dowolnością przedmiotowego grupowania. Nie do końca jednak „dowolność" ta ma miejsce, o czym będziemy poniżej przekonywać.

Kolejnym krokiem jest usunięcie wiązadeł o wartościach równych lub powyżej przyjętej wartości krytycznej (punkt 5). W ten sposób dendryt uspójniony w poprzednim etapie zostaje podzielony na odpowiednie części[68], które jednocześnie wyznaczają liczbę oraz skład podzbiorów analizowanych obiektów (etap finalny 7).

W dalszej części niniejszego podrozdziału prześledzimy konkretny przykład, który pozwolić powinien na ugruntowanie przedstawionej wyżej wiedzy, a także na stworzenie okazji do uzasadnienia, że z góry przyjmowana wartość krytyczna wskaźnika podobieństwa nie oznacza – wspominanej wcześniej – dowolności dokonywanego wydzielania względnie jednorodnych obiektów.

Posłużymy się tym samym przykładem, który prezentowany był w rozważaniach dotyczących metod syntetycznej oceny jednostek terytorialnych. Rozwiązywany wówczas problem był następujący: dokonać oceny ogólnego poziomu rozwoju województw. Zgromadzonych zostało 20 cech opisujących kryterium oceny (poziom rozwoju), z których – po analizie korelacji – do dalszego postępowania wybranych zostało pięć z nich. W obecnej metodzie celem jest podział niejednorodnego zbioru województw pod względem struktury osiągniętego poziomu rozwoju w podzbiory bardziej jednorodne.

Ponieważ cztery pierwsze etapy procedury metod taksonomicznych są zbieżne z etapami metod syntetycznej oceny, skorzystamy z wyników uzyskanych przy okazji analizowania przykładu w poprzedniej grupie metod. Jak wspominamy powyżej, w wyniku doboru cech diagnostycznych ustalony został zbiór pięciu cech[69], które to w przypadku omawianej metody taksonomicznej będą podstawą wyznaczenia – po uprzedniej ich standaryzacji – macierzy odległości taksonomicznych. Macierz tę prezentuje tabela 29[70].

Na podstawie wskaźników podobieństwa ujętych w macierzy odległości taksonomicznych (tabela 29), stosując się do wskazań punktu 2 i 3 algorytmu metody dendrytu wrocławskiego, otrzymujemy dendryt postaci prezentowanej rysunkiem 15.

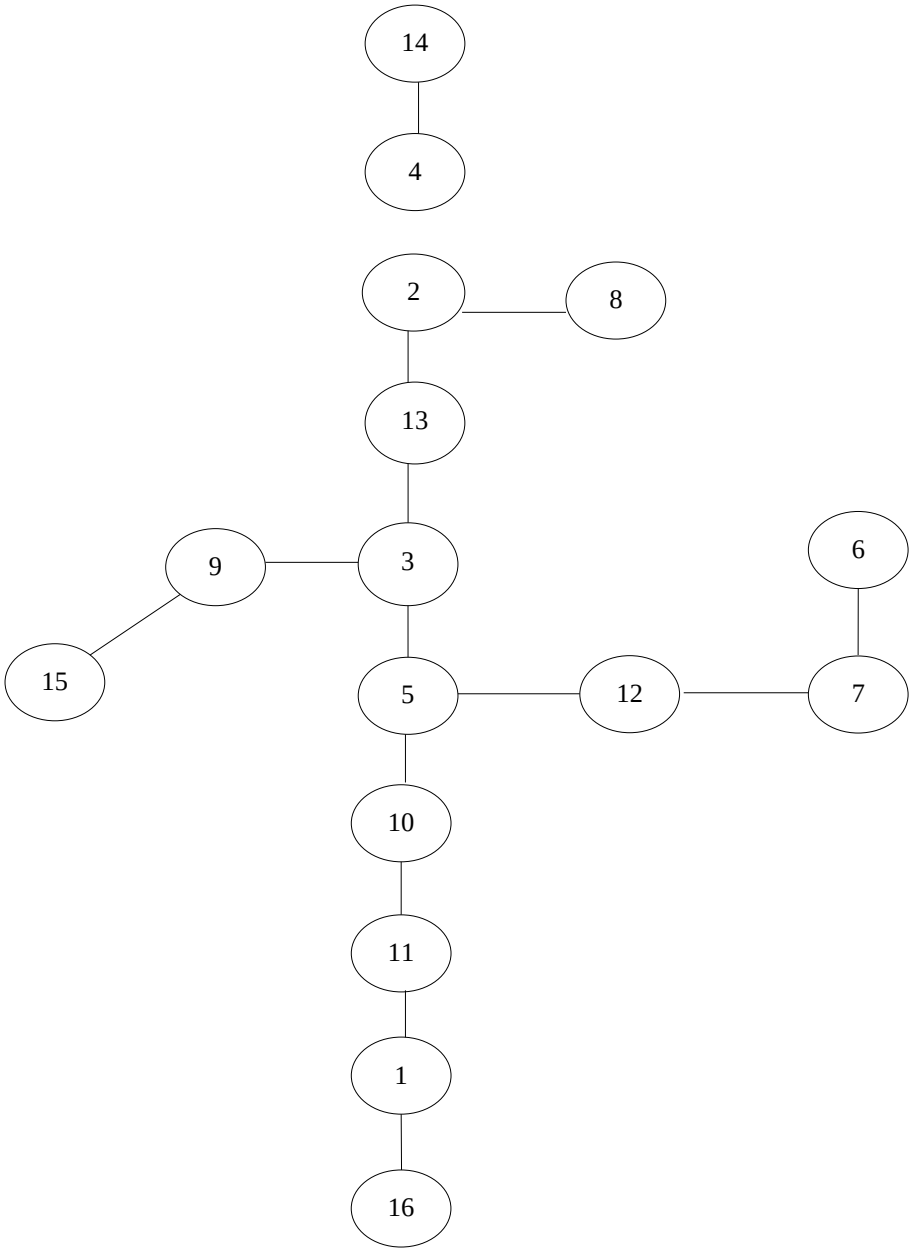
---

[68] Należy koniecznie zauważyć, że te dwa etapy: najpierw uspójnianie, a następnie pewnego rodzaju rozspójnianie, opierają się na zupełnie odmiennych podstawach. Usunięcie wiązadeł nie prowadzi bowiem do postaci dendrytu sprzed uspójniania.

[69] Spośród 20 cech charakteryzujących przyjęte kryterium wybranych zostało pięć następujących, spełniających warunek: nie są ze sobą wysoko skorelowane:

1) Stopa bezrobocia
2) Teatry i instytucje muzyczne
3) Plony zbóż
4) Turyści zagraniczni korzystający z noclegów na 1 000 ludności
5) Odsetek gospodarstw domowych posiadających więcej niż jeden samochód osobowy.

[70] Wskaźniki podobieństwa, które ujmuje prezentowana macierz wyznaczone zostały za pomocą wzoru przedstawianego w podrozdziale 3.1, dotyczącym istoty omawianych metod.



Rysunek 15. Dendryt wrocławski niespójny.
Źródło: opracowanie własne.

Tabela 29
Macierz odległości taksonomicznych

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0,000	1,355	1,719	1,329	1,610	1,540	1,851	1,468	1,987	1,632	0,917	1,282	1,823	1,670	1,978	1,127
2	1,355	0,000	0,713	0,655	0,768	1,801	1,850	1,365	0,915	0,947	0,906	1,190	0,542	0,913	1,510	1,531
3	1,719	0,713	0,000	0,958	0,557	1,577	1,649	1,447	0,389	0,766	1,135	1,087	0,480	1,362	0,967	1,956
4	1,329	0,655	0,958	0,000	1,136	1,578	1,949	1,467	1,083	1,136	0,986	1,482	0,896	0,649	1,562	1,171
5	1,610	0,768	0,557	1,136	0,000	1,511	1,285	1,747	0,900	0,650	0,864	0,805	0,683	1,390	1,271	1,849
6	1,540	1,801	1,577	1,578	1,511	0,000	1,086	2,076	1,794	1,553	1,261	1,262	1,912	1,999	1,405	1,699
7	1,851	1,850	1,649	1,949	1,285	1,086	0,000	2,453	1,909	1,622	1,344	0,968	1,916	2,181	1,654	2,097
8	1,468	1,365	1,447	1,467	1,747	2,076	2,453	0,000	1,461	1,836	1,728	1,596	1,638	2,012	1,420	2,191
9	1,987	0,915	0,389	1,083	0,900	1,794	1,909	1,461	0,000	1,117	1,486	1,379	0,621	1,471	0,927	2,185
10	1,632	0,947	0,766	1,136	0,650	1,553	1,622	1,836	1,117	0,000	0,847	1,190	0,814	1,432	1,511	1,781
11	0,917	0,906	1,135	0,986	0,864	1,261	1,344	1,728	1,486	0,847	0,000	0,917	1,196	1,250	1,680	1,139
12	1,282	1,190	1,087	1,482	0,805	1,262	0,968	1,596	1,379	1,190	0,917	0,000	1,373	1,854	1,214	1,905
13	1,823	0,542	0,480	0,896	0,683	1,912	1,916	1,638	0,621	0,814	1,196	1,373	0,000	1,100	1,429	1,902
14	1,670	0,913	1,362	0,649	1,390	1,999	2,181	2,012	1,471	1,432	1,250	1,854	1,100	0,000	2,098	1,150
15	1,978	1,510	0,967	1,562	1,271	1,405	1,654	1,420	0,927	1,511	1,680	1,214	1,429	2,098	0,000	2,426
16	1,127	1,531	1,956	1,171	1,849	1,699	2,097	2,191	2,185	1,781	1,139	1,905	1,902	1,150	2,426	0,000

Źródło: obliczenia własne.

Zauważmy, że jest to dendryt niespójny, gdyż składa się z dwóch części. Mniejsza z nich obejmuje dwa tylko wierzchołki (4 i 14). Od tej właśnie części rozpoczniemy uspójnianie dendrytu na zasadzie największego podobieństwa (najmniejszej wartości wskaźnika d_{ik}) – patrz punkt 4 algorytmu. Rozpoznanie podobieństwa województw reprezentowanych przez wierzchołki 4 i 14 z wierzchołkami drugiej części dendrytu prowadzi do wniosku, że wierzchołek 4 należy połączyć z wierzchołkiem 2. Uspójniony w ten sposób dendryt ujmuje rysunek 16. Dodatkowo umieszczone zostały obok wiązadeł tego dendrytu wartości wskaźników podobieństwa.

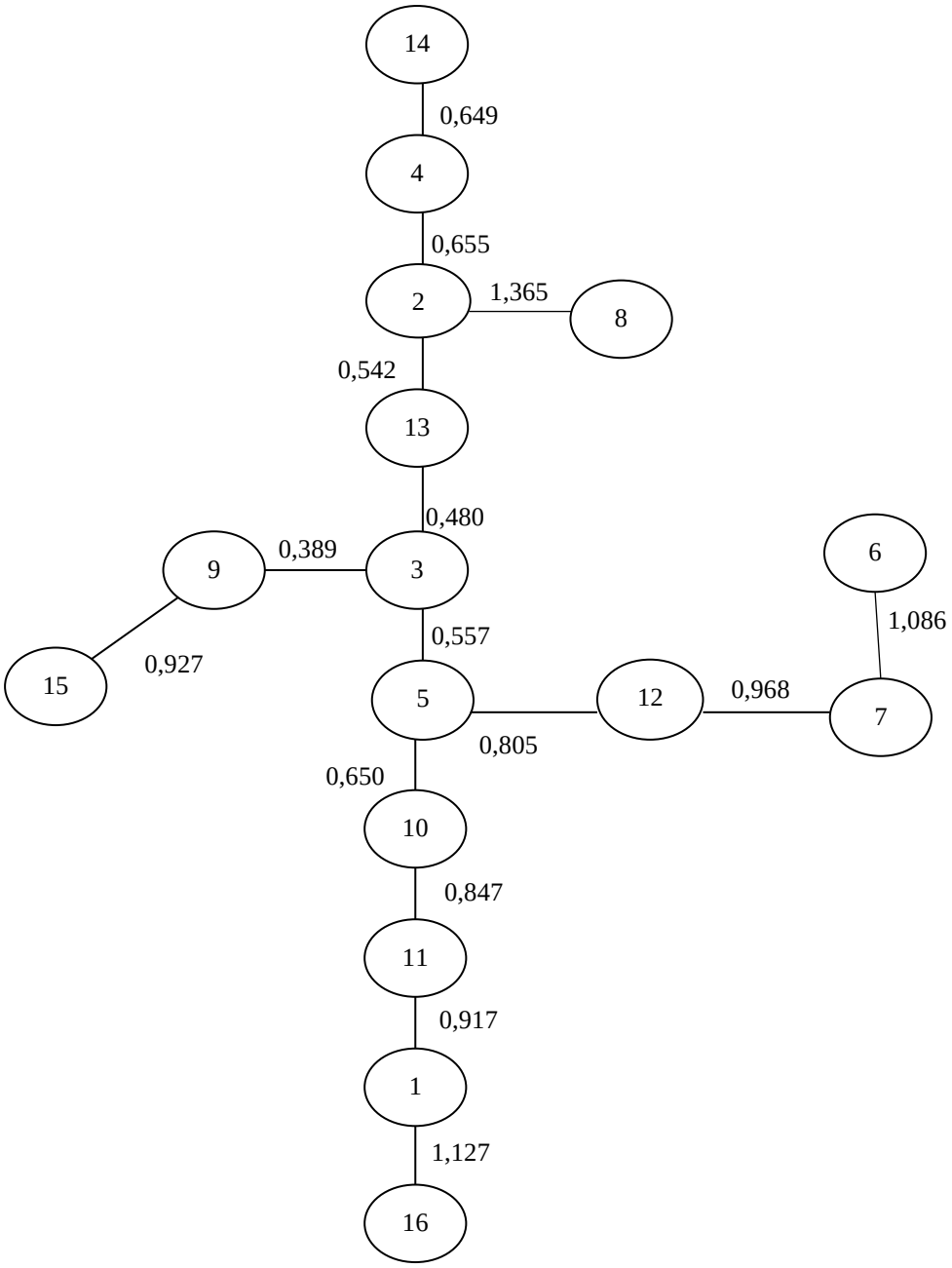
Kolejnym etapem jest ustalenie wartości krytycznej wskaźnika podobieństwa (punkt 5), na podstawie którego wiązadła o wartościach równych lub większych od d_k zostaną usunięte (punkt 6). W efekcie tego powstaną części dendrytu wskazujące jednoznacznie na liczbę i skład wydzielonych grup województw. Zanim zdecydujemy się na konkretną wartość krytyczną wskaźnika podobieństwa przeprowadźmy pewne rozumowanie. Zauważmy, że gdybyśmy dla naszego przykładu przyjęli np. $d_k \leq 0{,}389$[71], wówczas wszystkie wiązadła dendrytu (patrz rysunek 2) zostałyby usunięte, czyli powstałoby 16 jednoelementowych grup. Gdyby natomiast przyjąć $d_k > 1{,}365$[72], wtedy żadne wiązadło nie podlegałoby usunięciu, co oznacza, że powstałaby jedna grupa, czyli nie dokonalibyśmy żadnego podziału. Uogólniając możemy powiedzieć, że w każdym przypadku największa liczba możliwych grup wynosi „m" (jednoelementowe grupy), zaś najmniejsza liczba grup wynosi 1. Oczywiście te skrajne możliwości nie mają żadnego merytorycznego sensu/zastosowania. Praktyczną użyteczność ma liczba grup między 1 a „m". Wracając do naszego przykładu, zwróćmy uwagę, że gdyby przyjąć dowolną wartość d_k np. z przedziału (1,127; 1,365], wówczas otrzymamy dwie grupy obiektów. Zauważmy, że:

- ✓ Po pierwsze, dowolna wartość d_k należąca do tego przedziału prowadzi zawsze do tych samych dwóch grup. Zmieniając wartość d_k w ramach przedziału, nie ma więc „niebezpieczeństwa" wspominanej wcześniej dowolności podziału.

