Aleksandra Derda

e-mail: 177427@student.ue.wroc.pl

ORCID: 0009-0003-6022-2238

Uniwersytet Ekonomiczny we Wrocławiu

Wykorzystanie metody *k*-średnich w statystyce publicznej

DOI: 10.15611/2023.89.3.01
JEL Classification: C38, R1, R5

Streszczenie: Głównym celem pracy jest przedstawienie metody *k*-średnich oraz warunków, jakie powinny zostać spełnione, by uzyskać najlepsze grupowanie gmin w województwie dolnośląskim ze względu na ich sytuację finansową. Przeprowadzona analiza ukazuje zakres wykorzystania wspomnianej metody w statystyce publicznej. Dane dotyczące sytuacji finansowej gmin zaczerpnięto z Banku Danych Lokalnych prowadzonego przez Główny Urząd Statystyczny. W analizie wykorzystano oprogramowanie Statistica, Excel oraz język R z wykorzystaniem pakietu Factoextra. Dokonano dwóch różnych podziałów przy użyciu innych metod doboru początkowych środków ciężkości. Ostateczna klasyfikacja nie została uznana za dobrą, co uwypukliło wady metody.

Słowa kluczowe: metoda *k*-średnich, sytuacja finansowa gmin, statystyka publiczna

1. Wstęp

Statystyka publiczna to „system zbierania danych statystycznych, gromadzenia, przechowywania i opracowywania zebranych danych oraz ogłaszania, udostępniania i rozpowszechniania wyników badań statystycznych jako oficjalnych danych statystycznych" (Ustawa z dnia 29 czerwca 1995 r. o statystyce publicznej).

Nad działaniem tego procesu czuwają służby statystyki publicznej. W Polsce są to pracownicy instytucji podlegających Prezesowi Głównego Urzędu Statystycznego (GUS). Dodatkowo w realizacji zadań Prezes GUS wspierany jest przez Centrum Informatyki Statystycznej (CIS), Zakład Wydawnictw Statystycznych (ZWS), Centrum Badań i Edukacji Statystycznej (CBiES) oraz Centralną Bibliotekę Statystyczną (CBS) im. Stefana Szulca (GUS, b.d.a).

Dane zbierane są przede wszystkim po to, by jednostki państwowe czy samorządowe podejmowały na ich postawie decyzje. Jednak nie są to jedyne jednostki uprzywilejowane do korzystania z nich. Dane te są jawne, dlatego każdy z obywateli i przedsiębiorców również może je wykorzystać do podjęcia decyzji.

Zgodnie z Ustawą z dnia 29 czerwca 1995 r. o statystyce publicznej służby statystyki mogą zbierać dane z następujących źródeł:

- rejestrów urzędowych,
- systemów informacyjnych administracji publicznej,
- niepublicznych systemów informacyjnych,

- odpowiedzi respondentów,
- innych powszechnie dostępnych źródeł.

Oznacza to, że służby statystyki publicznej mogą przeprowadzać własne badania, ale i skorzystać z wyników badań czy baz danych tworzonych przez inne instytucje. Niezależnie jednak od tego, skąd pochodzą dane, wszystkie informacje podlegają tajemnicy statystycznej. Zgodnie z art. 10: „dane jednostkowe identyfikowalne zebrane w badaniach statystycznych podlegają bezwzględnej ochronie. Dane te mogą być wykorzystywane wyłącznie do opracowań, zestawień i analiz statystycznych oraz do tworzenia przez Prezesa GUS operatu do badań statystycznych; udostępnianie lub wykorzystywanie tych danych dla innych niż podane w ustawie celów jest zabronione (tajemnica statystyczna)". Oznacza to, że dane jednostkowe nie mogą być udostępniane, a tylko zbiorcze opracowania dla wszystkich obiektów badania mogą zostać upublicznione.

Wyniki uzyskane po opracowaniu danych publicznych udostępniane są zgodnie z zasadą 3R, co zostało przedstawione na rysunku 1. Niezależnie od tego, jakie byłyby to badanie, powinny spełnić zasady równoprawności, równorzędności i równoczesności. Oznacza to taką samą ważność wszystkich jednostek niezależnie od zajmowanych przez nie pozycji i zapewnia powszechność uzyskanych wyników.