- ✓ Po drugie, jednorodność[73] (podobieństwo) wewnątrz tych grup zawsze będzie na poziomie $d_k < 1{,}365$ (wartość mniejsza od górnej granicy przedziału[74]).

Przyjmując z kolei dowolną wartość d_k z przedziału (1,089; 1,365] otrzymamy trzy grupy obiektów. Podobnie jak powyżej, dowolna wartość d_k prowadzi zawsze do tych samych trzech grup, zaś maksymalna niejednorodność (przeciwieństwo jednorodności) będzie mniejsza od górnej granicy przedziału. Kontynuując rozumowanie, możemy znaleźć wartości krytyczne wskaźnika podobieństwa, prowadzące do podziału naszego zbioru województwa na cztery, pięć itd., aż do 16 grup.

[71] 0,385 jest najmniejszą wartością wskaźnika podobieństwa (wiązadła) w dendrycie prezentowanym na rysunku 16.
[72] 1,365 jest największą wartością wskaźnika podobieństwa (wiązadła) w dendrycie prezentowanym na rysunku 16.
[73] Przez jednorodność grupy obiektów w naszym przypadku rozumieć będziemy maksymalną wartość odległości dwóch dowolnych obiektów należących do tej grupy.
[74] Przy okazji zauważmy, że im mniejsze d_k, tym większa jednorodność elementów należących do poszczególnych grup. Nie jest to oczywiście relacja ciągła, lecz skokowa, gdyż nie każde zmniejszenie d_{ik} skutkuje zwiększeniem jednorodności. Ma to miejsce jedynie wówczas, gdy zmniejszenie wartości krytycznej prowadzi do powstanie większej liczby grup.



Rysunek 16. Dendryt wrocławski spójny.
Źródło: opracowanie własne.

Nietrudno zauważyć, że ważnym walorem omawianej metody taksonomicznej jest pełna kontrola prowadzącego analizy nad ilością wydzielanych grup. Analizując skład otrzymywanych grup, w tym poziom ich jednorodności, analityk musi w końcu zadecydować o ostatecznej wartości krytycznej wskaźnika podobieństwa. W następstwie tego otrzyma akceptowalną przez siebie liczbę grup oraz maksymalny poziom ich niejednorodności[75].

[75] Maksymalny poziom niejednorodności mierzony wskaźnikiem podobieństwa zawsze jest mniejszy od przyjmowanej wartość krytycznej d_k.

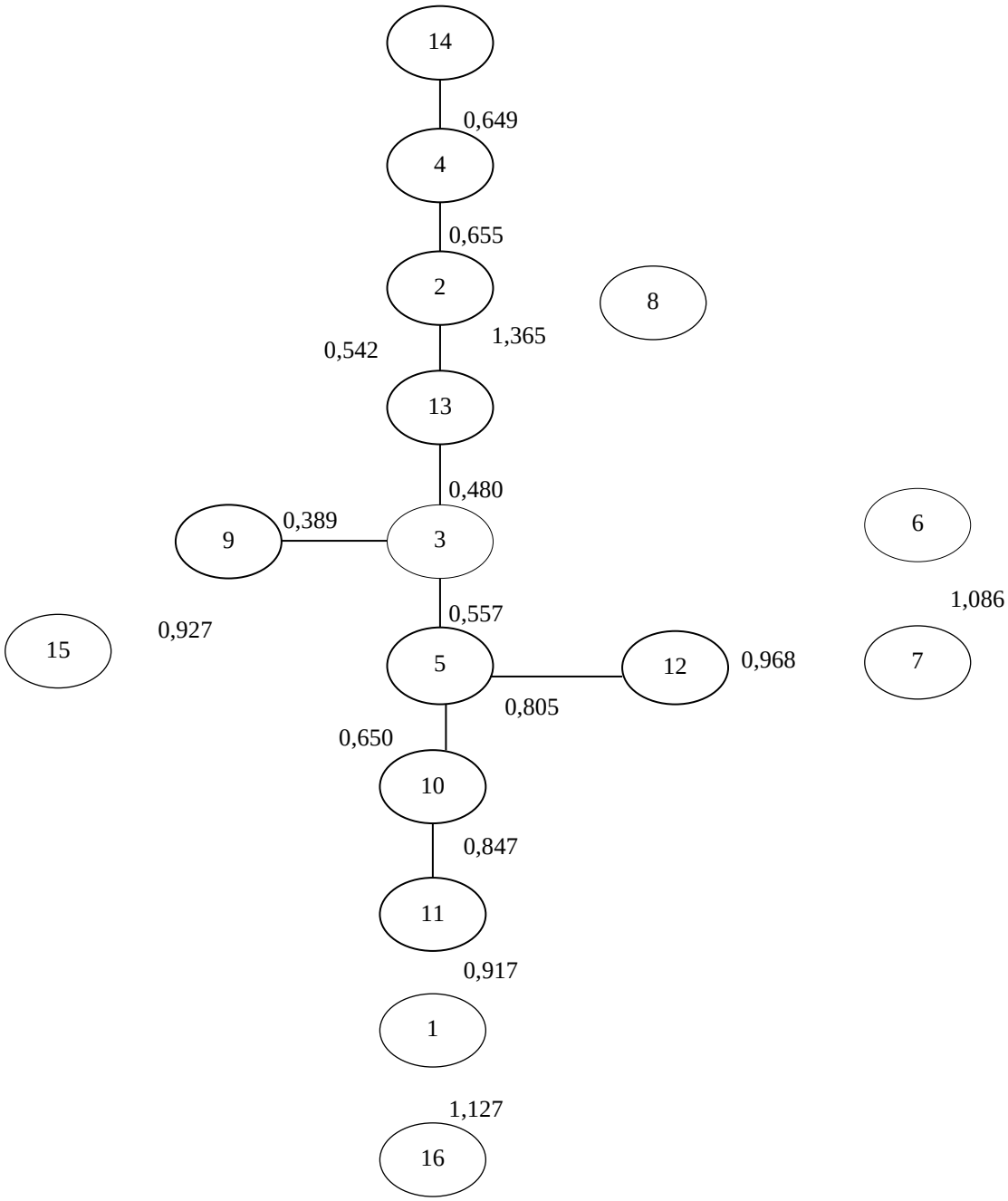
Zdecydujmy, że w naszym przykładzie $d_k = 0{,}9$. Finalny podział na grupy ilustruje rysunek 17.

Otrzymane części dendrytu wskazują na siedem grup, z których sześć obejmuje po jednym tylko województwie. Są to województwa reprezentowane przez wierzchołki o numerach[76]: 1, 6, 7, 8, 15, 16. Jedna natomiast grupa jest relatywnie duża, obejmując 10 województw. Trochę dziwaczny podział (grupowanie) wynika z pewnością z małolicznego zbioru obiektów podlegających grupowaniu. Ale z kolei operowanie w przykładzie dużym zbiorem obiektów mocno skomplikowałoby możliwość prezentacji oraz komentowania wyników w ramach opracowania, obniżając jednocześnie jego przejrzystość.

Na zakończenie wrócimy jeszcze do jednej ważnej, wcześniej już wspominanej, kwestii, a mianowicie uogólniania oraz upraszczania informacji zawartych w macierzy odległości taksonomicznych. Zapoznanie się z procedurą budowy dendrytu wrocławskiego, a następnie z zasadami wydzielania na jego podstawie względnie jednorodnych grup obiektów, pozwala dostrzec nam, jak daleko posunięte jest upraszczanie przez tę metodę informacji zawartych w macierzy odległości taksonomicznych. Macierz ta zawiera wskaźniki podobieństwa danego obiektu (w naszym przykładzie województwa) z <u>każdym</u> innym obiektem (zarówno wiersze, jak i kolumny[77] macierzy składają się z „m" elementów). Natomiast metoda dendrytu wrocławskiego, poszukując dla danego obiektu innego, do którego jest on najbardziej podobny, sięga jedynie do jednej z tych wartości w każdym wierszu (kolumnie), ignorując wszystkie pozostałe. Jedynie na etapie uspójniania, w którymś z wierzy (kolumn) sięga się do dalszych wskaźników podobieństwa. Pomimo, że w praktyce metoda ta jest dosyć często wykorzystywana, z dużą ostrożnością należy zatem podejść do wyników otrzymywanych na jej podstawie. Postuluje się, aby wyniki otrzymywane za pomocą omawianej metody traktować, jako wstępne grupowanie. Wrócimy do tego postulatu w następnym podrozdziale, przy okazji omawiania kolejnej metody taksonomicznej.

[76] Korzystając z alfabetycznego zestawienia województw prezentowanego w jednym z wcześniejszych podrozdziałów (2.2, tabela 3), oddzielne grupy stanowią następujące województwa: dolnośląskie, małopolskie, mazowieckie, opolskie, wielkopolskie, zachodniopomorskie.

[77] Należy przypomnieć, że macierz odległości taksonomicznych jest macierzą symetryczną, co oznacza, że elementy danego jej wiersza są identyczne, jak elementy kolumny o tym samym, co wiersz, numerze.



Rysunek 17. Dendryt wrocławski – podział na grupy względnie jednorodne.
Źródło: opracowanie własne.

Pytania/zadania kontrolne

1. Na czym polega w dendrycie wrocławskim upraszczanie informacji zawartych w macierzy odległości taksonomicznych?
2. W odniesieniu algorytmu dotyczącego dendrytu wrocławskiego, budowanego na macierzy odległości taksonomicznych, dokładnego przemyślenia wymagają jego różnice w stosunku do algorytmu dendrytu budowanego na bazie macierzy korelacji. Pierwszy z nich jest narzędziem podziału niejednorodnego zbioru obiektów w podzbiory bardziej do siebie podobne; drugi natomiast służy wyborowi/selekcji cech diagnostycznych spełniających określony warunek (nie są ze sobą wysoko skorelowane).
3. Ważnym etapem w algorytmie budowy dendrytu wrocławskiego jest „ustalenie wartość krytycznej wskaźników podobieństwa (dk)". Proszę o dokładne przemyślenie prawdziwości stwierdzenia: „arbitralność w ustalaniu wartości krytyczne dk nie skutkuje przypadkowością wydzielania grup podobnych do siebie obiektów".
4. Uzasadnij, że dendryt wrocławski umożliwia pełny przegląd możliwości podziału zbioru „m" na dowolną liczbę ich grup.

3.3. Metody taksonomiczne – diagram Czekanowskiego

W powyższym podrozdziale omówiona została metoda taksonomiczna, dendryt wrocławski. Przypomnijmy, że trzy pierwsze z omawianych metod taksonomicznych – prowadząc do podziału niejednorodnego zbioru obiektów w podzbiory obiektów do siebie podobnych – m.in. przedstawiają procedurę korzystania z informacji ujmowanej przez macierz odległości taksonomicznych (24).

W niniejszym podrozdziale omówiona zostanie kolejna metoda taksonomiczna, pod nazwą metoda Czekanowskiego, nazywana też diagramem Czekanowskiego[78]. Jej algorytm przedstawia się następująco:

1. Wyznaczenie macierzy odległości taksonomicznych.
2. Ustalenie przedziałów liczbowych na skali podobieństwa.
3. Przyporządkowanie przedziałom liczbowym na skali podobieństwa znaków graficznych.
4. Budowa diagramu nieuporządkowanego.
5. Porządkowanie diagramu (*warto wykorzystać wyniki dendrytu wrocławskiego*).
6. Ustalenie grup obiektów na podstawie diagramu uporządkowanego.

[78] Jako ciekawostkę warto podać, że nazwa metody pochodzi od nazwiska jej autora, polskiego naukowca, który po raz pierwszy zaproponował ją do klasyfikacji (typologii) antropologicznej.

Jak już zaznaczano, prezentowanie każdej z trzech pierwszych, kolejno omawianych, metod taksonomicznych rozpoczynamy od etapu *wyznaczenie macierzy odległości taksonomicznych*, ponieważ wszystkie wcześniejsze, włącznie z etapem wyznaczania macierzy, są takie same dla każdej z tych metod.

Etap *macierzy odległości taksonomicznych* jest nam już znany. Przechodzimy zatem do etapu następnego (2). Etap ten w dużym stopniu wyjaśnia istotę metody Czekanowskiego. Ujawnia kierunek uogólniania oraz upraszczania. Zwróćmy uwagę, że macierz odległości taksonomicznych – oprócz właściwości omawianych we wcześniejszych fragmentach – ukazuje pełną przestrzeń różnorodności (podobieństwa)[79] badanych obiektów. Każdemu obiektowi przypisane są wskaźniki podobieństwa z każdym innym, pod względem danego kryterium, czyli dla każdego obiektu przestrzeń różnorodności (podobieństwa) wynosi „m". W całym natomiast zbiorze obiektów wymiar omawianej przestrzeni[80] to: m*(m-1)/2. Bardziej obrazowo ujmując: gdyby nanieść na oś liczbową wszystkie wartości wskaźników, to ich liczba między najmniejszym i największym z nich ilustruje przestrzeń różnorodności. Jak już wiemy, ogrom informacji wyklucza możliwość wnioskowania wprost – bezpośrednio na podstawie macierzy odległości – dotyczącego wydzielania grup podobnych do siebie obiektów.

Informacja ujęta w etapie drugim algorytmu metody Czekanowskiego wychodzi z propozycją uproszczenia (zmniejszenia wymiaru) przestrzeni różnorodności. Zauważmy, że wśród wszystkich wskaźników podobieństwa danego obiektu znaleźć można ich zespoły o wartościach zbliżonych do siebie; niekiedy różniących się którymś miejscem po przecinku. Omawiany etap algorytmu proponuje zatem: zamiast operować konkretnymi, a zarazem bardzo licznymi, wartościami wskaźników użyć należy znacząco mniej licznych przedziałów liczbowych. Wyobraźmy sobie wspominaną wyżej oś liczbową z naniesionymi wartościami wskaźników podobieństwa. Zamiast rozróżniać konkretne wartości (punkty) na tej osi, zaznaczamy przedziały liczbowe, czyli:

- ✓ pierwszy przedział: charakteryzuje obiekty najbardziej podobne do siebie pod względem przyjętego kryterium,
- ✓ drugi przedział: charakteryzuje obiekty nieco mniej podobne do siebie,
- ✓ trzeci przedział: charakteryzuje obiekty o jeszcze mniejszym podobieństwie do siebie,

  ……

- ✓ ostatni przedział: charakteryzuje obiekty najbardziej do siebie niepodobne.

Z wyznaczaniem przedziałów liczbowych wiążą się trzy następujące pytania:

1) **Ile przedziałów należy ustalić?** Odpowiadając na to pytanie, należy przede wszystkim zwrócić uwagę, że nie ma żadnych podstaw teoretycznych pozwalających na wyznaczenie tej liczby (przedziałów). Musimy opierać się na doświadczeniach praktycznych wykorzystania omawianej metody. Nietrudno zauważyć, że z punktu widzenia celu, do którego dążymy (grupowanie obiektów), najważniejszymi przedziałami są pierwszy oraz ostatni. Pierwszy wskazuje na najbardziej podobne, zaś ostatni na najbardziej niepodobne do siebie obiekty. Te dwa przedziały nie wystarczają jednak do tworzenia grup obiektów podobnych. Przedziały pośrednie odgrywają rolę pomocniczą, jednak ważną – jak się później przekonamy – w przeprowadzanym grupowaniu. Jasnym chyba jest spostrzeżenie, że wraz ze wzrostem liczby przedziałów zmniejsza się zakres upraszczania ogółu informacji ujmowanych

[79] Różnorodność jest w pewnym sensie antonimem podobieństwa.
[80] Wzór omawiany był wcześniej.

w macierzy odległości, ale jednocześnie wzrasta trudność w grupowaniu obiektów. Dla satysfakcjonującego rozstrzygnięcia celu, jakim jest grupowanie obiektów w zupełności wystarcza relatywnie nieduża liczba przedziałów. Nie jest ona zależna od liczby obiektów (m). Obserwując zastosowanie omawianej metody taksonomicznej, można powiedzieć, że minimalna liczba przedziałów to trzy, zaś niezwykle rzadko jest ona większa niż siedem. Najczęściej spotykana w praktyce zawiera się między trzy a pięć.

2) **Czy interwały (rozpiętości) wszystkich przedziałów muszą być jednakowe?**

Odpowiedź: nie muszą. Zwróćmy uwagę na wcześniej wspominaną oś liczbową, na którą naniesione zostały wartości wszystkich wskaźników podobieństwa z macierzy odległości taksonomicznych. Otóż, nie mamy prawa zakładać, że punkty te są w miarę równomiernie rozmieszczone na tej osi. Jest wysoce prawdopodobne, że znajdować się będą miejsca ich większej koncentracji i miejsca o dużej ich rzadkości. Jednakowe interwały mogłyby prowadzić do sytuacji, że do niektórych przedziałów należałaby duża liczba punktów, natomiast inne, w skrajnych przypadkach, mogłyby być nawet puste. To właśnie powoduje, że interwały przedziałów na ogół muszą być odpowiednio dobierane, z pominięciem zasady ich równości.

Dla ułatwienia realizacji następnych punktów algorytmu (porządkowanie diagramu) nie jest obojętny rozkład liczby wskaźników należących do poszczególnych przedziałów liczbowych. Postulatem w tym względzie jest, aby do pierwszego i ostatniego przedziału należało więcej niż do przedziałów pośrednich. Jeżeli np. przyjąć cztery przedziały, wówczas do pierwszego oraz ostatniego należeć powinno po ok.[81] 30% wskaźników, zaś do dwóch pośrednich po ok. 20% wskaźników.

3) **Jak ustalić granice przedziałów, które respektować będą sugerowany wyżej rozkład liczby wskaźników podobieństwa należących do poszczególnych przedziałów?**

Przy większych rozmiarach macierzy odległości taksonomicznych (czyli dużej liczbie wskaźników podobieństwa) nie jest to łatwe. Metoda „prób i błędów" ustalania granic przedziałów najczęściej okazuje się niezwykle pracochłonna. Jedną z technik ułatwiających rozstrzygnięcie przedstawianego zadania jest uporządkowanie monotoniczne wszystkich wskaźników, by na tej podstawie w sposób bardzo już łatwy podjąć decyzję dotyczącą wartości granic przedziałów liczbowych. Zilustrowane to zostanie na konkretnym przykładzie przy końcu niniejszego podrozdziału.

Dwa kolejne punkty algorytmu (3 oraz 4) ściśle się ze sobą wiążą, warto więc omówić je łącznie. Wyznaczone przedziały liczbowe stanowią podstawę do tego, aby macierz odległości przekształcić w macierz (tablicę/diagram) znaków. W tym właśnie celu przedziałom liczbowym przypisujemy odpowiednie znaki graficzne, którymi następnie zastępujemy liczby w macierzy odległości, budując diagram nieuporządkowany (punkt 4 algorytmu).

W wyborze znaków graficznych ważną zasadą jest, aby wizualnie mocniej eksponować przedziały o wyższych podobieństwach. Inaczej ujmując, aby dla ułatwienia dalszych etapów, pierwszemu przedziałowi liczbowemu przypisać mocniej rzucający się w oczy znak graficzny niż drugiemu; z kolei drugiemu przedziałowi mocniej rzucający się w oczy niż trzeciemu itd. Należy zauważyć, że tabela (macierz/diagram) znaków składa z krateczek na przecięciu i-tego wiersza i k-tej kolumny. Można przykładowo przyjąć, że przy czterech przedziałach znak przypisany

[81] Podkreślenie ma na celu dodatkowe zwrócenie uwagi, że nie chodzi o dokładne 30% czy 20%. Rzecz w tym, aby było więcej w pierwszym i ostatnim w stosunku do pośrednich przedziałów.

pierwszemu przedziałowi to zakolorowana kratka, przypisany drugiemu to przecinające się po przekątnych linie, trzeciemu to myślnik, zaś czwarta kratka mogłaby być pusta.

Po zbudowaniu diagramu nieuporządkowanego kolejnym etapem jest jego porządkowanie (punkt 5 algorytmu). Porządkowanie diagramu polega na takim przestawianiu kolejności wierszy i kolumn, aby wokół głównej przekątnej skupić jak najwięcej znaków reprezentujących najwyższy stopień podobieństwa; im natomiast dalej od głównej przekątnej, tym więcej znaków dotyczących coraz mniejszych podobieństw. Mocnego wyeksponowania wymaga zasada, aby w przestawianiu kolejności wierszy i kolumn zachowana była symetryczność diagramu, tak jak symetryczna jest macierz odległości taksonomicznych. Oznacza to, że jeżeli zamieniamy miejscami wiersz np. o nr 3 z wierszem o nr 8, to jednocześnie musimy zamienić miejscami kolumnę nr 3 z kolumną nr 8.

Porządkowanie diagramu jest najbardziej mozolnym etapem, wymagającym pewnej orientacji przestrzennej w ocenie, jakie zmiany w kolejności wierszy i kolumn mogą poprawić uporządkowanie diagramu. Najczęściej porządkowanie przeprowadzane jest w kilku iteracjach[82]. Wynik każdej następnej iteracji, po jej analizie, jest poprawiany, tzn. podejmuje się próby jego udoskonalenia, aż do sytuacji, gdy już – w naszej ocenie – nie jest możliwe dalsze udoskonalanie porządku.

Doświadczenie podpowiada, że najtrudniejszy jest punkt startu w porządkowaniu, tzn. pierwsza decyzja dotycząca nieuporządkowanego diagramu. Poza jego symetrycznością trudno doszukać się, ukierunkowanych na poprawę uporządkowania, następstw zmian w kolejności wierszy/kolumn. W każdej natomiast postaci diagramu, w której ujawnia się już w jakimś stopniu żądany porządek, o wiele łatwiej doszukać się możliwości poprawy uporządkowania. Biorąc pod uwagę wspominaną trudność w podjęciu decyzji, co do zmiany kolejności wierszy/kolumn „w punkcie startu", proponowane jest wykorzystanie w tym celu wyników dendrytu wrocławskiego do wstępnego uporządkowania diagramu (patrz zapis punktu 5 algorytmu omawianej metody). W takiej sytuacji, dendryt wrocławski traktuje się jako metodę wstępnego porządkowania i jednocześnie pomocniczą względem metody Czekanowskiego. Sposób wykorzystania wyników dendrytu wrocławskiego do wstępnego porządkowania diagramu zilustrowany zostanie nieco dalej, przy okazji omawiania przykładu.

Warto w tym miejscu zauważyć, że taksonomiczną metodę Czekanowskiego cechuje znacząco mniejsze upraszczanie informacji zawartych w macierzy odległości niż metodę dendrytu wrocławskiego. Uogólnianie w tej metodzie (diagramu) polega na zastąpieniu indywidualnych wskaźników podobieństwa dla każdej pary obiektów przez przedziały liczbowe. Cały czas jednak korzysta się z pełnego spektrum różnic w ich podobieństwie. Rozważane jest podobieństwo danego obiektu z każdym innym obiektem przez wszystkie ustalone przedziały podobieństwa.

Ostatni, finalny, etap polega na wydzieleniu grup obiektów podobnych do siebie pod względem zadanego kryterium (punkt 6 algorytmu). Grupy te obrazowo sugeruje uporządkowany diagram, aczkolwiek najczęściej nie w sposób w pełni jednoznaczny. Konieczna jest decyzja osoby/zespołu przeprowadzającego przedmiotowe postępowanie, co do ostatecznego składu wydzielanych grup. Etap ten ilustrowany będzie na konkretnym przykładzie, prezentowanym poniżej.

[82] Liczba iteracji nie jest z góry znana. Zależy to od przeprowadzającego porządkowanie diagramu. Teoretycznie można sobie wyobrazić, że osoba doświadczona w stosowaniu metody może od razu przewidzieć najbardziej odpowiednią kolejność wierszy (kolumn). Przy diagramie większych rozmiarów jest to jednak bardzo mało prawdopodobne.

W dalszej części podrozdziału rozważymy przykład pozwalający na prześledzenie kolejnych etapów algorytmu metody Czekanowskiego. Przedmiotem rozważań będzie ten sam przykład, co powyżej, dotyczący poziomu rozwoju województw. Ponieważ nie ma już sensu powtarzania kwestii związanych z wyznaczaniem macierzy odległości taksonomicznych, przejdziemy od razu do etapu nr 2 algorytmu, czyli wyznaczenia przedziałów podobieństwa. Punktem wyjścia do rozwiązania tego zadania jest macierz odległości taksonomicznych wyznaczona w poprzednim podrozdziale (patrz tabela 29). Załóżmy, że decydujemy się na cztery przedziały podobieństwa. Pamiętając sformułowany wcześniej postulat: przedział pierwszy oraz ostatni powinny zawierać więcej wskaźników podobieństwa niż przedziały pośrednie, powinniśmy ustalić takie granice, które przy czterech przedziałach zapewnią, że do pierwszego i ostatniego z nich należeć będzie po ok. 30% wskaźników, zaś do drugiego i trzeciego po ok. 20%. Tak, jak było to wcześniej podpowiadane, pomocne będzie uporządkowanie monotoniczne wartości wskaźników podobieństwa. Uporządkowanie takie ujmuje tabela 30. Wyraźnego podkreślenia wymaga fakt, że jest to stablicowane ujęcie wektora wartości uporządkowanych (nie jest to żadna macierz). Nie jest bowiem możliwe przedstawienie długiego na ogół wektora, w formie jednego wiersza lub jednej kolumny na standardowej stronie książki. Zestawienie zaczyna się zatem w pierwszej kolumnie, począwszy od najmniejszej wartości wskaźnika podobieństwa, jest kontynuowane w kolumnie drugiej, a następnie trzeciej itd. Drobnego wyjaśnienia dotyczącego tabeli 30 wymagają dwie kwestie. Po pierwsze, powtarzające się te same liczby wynikają z tego, że macierz odległości jest macierzą symetryczną (poniżej i powyżej głównej przekątnej pojawiają się te same liczby)[83]. Po drugie, pominięto zera na głównej przekątnej, a więc tabela ujmuje 16*15 liczb, a nie 16*16.

Przechodząc do wyznaczania granic przedziałów, zauważmy, że wszystkich wskaźników ujętych w tabeli jest 240 (16*15). Jeżeli przyjąć proponowany wyżej rozkład procentowy liczby wskaźników należących do poszczególnych przedziałów, to do pierwszego należeć powinno[84] 72 (30% z 240), drugiego 48, trzeciego 48 i ostatniego 72. Patrząc na wartości tabeli, dolną granicą pierwszego przedziału jest 0,389, zaś górną 1,134 (72 pozycja). Granice kolejnych przedziałów to: drugi: [1,136; 1,404], trzeci: [1,420; 1,639], czwarty: [1,648; 2,453].

Ułatwiając sobie pracę, znaki graficzne dla kolejnych przedziałów podobieństwa zdecydowano zaczerpnąć z klawiatury komputera[85]. Są nimi:

0,389 - 1,134	#
1,136 - 1,404	+
1,420 – 1,639	-
1,648 – 2,453	

[83] Można było oczywiście ograniczyć się do wartości powyżej lub poniżej głównej przekątne, co prowadziłoby do identycznych ustaleń.

[84] Należy powtórzyć, że nie ma żadnego rygorystycznego wymogu, aby zastosować się dokładnie do wskazanego rozkładu procentowego. W przykładzie przyjęliśmy zasugerowany rozkład procentowy, gdyż nie było przeciwskazań, co do takiego właśnie rozstrzygnięcia (przeciwskazaniem mogłyby być np. powtarzające się wartości dla górnej granicy jednego i dolnej następnego przedziału).

[85] Przekształcając wskaźniki macierzy odległości w diagram znaków skorzystano z prostej formuły dostępnej w arkuszu Excela.

Tabela 30
Uporządkowane monotonicznie odległości taksonomiczne

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,389	0,684	0,897	0,948	1,100	1,172	1,272	1,373	1,460	1,512	1,610	1,699	1,836	1,912	2,077
2	0,389	0,684	0,897	0,948	1,100	1,172	1,272	1,373	1,460	1,512	1,610	1,699	1,836	1,912	2,077
3	0,481	0,713	0,899	0,957	1,118	1,190	1,281	1,377	1,467	1,532	1,621	1,720	1,849	1,916	2,097
4	0,481	0,713	0,899	0,957	1,118	1,190	1,281	1,377	1,467	1,532	1,621	1,720	1,849	1,916	2,097
5	0,541	0,767	0,906	0,968	1,127	1,190	1,286	1,391	1,467	1,541	1,632	1,728	1,849	1,948	2,097
6	0,541	0,767	0,906	0,968	1,127	1,190	1,286	1,391	1,467	1,541	1,632	1,728	1,849	1,948	2,097
7	0,557	0,769	0,912	0,968	1,134	1,196	1,328	1,404	1,471	1,554	1,639	1,746	1,851	1,957	2,180
8	0,557	0,769	0,912	0,968	1,134	1,196	1,328	1,404	1,471	1,554	1,639	1,746	1,851	1,957	2,180
9	0,622	0,805	0,915	0,986	1,136	1,214	1,344	1,420	1,483	1,563	1,648	1,780	1,854	1,977	2,185
10	0,622	0,805	0,915	0,986	1,136	1,214	1,344	1,420	1,483	1,563	1,648	1,780	1,854	1,977	2,185
11	0,648	0,814	0,917	1,082	1,136	1,250	1,355	1,429	1,485	1,576	1,652	1,793	1,903	1,986	2,191
12	0,648	0,814	0,917	1,082	1,136	1,250	1,355	1,429	1,485	1,576	1,652	1,793	1,903	1,986	2,191
13	0,651	0,847	0,917	1,087	1,138	1,261	1,362	1,431	1,509	1,579	1,670	1,800	1,905	1,999	2,426
14	0,651	0,847	0,917	1,087	1,138	1,261	1,362	1,431	1,509	1,579	1,670	1,800	1,905	1,999	2,426
15	0,655	0,863	0,926	1,087	1,149	1,261	1,366	1,447	1,512	1,597	1,679	1,822	1,910	2,012	2,453
16	0,655	0,863	0,926	1,087	1,149	1,261	1,366	1,447	1,512	1,597	1,679	1,822	1,910	2,012	2,453

Źródło: opracowanie własne.

W miejsce liczb w macierzy odległości taksonomicznych wprowadzamy powyższe znaki graficzne, otrzymując nieuporządkowany diagram Czekanowskiego; patrz rysunek 18.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	#	+		+	-	-		-		-	#	+				#
2	+	#	#	#	#			+	#	#	#	+	#	#	-	-
3		#	#	#	#	-		-	#	#	+	#	#	+	#	
4	+	#	#	#	+	-		-	#	+	#	-	#	#	-	+
5	-	#	#	+	#	-	+		#	#	#	#	#	+	+	
6	-		-	-	-	#	#			-	+	+			+	
7					+	#	#			-	+	#				
8	-	+	-	-				#	-			-	-		+	
9		#	#	#	#			-	#	#	-	+	#	-	#	
10	-	#	#	+	#	-	-		#	#	#	+	#	-	-	
11	#	#	+	#	#	+	+		-	#	#	#	+	+		+
12	+	+	#	-	#	+	#	-	+	+	#	#	+		+	
13		#	#	#	#			-	#	#	+	+	#	#	-	
14		#	+	#	+				-	-	+		#	#		+
15		-	#	-	+	+		+	#	-		+	-		#	
16	#	-		+							+			+		#

Rysunek 18. Diagram nieuporządkowany Czekanowskiego.
Źródło: opracowanie własne.