Rysunek 1. Zasada 3R

Źródło: (GUS, b.d.).

Szczególnym badaniem w statystyce publicznej jest Narodowy Spis Ludności. Zgodnie z encyklopedią spis ludności to podstawowe badanie i źródło danych z zakresu statystyki ludności, które ma na celu zebranie informacji o jej stanie i strukturze według ustalonych cech demograficznych i społeczno-zawodowych, w oznaczonym momencie, na określonym terytorium (*Encyklopedia PWN*, b.d.). By umożliwić

porównywanie wyników między państwami, spisy takie są przeprowadzane zgodnie z prawem unijnym, a ściślej – zgodnie z art. 4 rozporządzenia (WE) Parlamentu Europejskiego i Rady Nr 763/2008 z dnia 9 lipca 2008 r. w sprawie spisów powszechnych ludności i mieszkań.

W akcie prawnym wskazane są warunki, które muszą zostać spełnione podczas organizacji takiego spisu. Należą do nich: warunek powszechności, jednoczesności i imienności wynikające z prawa unijnego, a także periodyczności oraz bezpośredniości wynikające z prawa polskiego. Warunek powszechności oznacza, że spis powinien obejmować wszystkich mieszkańców państwa; warunek jednoczesności jest spełniony, gdy spis odbywa się w określonym czasie, a zebrane dane będą dotyczyły konkretnego momentu w czasie. Warunek imienności informuje o tym, że każda z osób musi być spisana imiennie. Warunek periodyczności, stosowany w Polsce, oznacza, że spis jest organizowany w konkretnych odstępach czasu. W Polsce zgodnie z zaleceniami ONZ spis odbywa się co 10 lat. Z kolei warunek bezpośredniości wiąże się z odpowiedziami na pytania, których powinna udzielać bezpośrednio osoba spisywana; jedynie w szczególnych przypadkach dopuszcza się, by w imieniu takiej osoby odpowiadali najbliżsi.

Dane pochodzące ze spisu są szczególnie ważne na arenie zarówno międzynarodowej, jak i wewnątrzkrajowej. Na podstawie zebranych danych przyznawane są dotacje unijne czy liczba miejsc w Parlamencie Europejskim, gdzie szczególne znaczenie ma przede wszystkim liczba osób zamieszkujących dane państwo. Badania te często są jedynym źródłem informacji na temat niektórych tematów, głównie dotyczących cech demograficzno-społecznych, takich jak wyznanie czy narodowość. Dane te służą również jako baza do losowania reprezentacyjnych prób w badaniach społecznych, a także na ich podstawie weryfikowane są coroczne badania bilansowe. Tak jak pozostałe wyniki badań przeprowadzanych przez służby statystyki publicznej, dane spisowe są wykorzystywane przez jednostki zarządzające w procesie podejmowania decyzji gospodarczych czy społecznych.

Celem prezentowanego badania jest wskazanie użyteczności wybranych sposobów analizy i wnioskowania na podstawie zaprezentowanych danych. Jest to element niezmiernie istotny, gdyż na tej podstawie można podejmować decyzje na szczeblu gminnym, regionalnym czy krajowym. Wyniki są również użyteczne w procesie zarządzania przedsiębiorstwami, gdyż uwzględniając sytuację społeczno-ekonomiczną gmin, możliwe jest podejmowanie działań lepiej dostosowanych do poszczególnych regionów.

Wykorzystując wiedzę teoretyczną na temat wskazanych metod grupowania, czyli przeprowadzając analizę sytuacji finansowej gmin w województwie dolnośląskim w 2018 roku, sprawdzano poprawność kilku hipotez badawczych:

- Czy występuje tendencja łączenia się obiektów w jedno liczne skupienie?
- Czy gminy miejskie powinny trafić do jednej grupy?
- Czy gmina Wrocław powinna być traktowana jak obserwacja odstająca?

2. Podstawowe zagadnienia wielowymiarowej analizy statystycznej

Obok wiarygodnego źródła danych równie ważne jest, by wybrane zmienne spełniały założenia zastosowanych metod, dlatego niezbędne jest poznanie zarówno wymagań poszczególnych metod, jak i sposobów sprowadzania zmiennych do wymaganej formy.