Potwierdza się wcześniej sformułowana uwaga, że diagram nieuporządkowany, zgodnie z nazwą, żadnego porządku nie ujawnia, poza jego symetrycznością. Kolejne wiersze i kolumny prezentują wzajemne podobieństwa województw ułożonych w porządku alfabetycznym. Nawet przy tak niewielkim wymiarze diagramu niezwykle trudno zorientować się, jaka zmiana kolejności wierszy/kolumn może przyczynić się do postępu w uporządkowaniu (czyli skupianiu znaków reprezentujących wyższe podobieństwa jak najbliżej głównej przekątnej). Potraktujmy zatem wyniki dendrytu wrocławskiego jako wstępne uporządkowanie. Korzystamy z dendrytu spójnego obrazowanego rysunkiem 16.

Analizując jego układ, zauważamy, że podpowiada on, iż pierwszy wiersz/kolumna powinna być o numerze 14 lub 16 (początek lub koniec dendrytu). Załóżmy, że rozpoczynamy od górnej części dendrytu. Wówczas następnym numerem wiersza/kolumny bez wątpliwości jest nr 4, a kolejnym 2 lub 8, gdyż pojawiają się rozgałęzienia. Właśnie rozgałęzienia wprowadzają niepewność w kwestii kolejności ułożenia wierszy/kolumn w diagramie. Ponieważ nie ma żadnej podpowiedzi, jak postąpić w takich przypadkach, jedynym rozwiązaniem jest przyjęcie jakiejś (dowolnej) kolejności i dopiero po ułożeniu diagramu zdecydować o jego ulepszeniu. W każdym bądź razie ułożenie kolejności wierszy/kolumn według podpowiedzi dendrytu prowadzi zawsze do zauważalnego porządku, co prawda najczęściej nie w pełni zadowalającego. Na podstawie rysunku 16 zdecydowano, że kolejność wierszy/kolumn w pierwszej iteracji porządkowania będzie następująca: 14; 4; 2; 8; 13; 3; 9; 15; 5; 12; 7; 6; 10; 11; 1; 16. Wynik tego porządku prezentuje rysunek 19.

	14	**4**	**2**	**8**	**13**	**3**	**9**	**15**	**5**	**12**	**7**	**6**	**10**	**11**	**1**	**16**
14	#	#	#		#	+	-		+				-	+		+
4	#	#	#	-	#	#	#	-	+	-		-	+	#	+	+
2	#	#	#	+	#	#	#	-	#	+			#	#	+	-
8		-	+	#	-	-	-	+		-					-	
13	#	#	#	-	#	#	#	-	#	+			#	+		
3	+	#	#	-	#	#	#	#	#	#		-	#	+		
9	-	#	#	-	#	#	#	#	#	+			#	-		
15		-	-	+	-	#	#	#	+	+		+	-			
5	+	+	#		#	#	#	+	#	#	+	-	#	#	-	
12		-	+	-	+	#	+	+	#	#	#	+	+	#	+	
7									+	#	#	#	-	+		
6		-				-		+	-	+	#	#	-	+	-	
10	-	+	#		#	#	#	-	#	+	-	-	#	#	-	
11	+	#	#		+	+	-		#	#	+	+	#	#	#	+
1		+	+	-					-	+		-	-	#	#	#
16	+	+	-											+	#	#

Rysunek 19. Diagram Czekanowskiego (uporządkowanie – 1).
Źródło: opracowanie własne.

Diagram powyższy (rysunek 19) ukazuje już pewne uporządkowanie, ale jednocześnie ujawnia również możliwość poprawy porządku, m.in. wiersz nr 10 należy przesunąć znacznie ku górze[86], a także warto zamienić miejscami parę wierszy/kolumn będących obok siebie. Wynik kolejnej iteracji porządkowania ujmuje rysunek 20.

Wynik tej iteracji nie jest jeszcze ostateczny. Pojawia się możliwość poprawy porządku poprzez przesunięcie wiersza/kolumny nr 8 na dalsze pozycje. Efekt tego przesunięcia prezentuje kolejny diagram (rysunek 21). Na rysunku tym do wierszy wprowadzono nazwy województw, przyjmując jednocześnie, że dalsze udoskonalanie porządku w zasadzie nie jest już możliwe. Co najważniejsze, uporządkowanie tego diagramu jest wystarczające do grupowania województwa na zasadzie największego podobieństwa.

[86] A jednocześnie kolumnę nr 10 przesunąć znacznie w stronę lewą diagramu.

	14	4	2	8	13	3	9	10	5	15	12	7	6	11	1	16
14	#	#	#		#	+	-	-	+					+		+
4	#	#	#	-	#	#	#	+	+	-	-		-	#	+	+
2	#	#	#	+	#	#	#	#	#	-	+			#	+	-
8		-	+	#	-	-	-			+	-				-	
13	#	#	#	-	#	#	#	#	#	-	+			+		
3	+	#	#	-	#	#	#	#	#	#	#		-	+		
9	-	#	#	-	#	#	#	#	#	#	+			-		
10	-	+	#		#	#	#	#	#	-	+	-	-	#	-	
5	+	+	#		#	#	#	#	#	+	#	+	-	#	-	
15		-	-	+	-	#	#	-	+	#	+		+			
12		-	+	-	+	#	+	+	#	+	#	#	+	#	+	
7								-	+		#	#	#	+		
6		-				-		-	-	+	+	#	#	+	-	
11	+	#	#		+	+	-	#	#		#	+	+	#	#	+
1		+	+	-				-	-		+		-	#	#	#
16	+	+	-											+	#	#

Rysunek 20. Diagram Czekanowskiego (uporządkowanie – 2).
Źródło: opracowanie własne.

Województwa	14	4	2	13	3	9	10	5	15	8	12	7	6	11	1	16
14. warmińsko-mazurskie	#	#	#	#	+	-	-	+						+		+
4. lubuskie	#	#	#	#	#	#	+	+	-	-	-		-	#	+	+
2. kujawsko-pomorskie	#	#	#	#	#	#	#	#	-	+	+			#	+	-
13. świętokrzyskie	#	#	#	#	#	#	#	#	-	-	+			+		
3. lubelskie	+	#	#	#	#	#	#	#	#	-	#		-	+		
9. podkarpackie	-	#	#	#	#	#	#	#	#	-	+			-		
10. podlaskie	-	+	#	#	#	#	#	#	-		+	-	-	#	-	
5. łódzkie	+	+	#	#	#	#	#	#	+		#	+	-	#	-	
15. wielkopolskie		-	-	-	#	#	-	+	#	+	+		+			
8. opolskie		-	+	-	-	-			+	#	-				-	
12. śląskie		-	+	+	#	+	+	#	+	-	#	#	+	#	+	
7. mazowieckie							-	+			#	#	#	+		
6. małopolskie		-			-		-	-	+		+	#	#	+	-	
11. pomorskie	+	#	#	+	+	-	#	#			#	+	+	#	#	+
1. dolnośląskie		+	+				-	-		-	+		-	#	#	#
16. zachodniopomorskie	+	+	-											+	#	#

Rysunek 21. Diagram Czekanowskiego (uporządkowanie – 3).
Źródło: opracowanie własne.

Na rysunku 21 kolorową linią zaznaczono wydzielone grupy województw. Sposób ich wyznaczenia jest dosyć jednoznacznie sugerowany przez kwadraty obrysowane wokół głównej przekątnej. Dwa ważne spostrzeżenia wymagają wyjaśnień. **Po pierwsze**, nie wszystkie grupy legitymują się najwyższymi wskaźnikami podobieństwa, a zatem należącymi do pierwszego przedziału (w kwadratach pojawiają się znaki reprezentujące drugi przedział podobieństwa). Przy czterech przedziałach podobieństwa dopuszczalne jest jednak pojawianie się podobieństw należących do drugiego przedziału, jednak bez ich liczebnej przewagi nad wskaźnikami należącymi do przedziału pierwszego. **Po drugie**, uporządkowany diagram może być podstawą nieco innego składu wydzielanych grup. Można bowiem przykładowo zdecydować, że województwo opolskie i wielkopolskie tworzą oddzielne jednoelementowe grupy (tak np. sugerowały wyniki dendrytu wrocławskiego). Nieco inny może być także skład grupy pierwszej i drugiej (patrząc od góry diagramu). Ten ostatni przypadek zaprezentowano na następnym, ostatnim już diagramie (rysunek 22). Różnica dotyczy województwa lubelskiego, które na podstawie zastosowanej metody należeć może zarówno do grupy pierwszej, jak też drugiej. Ostateczną decyzję podjąć musi analityk w oparciu o wiedzę merytoryczną dotyczącą tego województwa oraz województw grupy pierwszej i drugiej.

Województwa	**14**	**4**	**2**	**13**	**3**	**9**	**10**	**5**	**15**	**8**	**12**	**7**	**6**	**11**	**1**	**16**
14. warmińsko-mazurskie	#	#	#	#	+	-	-	+						+		+
4. lubuskie	#	#	#	#	#	#	+	+	-	-	-		-	#	+	+
2. kujawsko-pomorskie	#	#	#	#	#	#	#	#	-	+	+			#	+	-
13. świętokrzyskie	#	#	#	#	#	#	#	#	-	-	+			+		
3. lubelskie	+	#	#	#	#	#	#	#	#	-	#		-	+		
9. podkarpackie	-	#	#	#	#	#	#	#	#	-	+			-		
10. podlaskie	-	+	#	#	#	#	#	#	-		+	-	-	#	-	
5. łódzkie	+	+	#	#	#	#	#	#	+		#	+	-	#	-	
15. wielkopolskie		-	-	-	#	#	-	+	#	+	+		+			
8. opolskie		-	+	-	-	-			+	#	-				-	
12. śląskie		-	+	+	#	+	+	#	+	-	#	#	+	#	+	
7. mazowieckie							-	+			#	#	#	+		
6. małopolskie		-			-		-	-	+		+	#	#	+	-	
11. pomorskie	+	#	#	+	+	-	#	#			#	+	+	#	#	+
2. dolnośląskie		+	+				-	-		-	+		-	#	#	#
16. zachodniopomorskie	+	+	-											+	#	#

Rysunek 22. Diagram Czekanowskiego (uporządkowanie – 3a).
Źródło: opracowanie własne.

Pytania/zadania kontrolne

1. Na czym polega w metodzie diagramu Czekanowskiego upraszczanie informacji zawartych w macierzy odległości taksonomicznych?
2. Zastanowić się nad prawdziwością stwierdzenia: „Pomimo iż metoda diagramu Czekanowskiego nie sugeruje w pełni jednoznacznego grupowania badanych obiektów, to jednak decyzje dotyczące wyboru konkretnego wariantu nie są obarczone zbyt dużym ładunkiem subiektywizmu".
3. W jaki sposób dendryt wrocławski może być wykorzystany w roli wstępnego porządkowania diagramu Czekanowskiego?
4. W jakim zakresie wybór skali podobieństwa przesądzać może o wiarygodności wyników końcowego grupowania obiektów?

3.4. Metody taksonomiczne – metody aglomeracyjne

Ostatnią grupą metod taksonomicznych, z wcześniej zapowiadanego ich zestawu są metody aglomeracyjne. Uwagę zwraca liczba mnoga: „metody aglomeracyjne". W odróżnieniu od dotychczas omawianych, jest to cała grupa metod, których wspólną cechą jest sposób uogólniania informacji zawartych w macierzy odległości taksonomicznych, nawiązujący do użytego w nazwie określenia „aglomeracyjne", a konkretniej „aglomerować", czyli „skupiać".

Algorytm postępowania w stosowaniu metod aglomeracyjnych przedstawia się następująco:

1. Wyznaczenie macierzy odległości taksonomicznych.
2. Wyszukanie pary skupień (obiektów) „p" i „q" (p<q) najmniej odległych od siebie.
3. Połączenie skupienia „p" i „q" w jedno nowe skupienie o numerze „p" i usunięcie jednocześnie skupienia o numerze „q" (zmniejsza się o jeden numery skupień większe od „q").
4. Wyznaczenie odległości nowego skupienia od wszystkich pozostałych skupień według formuły:

$$D_{pr} = a_1 d_{pr} + a_2 d_{qr} + b d_{pq} + c \left| d_{pr} - d_{qr} \right| \qquad (25)$$

   - r przebiega wszystkie wartości różne od „p" i „q"
   - a1, a2, b, c są parametrami charakteryzującymi różne warianty metod aglomeracyjnych.
5. Powtarza się kroki 2-4, redukując stopniowo wymiar macierzy C o jeden, aż do momentu, gdy wszystkie obiekty utworzą jedną grupę.
6. Powstały w ten sposób „diagram drzewa" (dendrogram) jest podstawą do podziału niejednorodnego zbioru obiektów w podzbiory obiektów bardziej do siebie podobnych, przy jednoczesnej możliwości odczytu wielu parametrów charakteryzujących dokonywany podział.

Ogólny ogląd algorytmu wskazuje (patrz etap 5), że postępowanie związane ze stosowaniem metod aglomeracyjnych przebiega w sposób iteracyjny. Liczba iteracji jest z góry znana i wynosi[87] m-1. W każdej kolejnej iteracji dochodzi do zmniejszania o jeden wymiaru macierzy odległości taksonomicznych – o czym poniżej.

Pomijając omawianie etapu pierwszego ze zrozumiałych względów, najpierw skoncentrujmy się na dwóch kolejnych etapach. Są one kluczowe dla zrozumienia istoty omawianych metod; m.in. podpowiadają kierunek uogólniania/upraszczania informacji zawartych w macierzy odległości taksonomicznych. Odnosząc się do tych punktów algorytmu, wyjaśnienia wymagają dwie kwestie:

Po pierwsze, pojawia się nowe określenie – „skupienie", które występuje w kolejnych również etapach. W punkcie wyjścia – czyli w pierwszej iteracji – „skupienia" są tożsame z „obiektami" (u nas jednostkami terytorialnymi) podlegającymi grupowaniu. W każdej kolejnej iteracji dwa skupienia spełniające kryterium podane w punkcie 2 (najmniej odległe od siebie) są łączone (pochłaniane) przez jedno, prowadząc do jeszcze bardziej zaglomerowanego skupienia. Sens tego łącznia wyjaśnimy poprzez następujące rozumowanie. W wyjściowej macierzy odległości taksonomicznej, a więc w postaci macierzy nam znanej, znaleźć należy najmniejszą wartość (ignorując oczywiście zera na głównej przekątnej). Obrazuje ona odległość taksonomiczną dwóch najbardziej do siebie podobnych obiektów (jednostek terytorialnych) w całym rozważanym ich zbiorze (np. miast w Polsce, powiatów, województw, krajów). Ujmując teraz obrazowo, wręcz metaforycznie, wniosek nasuwający się z tego ustalenia możemy sformułować następująco: skoro te dwa obiekty są do siebie tak bardzo podobne[88], to nie rozróżniajmy ich jako dwóch odrębnych bytów, lecz jako jeden, zaglomerowany, abstrakcyjny byt (skupienie). Zauważmy, że w ten sposób postępując, skutkiem każdej kolejnej iteracji, jest wskazywana powyżej redukcja wymiaru macierzy odległości taksonomicznej o jeden. Postępujemy tak aż do otrzymania wymiaru macierzy 1x1 w ostatniej iteracji.

Po drugie, w omawianym punkcie algorytmu pojawiają się oznaczenie „p" oraz „q". Znaleziony w macierzy odległości taksonomicznych, w danej iteracji, najmniejszy wskaźnik podobieństwa znajduje się na przecięciu określonego wiersza i kolumny, których numery wyznaczają właśnie wartości p oraz g. Nałożony warunek, aby p było mniejsze od q wynika z potrzeby zachowania w każdej iteracji takiego samego postępowania, np. które skupienie (wiersz w macierzy) usuwamy, a w miejsce którego wpisujemy wskaźniki podobieństwa nowego skupienia. Łącząc dwa skupienia w jedno w danej iteracji, musimy dokonać zmian w układzie macierzy odległości taksonomicznych, o czym informuje punkt 3 algorytmu. Przede wszystkim należy zwrócić uwagę, że tworząc nowe skupienie powstałe z dwóch innych, zniknąć muszą wartości wskaźników z dwóch wierszy i dwóch kolumn, a więc o numerze p oraz q (przestają one być aktualne, gdyż obiekty/skupienia, które reprezentowały znikają). Musimy natomiast wyznaczyć nowe wskaźniki podobieństwa dla nowego skupienia (sposób ich wyznaczania wyjaśniamy poniżej), które umieścimy w wierszu i kolumnie o numerze p. Wiersz i kolumnę o numerze q odpowiednio usuwamy według sposobu wskazanego w punkcie 3 algorytmu.

[87] „m" jest liczbą obiektów podlegających grupowaniu.
[88] Używamy trochę przesadnego stwierdzenia („tak bardzo do siebie podobne") dla wyeksponowania zasadności łączenia obiektów.

Punkt 4 algorytmu podaje regułę wyznaczania wskaźników podobieństwa (odległości taksonomicznych) dla nowego skupienia, powstającego w każdej kolejnej iteracji. Łącząc w danej iteracji dwa skupienia, powstaje skupienie nowe, którego podobieństwa z pozostałymi skupieniami nie znamy, a zatem musimy je wyznaczać. W podanym wzorze pojawiają się parametry, które rozróżniają podejścia do wyliczania wskaźników podobieństwa, a więc rozróżniają konkretne rodzaje taksonomicznych metod aglomeracyjnych. Są to parametry: a_1, a_2, b, c, których postać omówiona zostanie nieco dalej. Najpierw jednak wyjaśnijmy ogólną ideę wyznaczania wskaźników podobieństwa dla skupienia nowopowstałego w każdej iteracji, w tym przyjęte oznaczenia.

Po lewej stronie równania (25) D_{pr} jest poszukiwaną wartością (wskaźnikiem podobieństwa) po dokonaniu już łączenia w danej iteracji, a zatem oznacza wskaźnik podobieństwa nowego skupienia „p" powstałego w każdej z kolejnych „k" iteracji, ze skupieniem „r". Należy zauważyć, że po wykonaniu w danej iteracji łączenia dwóch skupień, do wyliczenia pozostaje m-k wskaźników podobieństwa. Musimy wyliczyć wskaźniki podobieństwa nowego skupienia ze wszystkimi pozostałymi skupieniami, czyli r przebiega od 1 do m-k. Po prawej stronie równania występują parametry d_{pr}, d_{qr}, oraz d_{pq}, będące wskaźnikami podobieństw, których miejsce w macierzy odległości taksonomicznych definiują subskrypty p, r oraz q. Są to odpowiednie wartości podobieństw przed dokonaniem jeszcze łączenia.

Jak pamiętamy z powyższych wyjaśnień, wyliczone wskaźniki podobieństw umieszczamy w wierszu oraz kolumnie o numerze p. Po wykonaniu każdej iteracji powstaje nowa, nieco zmodyfikowana macierz odległości taksonomicznych. Różni się od macierzy poprzedniej iteracji nie tylko o jeden mniejszym wymiarem, ale również wartościami wskaźników umieszczonych w wierszu i kolumnie p.

Powtarzając wszystkie kolejne iteracje, w finale otrzymujemy tzw. „diagram drzewa", z którego możemy odczytać nie tylko liczbę i skład grup obiektów, ale także szereg innych ważnych charakterystyk dokonywanego grupowania. Będzie to dokładnie wyjaśnione przy okazji omawiania przykładu na końcu niniejszego podrozdziału.

Wróćmy do parametrów różnicujących konkretne taksonomiczne metody aglomeracyjne: a_1, a_2, b, c. Umieszczone w tabeli 31 ich wartości wyjaśniają, po części przynajmniej, istotę różnych metod aglomeracyjnych. W ogólnej ocenie poszczególne metody, oprócz różnic formalnych, powodowanych wartościami parametrów, cechują się różnym stopniem upraszczania. Największe upraszczanie ma miejsce w przypadku stosowania dwóch pierwszych metod, najbardziej zaś zaawansowaną koncepcyjnie i o najmniejszym stopniu upraszczania jest metoda ostatnia z wymienionych w tabeli 31.

Tabela 31

Warianty metod aglomeracyjnych stosownie do wartości przyjmowanych przez parametry: a1, a2, b, c

Metoda	**a_1**	**a_2**	**b**	**c**
Najbliższego sąsiedztwa	0,5	0,5	0	-0,5
Najdalszego sąsiedztwa	0,5	0,5	0	0,5
Skupienia parami	0,5	0,5	0	0
Mediany	0,5	0,5	-0,25	0
Środka ciężkości	$\frac{n_p}{n_p+n_q}$	$\frac{n_q}{n_p+n_q}$	$-a_1a_2$	0
Średniej grupowej	$\frac{n_p}{n_p+n_q}$	$\frac{n_q}{n_p+n_q}$	0	0
Warda	$\frac{n_p+n_r}{n_p+n_q+n_r}$	$\frac{n_q+n_r}{n_p+n_q+n_r}$	$\frac{-n_r}{n_p+n_q+n_r}$	0

Źródło: opracowanie własne z wykorzystaniem *Zastosowanie wybranych metod taksonomicznych w badaniach historycznych* (s. 129), L. Błażejczyk-Majka, 2018, Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu.

Wyjaśnienie: n_p, n_q, n_r oznaczają liczebność obiektów pierwotnych w skupieniu, odpowiednio p, q oraz r.

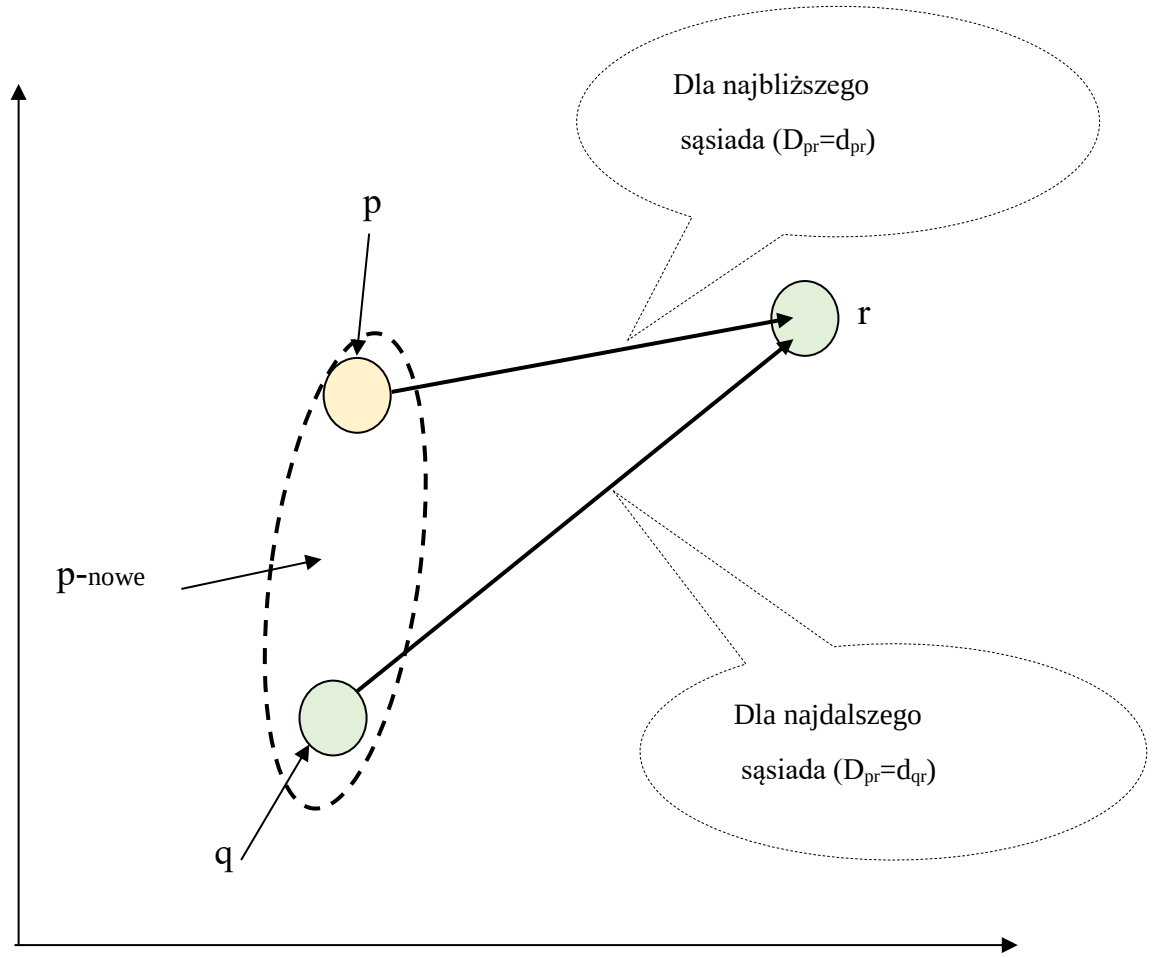
W interpretacji sposobów wyznaczania wskaźników podobieństwa dla nowo powstałego skupienia za pomocą poszczególnych metod aglomeracyjnych, pomocnym będzie schemat ujęty rysunkiem 23.

Z rysunku tego wynika, że w którejś iteracji[89] łączymy dwa skupienia: p oraz q i powstaje skupienie nowe (p-nowe). W dwóch pierwszych metodach, dla ustalenia podobieństwa nowo powstałego skupienia ze skupieniem r, niczego nie wyliczamy, lecz przyjmujemy z góry:

- w metodzie *najbliższego sąsiada* podobieństwo to reprezentowane jest przez podobieństwo skupienia najbliżej położonego względem skupienia r;
- zaś w metodzie *najdalszego sąsiada*, reprezentowane jest przez podobieństwo skupienia najdalej położonego.

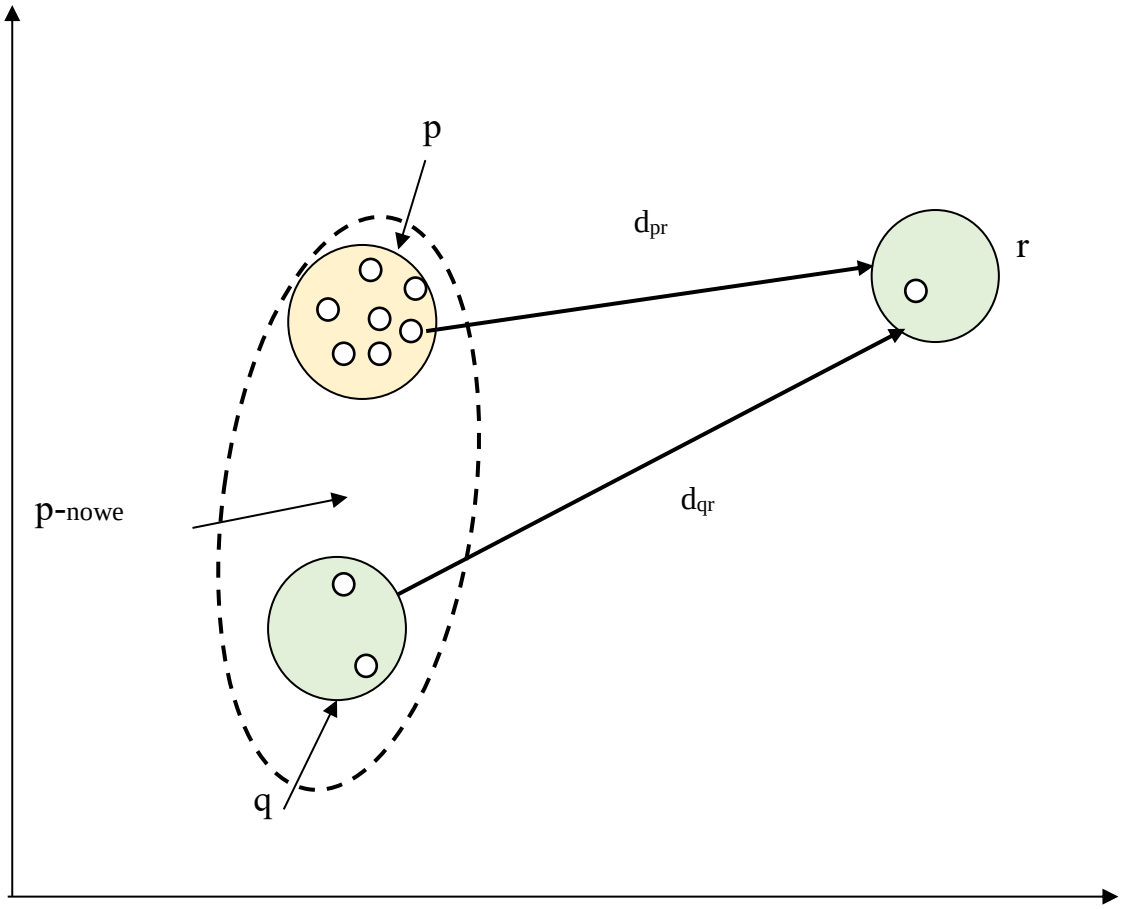
W metodzie *skupienia parami* wskaźniki podobieństwa wyznaczane są jako średnie arytmetyczna z wartości d_{pr} oraz d_{qr}. Również w metodzie mediany wartości parametrów (a_1, a_2, b, c) są z góry ustalone (patrz tabela 31). Nieco inaczej sytuacja wygląda w trzech ostatnich metodach. W ich charakterystyce pomocnym będzie rysunek 24, nieco zmodyfikowany w stosunku do rysunku wcześniejszego (23).

[89] Dla zrozumienia przedstawianych wyjaśnień nie jest ważne, która to jest iteracja.



Rysunek 23. Ilustracja wyznaczania wskaźnika podobieństwa metodą najbliższego i najdalszego sąsiada dla nowo powstałego skupienia.
Źródło: opracowanie własne.

Wyobraźmy sobie, że w pewnej iteracji łączymy ze sobą w jedno skupienie dwa najbardziej do siebie podobne, ale obejmujące różną liczbę obiektów pierwotnych (np. jednostek terytorialnych). Z zamieszczonego rysunku wynika, że skupienie p obejmuje 7 pierwotnych obiektów, zaś skupienie q tylko 2 obiekty. Nie trzeba chyba przekonywać, że skupienie p powinno mieć większy wpływ na wartość wskaźnika podobieństwa nowego skupienia (p-nowe) z każdym innym skupieniem, aniżeli skupienie q. Krótko charakteryzowane wcześniejsze metody tego nie zakładały. Natomiast w metodach *środka ciężkości, średniej grupowej oraz Warda* parametry a_1, a_2 oraz b są wyliczane z uwzględnieniem liczby obiektów pierwotnych w poszczególnych skupieniach danej iteracji, stanowiąc jednocześnie wagi w wyliczaniu wskazywanych parametrów stosownie do rozkładu liczby obiektów pierwotnych.



Rysunek 24. Ilustracja wyznaczania wskaźnika podobieństwa dla nowo powstałego skupienia, z uwzględnieniem wag łączonych skupień.
Źródło: opracowanie własne.