Statystyczna analiza wielowymiarowa to „grupa metod statystycznych, za pomocą których jednoczesnej analizie poddane są pomiary na przynajmniej dwóch zmiennych, opisujących każdy obiekt badania" (Gatnar i Walesiak, 2004, s. 17), w związku z czym najważniejszymi pojęciami dla tych metod są ,,zmienna" i ,,obiekt".

Zmienna opisuje zbiorowość obiektów. Za jej pomocą mierzy się opisywane zagadnienia (Jajuga, 1987). Zmienną można przyrównać do cechy, której wartość czy charakter są przypisane indywidualnie do każdego z badanych obiektów. W analizach statystycznych stosuje się różne rodzaje zmiennych. Głównym podziałem jest ten pod kątem sposobu opisywania zjawiska. Pod tym względem występują dwa rodzaje: zmienne metryczne (których wartości przedstawia się za pomocą symboli, inaczej nazywane kwalitatywnymi) oraz zmienne niemetryczne (których wartości można przedstawić za pomocą liczb będących jej realizacjami, inaczej nazywane kwantytatywnymi).

Obiekt to natomiast najmniejszy element poddany obserwacji, na podstawie której można uzyskać informacje niezbędne do przeprowadzenia analizy (Steczkowski i Zeliaś, 1981). Obserwacje zazwyczaj przeprowadza się na obiektach należących do grupy określonej na podstawie pewnych wspólnych cech. Obiektami badania mogą być jednostki administracyjne, osoby odpowiadające na zadane pytania czy osobniki należące do tego samego gatunku. Za pomocą wielowymiarowych metod analizy statystycznej możliwe jest badanie obiektów ze względu na kilka różnych cech jednocześnie. Sprawia to, że metody te są często stosowane w wielu dziedzinach nauki.

By rzetelnie opisać badane zjawisko, niezbędne jest prawidłowe wyodrębnienie i zmierzenie zmiennych reprezentujących możliwie najwięcej czynników, które wpływają na zmienność i kształtowanie się tego zjawiska. O ile wyodrębnienie zmiennych często wymaga doświadczenia i intuicji, o tyle zmierzenie ich wydaje się procesem łatwiejszym do przeprowadzenia. Do tego używa się skal pomiarowych. Występują cztery główne rodzaje skal: nominalna i porządkowa dla danych niemetrycznych, a także przedziałowa i ilorazowa dla danych metrycznych. Różnią się one przede wszystkim szczegółowością prezentacji informacji. Wartości uzyskane na silniejszych skalach można przekształcać na pomiary na słabszych. Co więcej, wszystkie metody, które można stosować na skalach słabszych, są również dozwolone dla pomiarów z silniejszych skal. Skala, na której zmierzono zmienne, determinuje, jakie metody analiz można stosować na tych danych. Dlatego kluczowe jest określenie, czy zmienne, które zostaną zastosowane do dalszych analiz, zostały przedstawione

na skali zgodnej z założeniami metody. Szczegółową informację o skalach pomiarowych można znaleźć w (Adams i in., 1965; Sobczak, 2006).

Należy zwrócić uwagę, że w trakcie prowadzenia analizy metodą *k*-średnich zmienne muszą być metryczne.

Charakter zmiennej informuje, jakie wartości są w jej przypadku pożądane. Stymulanta to cecha, której wartość maksymalna uznawana jest za najbardziej, a minimalna – za najmniej korzystną dla badanych obiektów (Walesiak, 1990). Taką zmienną może być na przykład zysk z punktu widzenia firmy podsumowującej rok bilansowy. Kolejną grupą są destymulanty, u których preferowane są jak najniższe wartości. Przykładem może być liczba błędów w wydawanej książce. Występują również nominanty, czyli zmienne, „które mają charakter stymulanty do pewnego punktu, zwanego wartością nominalną, a później przyjmują charakter destymulanty. Przyjęcie przez nominantę wartości większych lub mniejszych od wartości nominalnej oznacza spadek oceny. Inaczej mówiąc, jest to zmienna diagnostyczna, dla której najbardziej pożądane są wartości przeciętne" (Serafin i Luściński, b.d., s. 1012). Taką zmienną jest na przykład masa ciała przeciętnego niemowlaka, której zarówno za wysokie, jak i za niskie wartości nie są pożądane. Występują również zmienne neutralne, czyli obojętne dla badanego zjawiska (Gatnar i Walesiak, 2004).

Niektóre z metod wielowymiarowej analizy statystycznej wymagają, by wszystkie ze zmiennych miały określony charakter, najczęściej stymulanty. Sposoby ujednolicania charakteru zmiennych szczegółowo opisują Gatnar i Walesiak (2004). W analizie wykorzystano jedynie dane o charakterze stymulant i destymulant, które nie wymagały ujednolicania.

Niektóre z metod wielowymiarowej analizy statystycznej, takie jak metody klasyfikacji, skalowania wielowymiarowego czy porządkowania liniowego, wymagają, by zmienne były porównywalne, czyli pozbawione mian z ujednoliconymi rzędami wielkości. Dlatego w takich przypadkach przeprowadza się proces normalizacji. Wiele przykładów normalizacji zmiennych prezentuje Walesiak (2004). W analizie wykorzystano standaryzację obliczaną zgodnie ze wzorem:

$$\frac{\left(x_{ij} - \overline{x}_j\right)}{s_j},$$

gdzie: x_{ij} – rzeczywista wartość *j*-tej zmiennej dla *i*-tego obiektu, $\overline{x}_j$ – średnia arytmetyczna *j*-tej zmiennej, s_j – odchylenie standardowe *j*-tej zmiennej.

Po wykorzystaniu tej formuły średnia arytmetyczna wynosi 0, a odchylenie standardowe 1 dla każdej ze zmiennych.

Każda z wielowymiarowych metod statystycznych stosowana jest w różnych celach. Najczęściej to, jaką metodę należy wybrać do analiz, determinowane jest przez cel badania i dane, jakimi się dysponuje. Za pomocą metody *k*-średnich obiekty łączy się w grupy. Stosowana jest ona w celu zbadania współwystępowania, po jej wykorzystaniu nie otrzymuje się informacji na temat zależności między obiektami

i zmiennymi. Dodatkowo metoda ta należy do grupy metod eksploracyjnych, czyli używana jest do wykrywania wniosków z danych. Metodę *k*-średnich można stosować jedynie na danych metrycznych. Zmienne brane pod uwagę w badaniu są traktowane równoważnie, nie występuje podział na zmienne endo- i egzogeniczne.

3. Wstęp do klasyfikacji

Klasyfikacja to „zbiór klas odpowiednio wyróżnionych z klasyfikowanego zbioru obiektów" (Gatnar i Walesiak, 2004, s. 317). Na podstawie tej definicji można wywnioskować, że metody klasyfikacji to takie, po zastosowaniu których obiekty są podzielone na grupy ze względu na pewne cechy odpowiedające w tym przypadku określonym zmiennym. Metody te służą jedynie do wykrywania struktur danych, a nie znajdowania przyczyn ich występowania (StatSoft, b.d.).

Uzyskane zbiory obserwacji określane są klasami, skupieniami, klastrami, typami czy grupami. Wyodrębnione klasy powinny spełniać dwa kryteria: wewnętrznej spójności i zewnętrznej izolacji (Gatnar i Walesiak, 2004, s. 318), czyli obiekty należące do danego skupienia byłyby do siebie jak najbardziej podobne ze względu na badane cechy, natomiast jak najbardziej różne od pozostałych obiektów należących do innych grup.

Występuje wiele podziałów metod klasyfikacji. Tym, który najbardziej odzwierciedla temat tego badania, jest podział przedstawiony przez Zalewską (2017):

- metody hierarchiczne, w których elementy łączy się w podgrupy, rozpoczynając od skupień z jednym elementem, kończąc na jednym klastrze, do którego należą wszystkie obiekty;
- metody niehierarchiczne, w których podziału dokonuje się na z góry określoną liczbę klas;
- metody rozmytej analizy skupień, które pozwalają na przydział do kilku klas, uwzględniając prawdopodobieństwo przynależności.