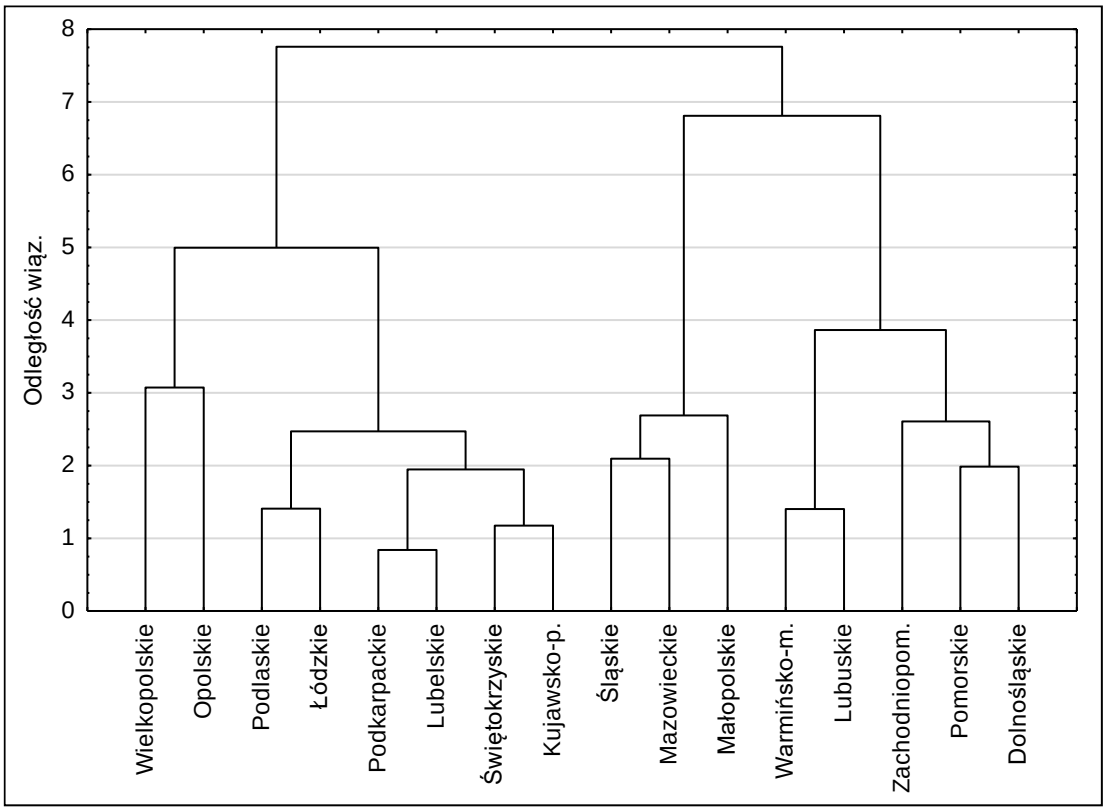
Najbardziej zaawansowaną koncepcyjnie jest metoda Warda. Najogólniej biorąc, do oszacowania odległości między skupieniami wykorzystuje podejście analizy wariancji. Mówiąc krótko, metoda ta zmierza do minimalizacji sumy kwadratów odchyleń dowolnych dwóch skupień, które mogą zostać uformowane na każdym etapie (iteracji). Choć spośród wszystkich metod aglomeracyjnych ta właśnie metoda jest powszechnie rekomendowana do wykorzystania w praktyce, jej założenia oraz postać analityczna nie będą jednak przedmiotem szczegółowych rozważań. Wykracza to poza obowiązujący zakres wiedzy z zakresu statystyki matematycznej wymagany na większości kierunków studiów, dla których rekomendowany jest przedstawiany podręcznik (zob. informacja przedstawiona we wstępie). Nie jest to jednak przeszkodą w nabyciu umiejętności posługiwania się metodą Warda. Metoda ta ma bowiem doskonałe oprogramowanie w pakietach statystycznych, np. *statistica*, stąd więc łatwość jej użycia w konkretnym postępowaniu badawczym. Jest ona uważana za jedną z najbardziej efektywnych metod statystycznych, co przekłada się na częstotliwość jej zastosowań w praktyce. Ważne walory tej metody zostaną wyjaśnione przy okazji prezentowania konkretnego przykładu.

Na zakończenie parę ważnych ustaleń dotyczących ostatniego etapu, a mianowicie diagramu drzewa, inaczej dendrogramu, który jest podstawą wydzielania grup podobnych do siebie obiektów. Obrazuje on graficznie wyniki wszystkich iteracji, informując o skupieniach łączonych w kolejnych iteracjach, stopniu ich podobieństwa, a przede wszystkim o możliwych do wydzielenia grupach obiektów podobnych do siebie. Prześledzimy to z wykorzystaniem konkretnego przykładu, prezentowanego w poprzednich podrozdziałach dotyczących metod taksonomicznych. Warto przypomnieć, że zadaniem ujętym w przykładzie jest podział polskich województw w grupy względnie jednorodne, z punktu widzenia kryterium, którym jest struktura osiągniętego poziomu rozwoju[90]. Przechodząc kolejne etapy metody aglomeracyjnej Warda, w efekcie otrzymujemy dendrogram postaci prezentowanej na rysunku 25. Oś pionowa mierzy odległości wiązania, czyli poziom podobieństwa łączonych skupień. Należy więc zwrócić uwagę, że długość linii pionowych nie jest przypadkowa. Jest ona proporcjonalna do wartości wskaźnika podobieństwa, na podstawie którego łączone są kolejne skupienia.

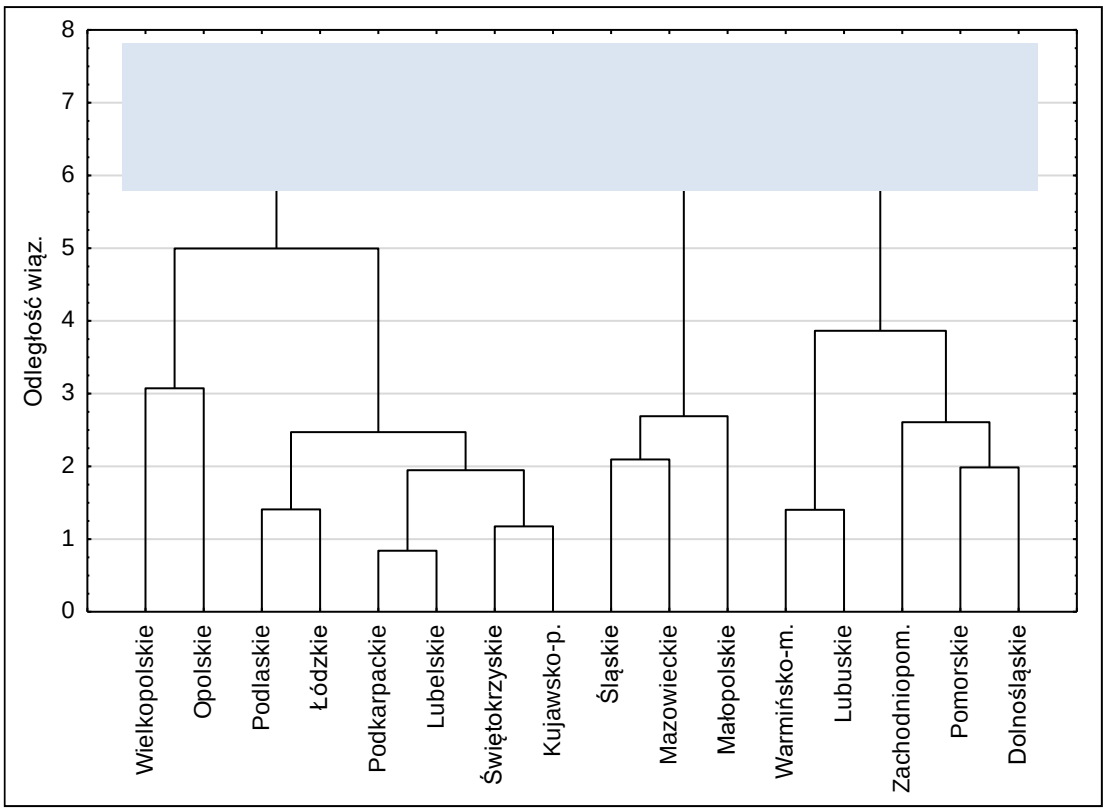
Z przedstawionego dendrogramu odczytać można bardzo ważne informacje dotyczące klasyfikowanych obiektów. Przede wszystkim łatwo ustalić można liczbę oraz skład grup, które obejmują województwa podobne do siebie pod względem rozważanego kryterium, z jednoczesnym odczytaniem poziomu jednorodności podziału. Należy ponownie wyjaśnić, że poziom jednorodności podziału jest wyznaczany przez najmniej jednorodną grupę (wrócimy do tego zagadnienia poniżej). Przykładowo, gdyby podjąć decyzję, że chcemy wydzielić trzy grupy, wynik takiej decyzji prezentuje rysunek 26. Wystarczy w tym celu, jak zaznaczono na rysunku obrazującym diagram drzewa, przyłożyć kartkę, zasłaniając górną część drzewa w ten sposób, aby przerwać trzy wiązadła. Rysunki 27 i 28 ilustrują podział, odpowiednio na cztery oraz pięć grup województw.

Zauważamy zatem, że diagram stwarza możliwość ustalenia dowolnej liczby grup pomiędzy 1 a „m" (u nas m=16). Dla każdego jednocześnie podziału ustalić można poziom jego jednorodności według zasady wcześniej wspominanej, a mianowicie, że o poziomie tym decyduje najmniej jednorodna grupa. Pamiętać przy tym należy, że długość linii drzewa jest proporcjonalna do stopnia podobieństwa skupień, które łączy. Poziom jednorodności odczytujemy na osi pionowej. Przy podziale na trzy grupy (rysunek 26) najmniej jednorodna grupa obejmuje województwa od wielkopolskiego do kujawsko-pomorskiego.

[90] Konkretne cech szczegółowe, sposoby i wyniki ich standaryzacji były omawiane we wcześniejszym rozdziale.



Rysunek 25. Diagram drzewa klasyfikacji województw pod względem poziomu rozwoju metodą Warda.
Źródło: opracowanie własne.



Rysunek 26. Trzy grupy województw wydzielone pod względem podobieństwa poziomu rozwoju.
Źródło: opracowanie własne.

Jej poziom jednorodności wynosi 5. Możemy zatem powiedzieć, że taki też jest poziom jednorodności podziału na trzy grupy. Przy podziale na cztery grupy (rysunek 27) poziom jednorodności podziału jest determinowany przez grupę obejmującą województwa od warmińsko-mazurskiego do dolnośląskiego i wynosi nieco poniżej 4. Podziałowi na pięć grup możemy zaś przypisać poziom jednorodności wynoszący nieco ponad 3 (województwo wielkopolskie i opolskie).

Z diagramu odczytać można wiele innych informacji. Łatwo ustalić można, jakie skupienia łączone były w każdej iteracji. Przykładowo, w iteracji pierwszej łączone były województwa lubelskie i podkarpackie. Informuje o tym łączenie położone najbliżej osi poziomej wykresu. W iteracji drugiej łączone były województwa świętokrzyskie i kujawsko-pomorskie. Iteracja następna, trzecia, stwarza pewne problemy z ustaleniem obiektów (skupień) łączonych z uwagi na wysoką zbieżność na rysunku łączenia województw: podlaskiego i łódzkiego oraz warmińsko-mazurskiego i lubuskiego. W tym celu trzeba byłoby sięgnąć do macierzy odległości (tabela 29), z której wynika, że w trzeciej iteracji łączone były województwa warmińsko-mazurskie oraz lubuskie[91]. W iteracji przedostatniej łączone były dwa skupienia: śląskie, mazowieckie, małopolskie ze skupieniem: warmińsko-mazurskie, lubuskie, zachodniopomorskie, pomorskie, dolnośląskie. W iteracji natomiast ostatniej łączone były skupienia: jedno zaczynające się od województwa wielkopolskiego i kończące na województwie kujawsko-pomorskim, drugie zaczynające się od województwa śląskiego i kończące na województwie dolnośląskim.

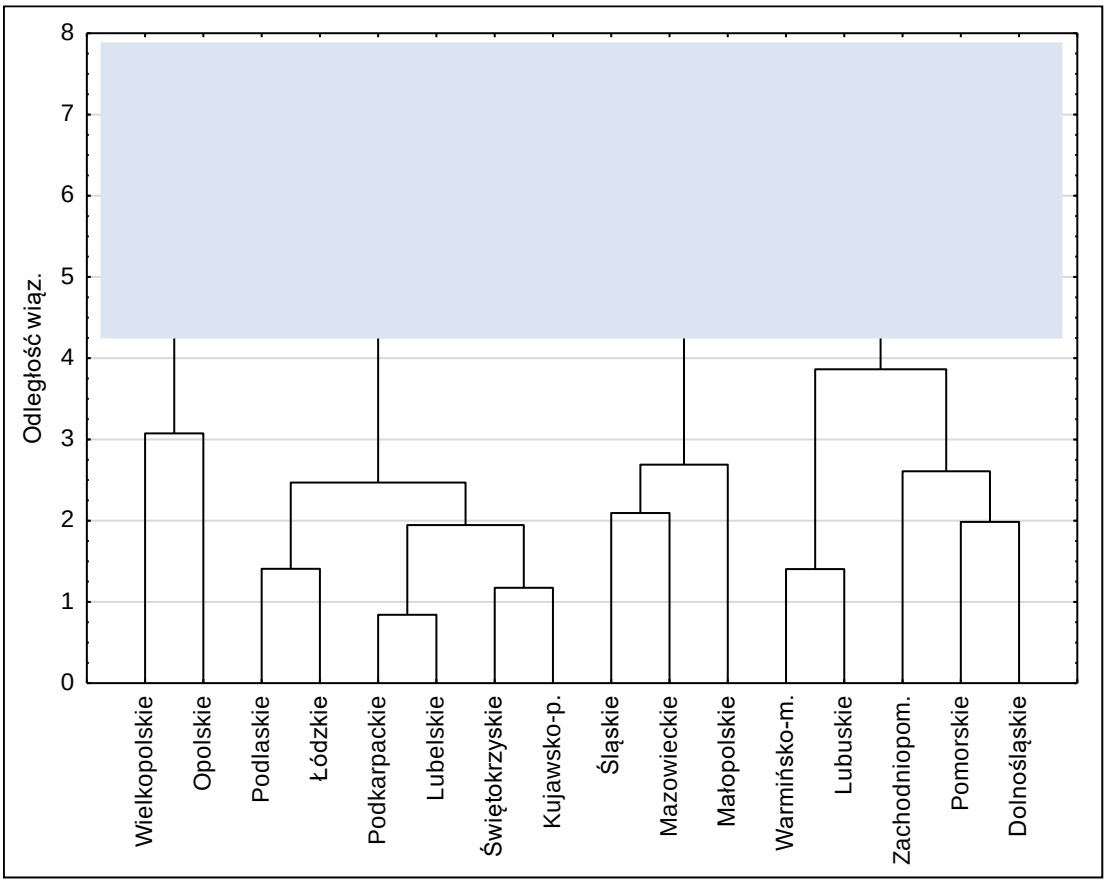
Diagram pozwala również na ustalenie parametrów n_p, n_q. Przypomnijmy, że parametry te oznaczają liczebność obiektów pierwotnych (u nas województw) w skupieniu p oraz q. Od razu należy wyjaśnić, że dendryt stwarza możliwość odczytania wartości tych parametrów, ale bez możliwości rozróżniania, który jest o numerze p, a który jest o numerze q. W początkowych iteracjach, a na pewno w iteracji pierwszej, omawiane parametry przyjmują wartość 1.

Wracając do naszego przykładu, zauważmy, że w iteracji ostatniej wartość jednego i drugiego parametru wynosi 8. W iteracji przedostatniej jeden z tych parametrów przyjmuje wartość 3 drugi zaś wartość 5. W iteracji trzeciej od końca n_p, n_q przyjmują wartość odpowiednio 2 oraz 6.