4. Metoda *k*-średnich

Jest to jedna z najczęściej stosowanych niehierarchicznych metod grupowania. Występuje wiele różnych algorytmów jej przeprowadzenia, ale najpowszechniej stosowany jest jednak ten zaproponowany przez Hartigana i Wonga (1979). Celem tej analizy jest podział zbioru obiektów na z góry narzuconą liczbę klas (oznaczaną jako *k*), takich, dla których wewnętrzne zmienności będą jak najmniejsze. Dąży się, by całkowita zmienność wewnątrzgrupowa, obliczana z poniższego wzoru, była jak najmniejsza (Kassambara, 2017):

$$tot.withinss = \sum_{k=1}^{k} \sum_{x_j \in c_k} \left(x_i - \mu_k \right)^2,$$

gdzie: *tot.withinss* – całkowita zmienność wewnątrzgrupowa, x_i – wartość zmiennej *X* dla obiektu należącego do skupienia C_k, μ_k – średnia wartość zmiennych dla obiektów należących do skupienia C_k, k – liczba skupień.

Algorytm metody *k*-średnich rozpoczyna się od wyznaczenia liczby klas. Istnieje wiele sposobów, na podstawie których można oszacować optymalną liczbę skupień. Są to na przykład metoda elbow, indeks Calińskiego i Harabasza (Caliński i Harabasz, 1974; Walesiak, 2004), indeks Huberta i Levine'a (Hubert i Levine, 1976; Walesiak, 2004) czy decyzja oparta na doświadczeniu badacza.

Kolejnym krokiem algorytmu jest wybór środków ciężkości, na podstawie których w następnych krokach tworzone będą skupienia. W tym przypadku także występują różne metody do ich wyboru. Wymienione zostały one w pracy Kajsturego (2013):

- wybór *k* pierwszych obserwacji;
- losowy wybór *k* obiektów, które zostają środkami ciężkości;
- uwzględnienie początkowego położenia punktów, wybrane mogą zostać na przykład obiekty położone najdalej od siebie;
- przyjęcie środków skupień uzyskanych przy wcześniejszych badaniach, często za pomocą metod hierarchicznych.

W kolejnym kroku algorytmu następuje przydzielenie każdego z obiektów do najbliższego mu klastra na podstawie odległości między nimi a środkami ciężkości poszczególnych skupień.

Przez niedokładny dobór początkowych środków ciężkości podział na klasy powstały w poprzednim kroku może nie być optymalny. Z tego powodu przeprowadza się kolejny etap algorytmu – ustalenie nowych środków skupień. Najczęściej są to punkty o współrzędnych równych wartościom średnim dla wszystkich obiektów danego skupienia. Znając nowe środki ciężkości, ponownie wykonuje się krok trzeci algorytmu, czyli obiekty są ponownie przypisywane do grup. Kroki trzeci oraz czwarty powtarza się do momentu, gdy kolejne powtórzenia, zwane iteracjami, nie powodują zmiany przydziału obiektów do grup lub gdy wykonana została założona wcześniej liczba iteracji. Wartość ta jest subiektywnie wybierana przez analityka.

Główną zaletą tego algorytmu jest niska złożoność, szczególnie ważna przy dużych zbiorach danych, natomiast do najważniejszych wad metody zalicza się brak jednoznacznego sposobu wybrania optymalnej liczby skupień oraz losowość wyboru początkowych środków ciężkości. Dużym ograniczeniem jest również fakt, że algorytm ten jest odpowiedni dla skupisk o sferycznym kształcie i jednorodnej gęstości.

Istnieje jednak sposób, by uniknąć nieprawidłowego doboru początkowych skupień. Jest to tak zwana metoda *k*-medoidów, w której początkowe środki skupień wybierane są w taki sposób, by zostały nimi te punkty, których suma różnic w stosunku do wszystkich obiektów w klastrze jest najmniejsza. Dzięki takiemu doborowi metoda *k*-medoidów jest bardziej odporna na wartości odstające, niestety jest ona jednocześnie mniej wydajna obliczeniowo, co może być szczególnie niekorzystne przy analizie dużych zbiorów danych.