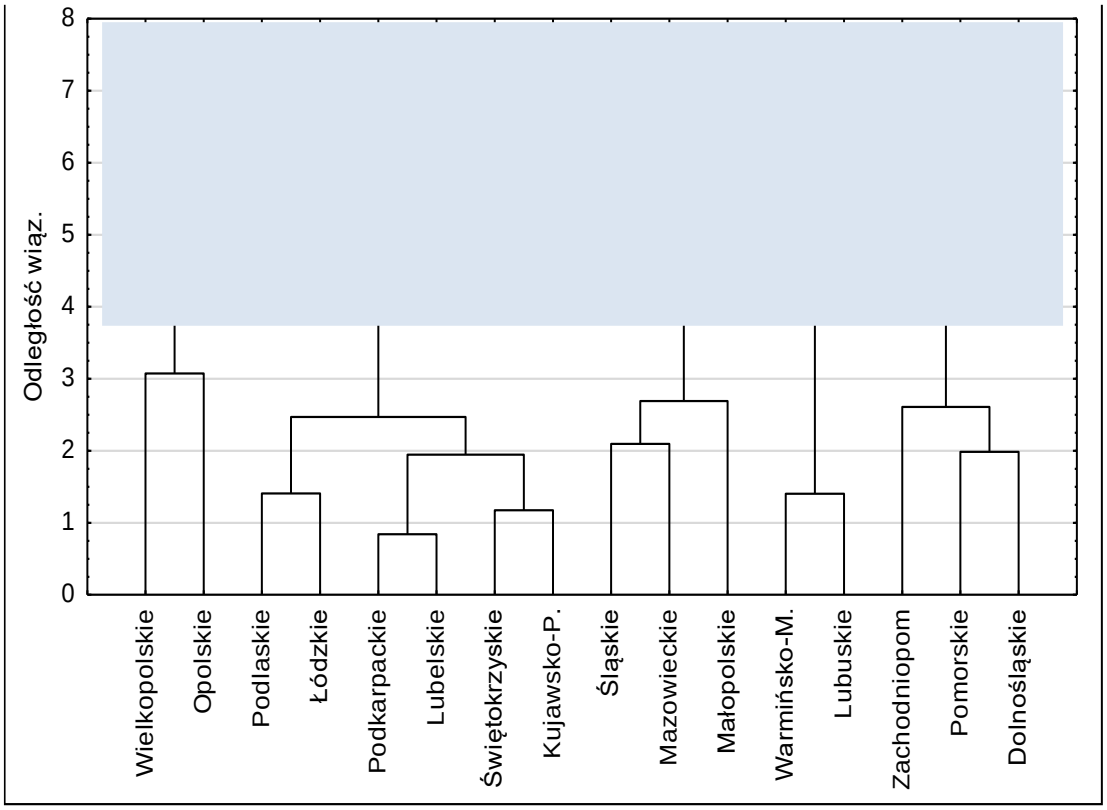
Na zakończenie, przedstawionych zostanie kilka ogólnych i podsumowujących uwag dotyczących metod aglomeracyjnych, ze szczególnym uwzględnieniem metody Warda.

1. Metody aglomeracyjne, a w szczególności metoda Warda, w najszerszym zakresie – w stosunku do metod wcześniej omawianych – wykorzystują informacje zawarte w macierzy odległości taksonomicznych. Inaczej ujmując, stopień upraszczania przy ich wykorzystywaniu jest najmniejszy.

[91] Należy od razu wyjaśnić, że ustalenie to jest możliwe na podstawie macierzy odległości, gdyż – jak się okazuje – w trzeciej iteracji wszystkie cztery wymienione skupienia były jeszcze jednoelementowe.



Rysunek 27. Cztery grupy województw wydzielone pod względem podobieństwa poziomu rozwoju.
Źródło: opracowanie własne.



Rysunek 28. Pięć grup województw wydzielonych pod względem podobieństwa poziomu rozwoju.
Źródło: opracowanie własne.

2. Diagram drzewa, ukazując pełną historię (przebieg) aglomerowania, stwarza możliwość wyboru najbardziej trafnej liczby grup, a więc dokonania racjonalnego podziału niejednorodnego zbioru obiektów w podzbiory obiektów bardziej do siebie podobnych. Przesądza o tym łatwy do ustalenia poziom jednorodności podziału, który odczytać możemy dla każdego z możliwych podziałów (a więc liczby przyjętych grup). Jest oczywistym, że decydując się na mniejszą liczbę grup, zawsze musimy liczyć się z tym, że poziom jednorodności podziału będzie mniejszy niż w sytuacji, gdyby tych grup było więcej. Jest to ogólna prawidłowość niewymagająca chyba dalszych wyjaśnień. Zwrócić natomiast należy uwagę na to, że dendrogram ujawnia nie tylko pełne spektrum możliwości podziału – począwszy od liczby grup wynoszącej 1, aż do liczby „m" grup – ale także pozwala przypisać każdemu podziałowi parametr przesądzający o jego jakości. Parametrem tym jest właśnie poziom jednorodności podziału. Jak pamiętamy, metoda dendrytu wrocławskiego również oferowała pełny przegląd możliwych podziałów (grupowań). Rzecz tylko w tym, że wybór wariantu podziału musiał być robiony niejako „w ciemno", a więc bez możliwości oceny jego jakości.
3. Jak już wspominano wcześniej, metoda Warda jest metodą najbardziej zaawansowaną koncepcyjnie. Odnosi się to do sposobu ustalania pary skupień łączonych w każdej iteracji, jak również do wyliczania wskaźników podobieństwa nowo powstałego skupienia ze skupieniami pozostałymi.
4. Wskazywane wyżej walory metod aglomeracyjnych przesądzają o tym, że są one obecnie najczęściej stosowanymi metodami taksonomicznymi w praktyce. Dokonywane na ich podstawie klasyfikacje budzą duże zaufanie. Zdecydowanym „liderem" w tym względzie jest oczywiście metoda Warda. O powszechności stosowania omawianych metod przesądza dodatkowo pełne ich oprogramowanie w specjalistycznych pakietach statystycznych, np. w *statistica*. Ich stosowanie jest więc relatywnie łatwe, co nie jest bez znaczenia dla wspominanej powszechności ich wykorzystania.

Pytania/zadania kontrolne

1. Na czym polega w metodach aglomeracyjnych upraszczanie informacji zawartych w macierzy odległości taksonomicznych?
2. Spróbuj uzasadnić, że w stosunku do metod wcześniej omawianych metody aglomeracyjne cechują się najmniejszym stopniem upraszczania.
3. Metody aglomeracyjne, podobnie jak dendryt wrocławski, umożliwiają pełny przegląd liczby możliwych grup podziału badanych obiektów. Jaka jest jednak przewaga w tym względzie metod aglomeracyjnych?
4. Abstrahując od praktycznej użyteczności dokonywanego podziału, przy jakiej liczbie grup otrzymywanych za pomocą metod aglomeracyjnych poziom jednorodności podziału jest największy, a przy jakiej liczbie grup poziom ten jest najmniejszy?
5. Co jest główną mocną stroną taksonomicznych metod aglomeracyjnych?

3.5. Metody taksonomiczne – metoda k-średnich

W przedstawionych powyżej konkluzjach dotyczących metod aglomeracyjnych podkreślone zostały ważne walory przesądzające o ich praktycznej użyteczności. Celem pełniejszej charakterystyki tych metod wymienione zalety uzupełnić jednak trzeba wskazaniem także ich słabych stron. **Po pierwsze**, metody te nie dają ostatecznej odpowiedzi, który podział (grupowanie) jest optymalny. W wyniku ich zastosowania uzyskiwany jest układ skupień tworzących swoistą hierarchię[92], tzn. ukazywana jest możliwość podziału zbioru obiektów na dowolną liczbę grup, począwszy od jednej, a skończywszy na m grupach; a nieco inaczej ujmując, począwszy od znalezienia dwóch obiektów najbardziej podobnych do siebie w pierwszej iteracji, aż do połączenia wszystkich obiektów w jedną grupę w iteracji ostatniej. Jak pamiętamy, na podstawie wizualnej oceny dendrogramu, uzupełnionej ewentualnie dodatkowo wskazaniami wynikającymi z celu (przydatności) przeprowadzanego grupowania, podejmujemy decyzję, który podział, a więc układ grup jest najbardziej odpowiedni. Dla ilustracji problemu odwołajmy się do naszego przykładu, w którym do grupowania województw zastosowano metodę Warda. Zwróćmy uwagę na otrzymane wyniki obrazowane rysunkiem 27 i 28. Jak już wspominano, metoda Warda zmierza do minimalizacji sumy kwadratów odchyleń wewnątrz skupień, a więc celem tej metody jest takie łączenie obiektów, żeby w powstałych skupieniach wariancja wewnątrzgrupowa zmiennych opisujących obiekty była możliwie mała. Przywołane dwa rysunki sugerują podział badanego zbioru obiektów, odpowiednio na cztery i pięć grup. Metoda Warda nie dostarcza jednak jednoznacznych wskazań, który z tych dwóch alternatywnych podziałów jest lepszy. Badacz musi zatem podjąć decyzje o wyborze, jego zdaniem, najlepszego podziału. **Po drugie**, słabą stroną metod aglomeracyjnych jest to, że są one użyteczne dla małych zbiorów grupowanych obiektów. W przypadku większej ich liczby użyteczność omawianych metod spada ze względu czytelność dendrogramu. To właśnie dendrogram jest główną podstawą intuicyjnego, wzrokowego odczytywania „najlepszego", w opinii badacza, zarówno podziału, jak i składu poszczególnych grup. Przy większej liczbie obiektów, przykładowo grupując polskie powiaty, nie ma możliwości odczytania składów grup wprost z dendrogramu.

Tych właśnie słabych stron nie posiada metoda k-średnich, której charakterystyka przedstawiona zostanie w niniejszym podrozdziale. Najogólniej ujmując, istota tej metody polega na ulepszaniu przyporządkowania obiektów do poszczególnych skupisk (grup) w kolejno przeprowadzanych iteracjach. Ulepszanie to dokonywane jest w myśl kryterium: minimalizowania wariancji wewnątrzgrupowych przy jednoczesnym maksymalizowaniu wariancji międzygrupowej. Oznacza to, że przy ustalonej liczbie grup otrzymujemy uporządkowanie, w którym grupy cechują się możliwie największym wewnętrznym podobieństwem skupianych obiektów i równocześnie możliwie najmniejszym podobieństwem między grupami.

Na początku omawiania metod taksonomicznych wspomniane było, że metoda k-średnich różni się od metod taksonomicznych powyżej scharakteryzowanych jednym ważnym szczegółem – nie wymaga wyznaczenia macierzy odległości taksonomicznych postaci (24). Procedura jej stosowania opiera się na odmiennej od wcześniej omawianej idei, a mianowicie nie sprowadza się do swoistego pomysłu uogólniania informacji zawartych w macierzy odległości taksonomicznych. Rozbudowany, pełny algorytm postępowania w stosowaniu metody k-średnich prezentuje zestawienie ujęte w tabeli 32.

[92] Dlatego też metody aglomeracyjne nazywane są również taksonomicznymi metodami hierarchicznymi.

Tabela 32

Metody taksonomiczne w zastosowaniu do klasyfikacji jednostek terytorialnych – główne etapy procedury metody k-średnich

1. **Sformułowanie problemu (określenie kryterium klasyfikacji):**
   a) Terytorialny system społeczno-gospodarczy [TSSG] (np. kraj, region)
   b) Jednostki tworzące ten system
   c) Kryterium oceny (zjawisko podlegające ocenie w ramach TSSG).
2. **Dyskusja cech – mierników szczegółowych grupowania.**
3. **Wybór cech diagnostycznych – metody:**
   a) metoda grafu
   b) metoda dendrytowa.
4. **Standaryzacja cech diagnostycznych; metody:**
   a) *zero-jedynkowa*
   b) uproszczona
   c) *min-max*.
5. **Zdefiniowanie kluczowych założeń wyjściowych oraz pierwsze porządkowanie (grupowanie) obiektów:**
   a) ustalenia liczby skupień (grup)[93]
   b) zdefiniowanie pierwszych centroidów (środków ciężkości)
   c) wyznaczenie odległości wszystkich obiektów od zdefiniowanych pierwszych centroidów
   d) przyporządkowanie obiektów do właściwych skupień.
6. **Porządkowanie obiektów w oparciu o przyjęte kryterium**
   a) wyliczenie współrzędnych kolejnych środków ciężkości dla nowo powstałych skupień
   b) wyliczenie odległości wszystkich obiektów od wyznaczonych centroidów
   c) ponowne przyporządkowanie obiektów do właściwych skupień
   d) powtarzamy kroki 6a – 6c, aż do wyczerpania liczby zadanych iteracji albo do momentu spełnienie zadanego kryterium.
7. **Interpretacja wyników.**

Źródło: opracowanie własne.

[93] W metodzie tej określenie „*skupienie*" stosowane jest zamiennie z określeniem „*grupa*".

Przytoczony pełny algorytm postępowania, począwszy od etapu „sformułowanie problemu", przypomnieć ma o omawianej już wcześniej zbieżności metod taksonomicznych z metodami syntetycznej oceny, a tym samym o możliwości pominięcia etapów, które były już dokładnie charakteryzowane. Są to etapy 1-4. Możemy więc przyjąć, że punktem wyjścia do omawiania taksonomicznej metody k-średnich jest zbiór standaryzowanych cech diagnostycznych. Przejdźmy zatem do omawiania kolejnych etapów zaprezentowanej procedury postępowania.

Ad. 5) Kluczowe założenia wyjściowe dotyczą dwóch ważnych kwestii:

- ustalenia liczby grup (5a),
- zdefiniowanie ich środków ciężkości (5b).

Z istoty omawianej metody wynika, że wymaga ona „zadeklarowania" z góry liczby grup, na które zamierzamy podzielić wyjściowy zbiór obiektów. Jest to niewątpliwie słaba strona metody k-średnich w stosunku do metod wcześniej zaprezentowanych, w których to liczba grup była jednym z parametrów wynikowych odpowiedniego postępowania. Wskazać można dwa podejścia do ustalania liczby grup. Jedno ma charakter losowy, tzn. przyjmujemy jakąś liczbę grup i uruchamiamy dalsze etapy postępowania, aż do otrzymania ostatecznych wyników, tzn. przyporządkowania obiektów do odpowiednich grup. Ewentualnie, na podstawie przesłanek merytorycznych dokonujemy oceny tak otrzymanych składów grup i podejmujemy decyzję o powtórzeniu procedury grupowania, ale przy innej ich liczbie. Drugie podejście do wyznaczania liczby grup – często rekomendowane – polega na wykorzystaniu metody aglomeracyjnej (np. Warda), jako narzędzia do podpowiedzi w tym względzie. Jak już wiemy, przy dużej liczbie grupowanych obiektów z dendrogramu trudno byłoby odczytać skład poszczególnych grup. Można natomiast relatywnie łatwo odczytać sugerowaną przez taki dendrogram liczbę grup.

Odrębną kwestią wymagającą wstępnego przyjęcia są parametry ustalające środki ciężkości (centroidy) poszczególnych, przyjętych grup. Środki ciężkości są jednym z wiodących pojęć mających zastosowanie w omawianej metodzie. Rezygnując z precyzyjnego definiowania środka ciężkości, ograniczymy się do bardziej obrazowej jego charakterystyki. Środkiem ciężkości dla grupy „k" punktów o współrzędnych opisanych wektorami (26)

$$
\begin{aligned}
&P_1: \{t_{11}, t_{12}, t_{13}, \ldots, t_{1n}\} \\
&P_2: \{t_{21}, t_{22}, t_{23}, \ldots, t_{2n}\} \\
&P_3: \{t_{31}, t_{32}, t_{33}, \ldots, t_{3n}\} \qquad (26) \\
&\quad \ldots\ldots\ldots\ldots\ldots \\
&P_k: \{t_{k1}, t_{42}, t_{43}, \ldots, t_{kn}\}
\end{aligned}
$$

będzie inny punkt o takiej samej liczbie współrzędnych, które to współrzędne zostają wyznaczone według reguły poniżej zaprezentowanej.

Wykonanie pierwszego kroku etapu grupowania (etap 5a powyższego algorytmu) prowadzi zatem do ustalenia liczby grup, na które chcemy dzielić nasze obiekty według zasady jak największego podobieństwa. Oznaczmy tę przyjętą liczbę grup przez „g". Nie znamy jednak w tej fazie postępowania składu tych grup. Z konieczności więc współrzędne środka ciężkości ustalamy w sposób losowy. Pamiętając, że współrzędnymi są wartości standaryzowane cech, korzystamy jedynie ze znanych właściwości zmiennych standaryzowanych. Przykładowo, jeżeli zmienna standaryzowana jest formułą *zero-jedynkową*, to wówczas niewielkie jest

prawdopodobieństwo tego, że wartości te będą spoza przedziału [-3; 3] (reguła trzech sigm). Mało uzasadnione byłoby więc definiowanie poszczególnych środków ciężkości współrzędnymi o wartościach spoza tego przedziału. W omawianym etapie musimy więc z góry ustalić „g" środków ciężkości (etap 5b), czyli „g" wektorów, każdy o „n" liczbie współrzędnych (n jest liczbą cech poprzez które badamy podobieństwo obiektów).