By klasyfikację można było uznać za prawidłową, sprawdzone powinny zostać prawidłowość zaklasyfikowania poszczególnych obiektów do klas, prawidłowość

wyodrębniania poszczególnych klas oraz ogólna jakość klasyfikacji, rozumiana jako relatywna zwartość i separowalność klas (Walesiak, 2004). Wskaźnik, za pomocą którego możliwe jest sprawdzenie wszystkich wymienionych cech, podsunął Rousseeuw (Rousseeuw, 1987; Kaufman i Rousseeuw, 1990). Wyróżniony przez badacza wskaźnik, nazywany również *silhouette index* – wskaźnikiem sylwetkowym, służy do sprawdzenia prawidłowości zaklasyfikowania poszczególnych obiektów do klas.

Wyniki uzyskane na podstawie tego wzoru mogą zostać poddane interpretacji. Subiektywną ocenę przedziałów wartości indeksu $S(P)$ przedstawił Walesiak (2004). Zgodnie z opinią twórcy strukturę można uznać za dopuszczalną, gdy współczynnik $S(P)$ wyniesie powyżej 0,5 (wtedy występuje poważna struktura klas), jednak dopiero gdy indeks ten przyjmuje wartości powyżej 0,7, możemy uznać podział za bardzo dobry, czyli że występuje silna struktura klas. Dla wyników poniżej 0,5 powinno się zastosować inną metodę.

5. Dane

Grupowaniu poddano gminy województwa dolnośląskiego ze względu na ich sytuację finansową w roku 2018. Na podstawie badań sytuacji finansowej w gminach innych województw przeprowadzonych przez Standarę (2017) oraz Wiśniewskiego (2011) wstępnie wybrano zmienne, które opisują badane zjawisko:

X_1 – wydatki ogółem na 1 mieszkańca,
X_2 – dochody ogółem na 1 mieszkańca,
X_3 – środki z Unii Europejskiej na finansowanie programów i projektów unijnych (2014-2018) w przeliczeniu na 1 mieszkańca,
X_4 – liczba ludności,
X_5 – wydatki majątkowe inwestycyjne w przeliczeniu na 1 mieszkańca,
X_6 – wydatki bieżące jednostek budżetowych ogółem w przeliczeniu na 1 mieszkańca,
X_7 – wydatki bieżące na zakup materiałów i usług w przeliczeniu na 1 mieszkańca,
X_8 – wydatki bieżące na wynagrodzenia w przeliczeniu na 1 mieszkańca,
X_9 – świadczenia na rzecz osób fizycznych w przeliczeniu na 1 mieszkańca,
X_{10} – dotacje ogółem w przeliczeniu na 1 mieszkańca,
X_{11} – dochody podatkowe (podatek od nieruchomości) w przeliczeniu na 1 mieszkańca,
X_{12} – udziały w podatkach stanowiących dochody budżetu państwa (podatek dochodowy od osób fizycznych) w przeliczeniu na 1 mieszkańca,
X_{13} – dochody z majątku w przeliczeniu na 1 mieszkańca,
X_{14} – zadłużenie w przeliczeniu na 1 mieszkańca.

Zmienne X_1-X_{13} to stymulanty, natomiast zmienną X_{14} uznano za destymulantę (Standar, 2017).

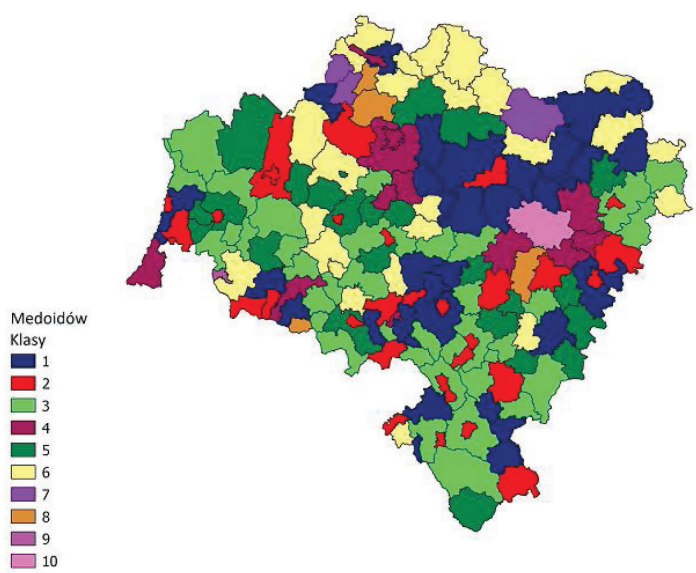
Dla każdej z wymienionych zmiennych obliczono współczynnik zmienności, by wykluczyć zmienne zachowujące się jak stałe lub *quasi*-stałe, czyli takie, które nie

wykazują zróżnicowania w obiektach. Żadnej ze zmiennych nie pominięto na tym etapie analizy.