Mając zdefiniowane pierwsze środki ciężkości, w kolejnym kroku (5c) wyliczane są odległości taksonomiczne punktów reprezentujących grupowane obiekty (przyjęliśmy, że ich liczbę oznaczamy przez „m") od każdego środka ciężkości. W efekcie otrzymujemy macierz odległości o wymiarach g x m. Korzystamy ze znanego już wzoru na odległość euklidesową[94] (wzór 19), z wykorzystaniem oczywiście oznaczeń przyjętych powyżej:

$$d_{is} = \sqrt{\sum_{j=1}^{n}(t_{ij} - t_{sj})^2} \quad \text{gdzie:}$$

d_{is} jest odległością i-tego punktu od s-tego środka ciężkości „s"; s=1, 2,…, g

t_{ij}; t_{js} – standaryzowane wartości cechy „j", punktu „i" oraz środka ciężkości „s".

Kolejnym krokiem (5d) jest przydzielenie każdego punktu (reprezentującego odpowiedni obiekt) do właściwej grupy. „Właściwą" grupą jest ta, której środek ciężkości jest najbliżej położony danego punktu. Efektem tej fazy jest podział całego zbioru obiektów na „g" grup. Pamiętamy jednak, że środki ciężkości, wokół których w tym etapie „zgromadzone" zostały punkty wyznaczone zostały arbitralnie. Należy więc dla każdej grupy wyznaczyć nowe ich środki ciężkości, tym razem wyliczając je na podstawie współrzędnych danej grupy punktów; jest to następny etap postępowania (etap 6 przedstawionego algorytmu), opisany poniżej.

Ad. 6) Istotą metod taksonomicznych jest podział (grupowanie) niejednorodnego zbioru obiektów w podzbiory cechujące się jak największym podobieństwem. Podobieństwo z kolei opisywane jest tzw. odległością taksonomiczną. Było to już przedmiotem szczegółowego omawiania. Jeżeli więc weźmiemy pod uwagę jakąś grupę obiektów, z których każdy reprezentowany jest przez punkt o współrzędnych (t_1, t_2, t_3,…, t_n), gdzie n oznacza liczę cech, wówczas możemy przyjąć, że środkiem ciężkości (centroidem) tej grupy jest punkt, którego położenie względem wszystkich innych punktów grupy gwarantuje „jakąś optymalność". „Optymalność" ta może być na kilka sposobów definiowana. Najprostsze i najbardziej przemawiające do wyobraźni może być rozwiązanie, że środkiem ciężkości grupy obiektów jest taki punkt, który zapewnia minimum sumy odległości od niego wszystkich pozostałych punktów danej grupy. Załóżmy, że nasze zmienne, a więc cechy opisujące kryterium podobieństwa, były standaryzowane formułą *zero-jedynkową* (wzór 7). Możemy przyjąć, że współrzędne środka ciężkości są średnią arytmetyczną współrzędnych punktów danego skupienia (s), czyli:

[94] Podobnie jak to wyjaśnialiśmy w innej części podręcznika, możliwe jest korzystanie z innych formuł ustalania odległości taksonomicznych. Muszą jedynie spełniać warunki opisane relacjami (15-18).

$$t_{sj} = \frac{\sum_{i=1}^{k_s} t_{ij}^{(s)}}{k_s} \qquad (27)$$ gdzie:

t_{sj} – standaryzowana wartość j-cechy (a więc j-ta współrzędna) punktu ciężkości „s”
k_s – liczba obiektów (punktów) w grupie „s”, dla której wyznaczany jest środek ciężkości.

Oznaczmy przez A_s sumę odległości wszystkich punktów należących do grupy „s” od ich środka. Możemy wtedy zapisać wzorem:

$$A_s = \sqrt{\sum_i^{k_s} (\sum_{j=1}^{n} (t_{sj} - t_{ij}^{(s)})^2)} \qquad (28)$$

Po wyliczeniu środków ciężkości wzorem 27 (krok 6a algorytmu), dwoma następnymi etapami są: wyliczenie taksonomicznych odległości wszystkich punktów od środków ciężkości (etap 6b) oraz w ślad za otrzymanymi wynikami, przyporządkowanie punktów do właściwych środków ciężkości (6c), a tym samym otrzymanie nowego, skorygowanego podziału na „g” grup obiektów.

Etapy 6a do 6c powtarzamy aż *do momentu spełnienie zadanego kryterium*. Kryterium tym może być z góry zadana liczba iteracji (powtórzeń, czyli powrotów do etapu 6a) lub też otrzymanie w danej iteracji wyników (czyli przyporządkowania obiektów do danych skupień) identycznych z iteracją poprzednią.

Całość procedury grupowania ilustruje schemat ujęty rysunkiem 29.

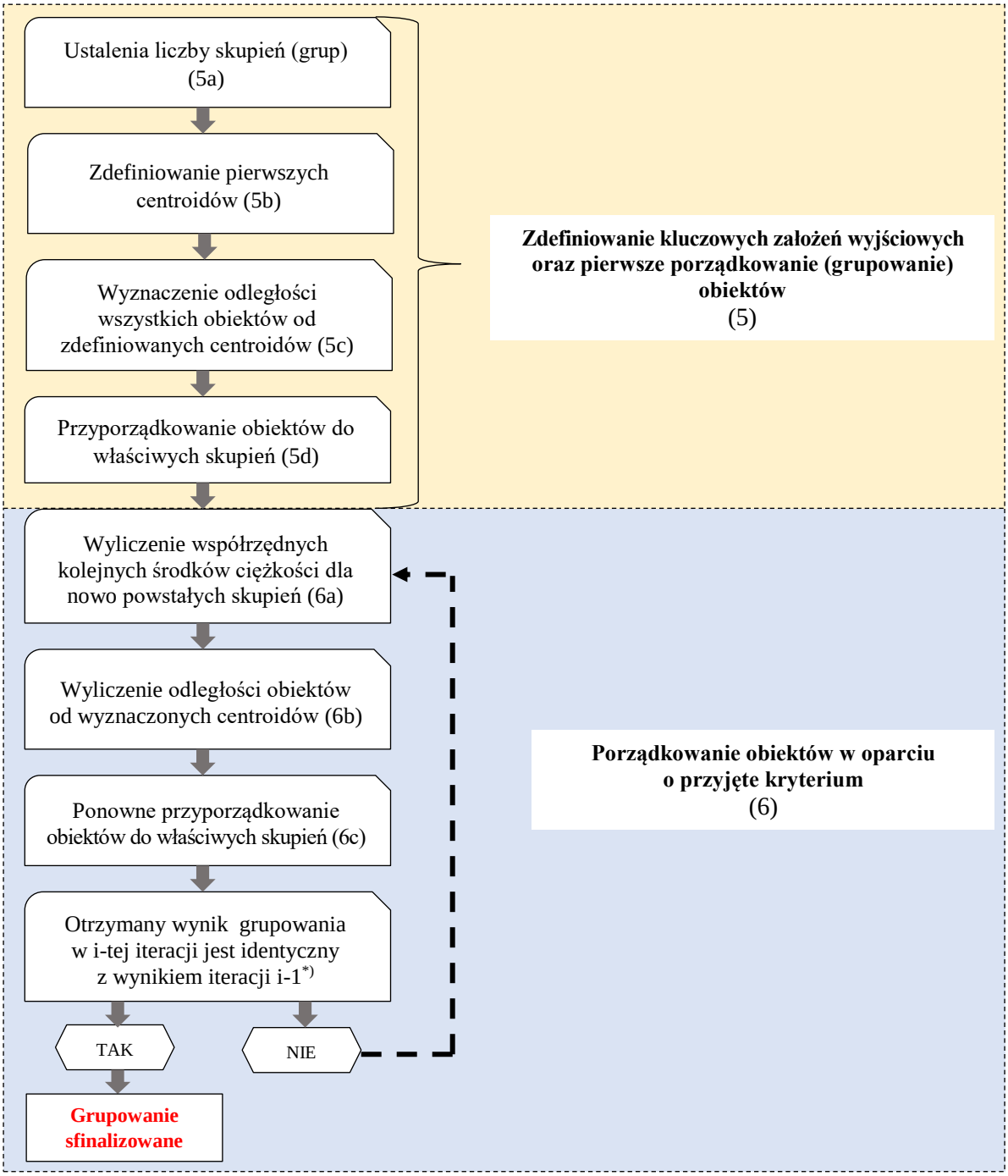
Dla celów porównawczych w tabelach 33 i 34 zestawiono wyniki grupowania odnoszące się do naszego przykładu, otrzymane na podstawie metody aglomeracyjnej Warda oraz metody k-średnich. Jak można zauważyć, wyniki grupowania w oparciu o te dwie metody są w pełni zbieżne.

W podsumowaniu rozważań dotyczących metody k-średnich warto zwrócić uwagę na następujące kwestie:

1. Metoda ta posiada szereg zalet w stosunku do metod wcześniej omawianych. W procedurze jej stosowania nie pojawiają się założenia wymuszające daleko idące uproszczenie we wnioskowaniu o podobieństwie grupowanych obiektów. Ma jasno zdefiniowane kryterium oceny jakości dokonywanego grupowania. Są to ważne walory przesądzające o częstym jej wykorzystywaniu w praktyce. Omawiana metoda ma jednak również istotne słabe strony. Dotyczą one głównie konieczności z góry przyjmowanej deklaracji o liczbie wydzielanych grup, a także zdefiniowania pierwszych środków ciężkości – była o tym mowa powyżej.
2. Słabe strony metody k-średnich przesądzają, że w praktyce bardzo często metoda k-średnich stosowana jest w duecie z metodą aglomeracyjną Warda. Wyniki tej drugiej traktowane są jako przegląd/podpowiedź następstw przyjęcia określonej liczby grup.
3. Kluczową kategorią w metodzie k-średnich jest centroid (środek ciężkości). Należy zauważyć, że badanie podobieństwa porządkowanych obiektów do odpowiednich centroidów, zamiast bezpośredniego podobieństwa pomiędzy obiektami (jak to było w przypadku wcześniej omawianych metod), jest jednak pewnym uproszczeniem.

Pytania/zadania kontrolne

1. Dokonaj zestawienia porównawczego mocnych i słabych stron taksonomicznych metod aglomeracyjnych i metody k-średnich.
2. Czy metoda k-średnich może prowadzić do wydzielania jednoelementowych grup obiektów?
3. Jaką ważną informację/podpowiedź dla metody k-średnich zawierają wyniki grupowania za pomocą metod aglomeracyjnych?



Rysunek 29. Metoda k-średnich – schemat procedury postępowania.
Źródło: opracowanie własne.
Kryterium to może również brzmieć: „*Czy wykonywana iteracja jest ostatnią z założonych*?”.

Tabela 33

Wyniki grupowania województw pod względem osiągniętego poziomu rozwoju (podział na trzy grupy)

Aglomeracyjna Warda	K-średnich
Kujawsko-pomorskie Lubelskie Łódzkie Opolskie Podkarpackie Podlaskie Świętokrzyskie Wielkopolskie	Kujawsko-pomorskie Lubelskie Łódzkie Opolskie Podkarpackie Podlaskie Świętokrzyskie Wielkopolskie
Małopolskie Mazowieckie Śląskie	Małopolskie Mazowieckie Śląskie
Dolnośląskie Lubuskie Pomorskie Warmińsko-mazurskie Zachodniopomorskie	Dolnośląskie Lubuskie Pomorskie Warmińsko-mazurskie Zachodniopomorskie

Źródło: opracowanie własne.

Tabela 34

Wyniki grupowania województw pod względem osiągniętego poziomu rozwoju (podział na cztery grupy)

Aglomeracyjna Warda	K-średnich
Opolskie Wielkopolskie	Opolskie Wielkopolskie
Kujawsko-pomorskie Lubelskie Łódzkie Podkarpackie Podlaskie Świętokrzyskie	Kujawsko-pomorskie Lubelskie Łódzkie Podkarpackie Podlaskie Świętokrzyskie
Małopolskie Mazowieckie Śląskie	Małopolskie Mazowieckie Śląskie
Dolnośląskie Lubuskie Pomorskie Warmińsko-mazurskie Zachodniopomorskie	Dolnośląskie Lubuskie Pomorskie Warmińsko-mazurskie Zachodniopomorskie

Źródło: opracowanie własne.

Bibliografia

Apanowicz, J. (2002). *Metodologia ogólna*. Gdynia: Wydawnictwo Bernardinum.

Babbie, E. (2005). *Badania społeczne w praktyce*. Warszawa: Wydawnictwo Naukowe PWN.

Błażejczyk-Majka, L. (2018). *Zastosowanie wybranych metod taksonomicznych w badaniach historycznych*. Poznań : Uniwersytet im. Adama Mickiewicza w Poznaniu.

Chojnicki, Z., Czyż, T. (1973). *Metody taksonomii numerycznej w regionalizacji geograficznej*. Warszawa: PWN.

Grabiński, T. (1984). *Wielowymiarowa analiza porównawcza w badaniach dynamiki zjawisk ekonomicznych*. Kraków. Zeszyty Naukowe AE, Seria specjalna: Monografie, nr 61.

Grabiński, T., Wydymus, S., Zeliaś, A. (1993). *Metody prognozowania rozwoju społeczno--gospodarczego*. Kraków: AE w Krakowie.

Greń, J. (1982). *Statystyka Matematyczna. Modele i zadania*. Warszawa. Państwowe Wydawnictwo Naukowe.

Krajewski, M. (2010). *O metodologii nauk i zasadach pisarstwa naukowego. Uwagi podstawowe*. Gliwice: Uniwersytet Śląski.

Kudłacz, T. (1992*). Modelowanie rozwoju społeczno-gospodarczego w układach regionalnych. Pomocnicze materiały dydaktyczne*. Kraków: AE w Krakowie.

Kukuła, K. (2000). *Metoda unitaryzacji zerowanej*. Warszawa: Wydawnictwo Naukowe PWN.

Łobocki, M. (2000). *Metody i techniki badań pedagogicznych*. Kraków: Oficyna Wydawnicza „Impuls".

Malchar, J., Zielińska-Sitkiewicz, M. (2017). Metody klasyfikacji w analizie porównawczej rozwoju społeczno-gospodarczego polskich województw w latach 2010 i 2014 – wpływ procedury normalizacji na wynik ranking. *Metody Ilościowe w Badaniach Ekonomicznych, XVIII*/4, 643-652.

Młodak, A. (2006). *Analiza taksonomiczna w statystyce regionalnej*. Warszawa: Difin.

Parysek, J., Wojtasiewicz, L. (1979). *Metody analizy regionalnej i planowania regionalnego*. Warszawa: PWE.

Poradnik kwalifikowania zadań w projektach B+R o charakterze społecznym i ekonomicznym. (2018). Narodowe Centrum Badań i Rozwoju. Politechnika Warszawska.

Sagan, A. (2016). *Metodologia badań ekonomicznych*. Kraków: Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie.

Strahl, D. (1990). *Metody programowania rozwoju społeczno-gospodarczego*. Warszawa: PWE.

Strahl, D. (2006). *Metody oceny rozwoju regionalnego*. Wrocław: Wydawnictwo Akademii Ekonomicznej we Wrocławiu.

Suchecki, B. (red.). (2010). *Ekonometria przestrzenna. Metody i modele analizy danych przestrzennych*. Warszawa: Wydawnictwo C.H. Beck.

Wiśniewski, E. (1983). *Metodyka wojskowych badań naukowych*. Warszawa: ASG WP.

Spis tabel

Spis rysunków