Oprócz zmienności sprawdzona została również korelacja między zmiennymi. Do jej obliczenia zastosowano współczynnik korelacji Pearsona. Po przeanalizowaniu wartości dla wszystkich par zmiennych zdecydowano, że jako maksymalną dopuszczalną wartość współczynnika przyjęto 0,8, by uniknąć powtarzania informacji. Po przeanalizowaniu otrzymanych wartości zauważono, że występują zmienne zbyt mocno ze sobą skorelowane, dlatego powinno się zredukować ich liczbę. Z tego powodu ze zbioru zmiennych wykluczono zmienne X_2, X_5, X_6, X_7 oraz X_{11}. Pozostałe zmienne poddano standaryzacji.

6. Wyniki analizy metodą *k*-średnich

Dla gmin województwa dolnośląskiego, które zostały wybrane do analizy, ustalono wstępny podział na 10 klas. By uniknąć błędów wynikających z losowości doboru środków skupień, zdecydowano, że wstępnymi punktami będącymi podstawami tworzonych klas zostaną medoidy, czyli takie obiekty, dla których suma różnic w stosunku do wszystkich obiektów w klastrze jest najmniejsza. Obiektami tymi mogą być tylko gminy, które należą do początkowego zbioru danych. Metoda ta bywa nazywana również metodą *k*-medoidów. Otrzymane wyniki porównano z podziałem uzyskanym po losowym doborze początkowych środków skupień. W obu przypadkach jako metodę obliczania odległości zastosowano formułę euklidesową.



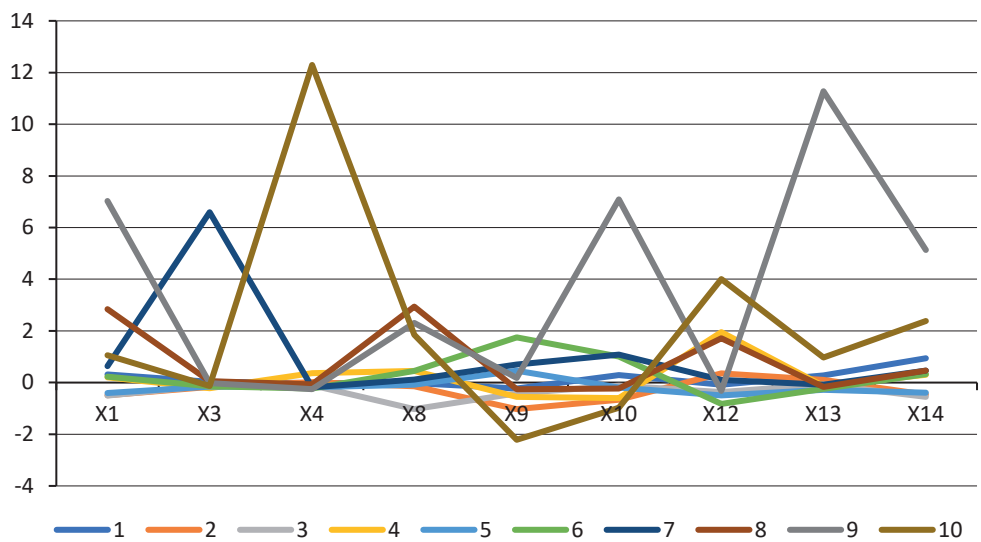
Rysunek 2. Przydział gmin do klas na mapie województwa dolnośląskiego zgodnie z metodą medoidów

Źródło: opracowanie własne.

Po przeprowadzeniu obliczeń zgodnie z algorytmem dla wstępnych środków skupień opartych na medoidach otrzymano podział przedstawiony na mapie na rysunku 2, na którym kolorami zaznaczono przynależność poszczególnych gmin do skupień.

Na podstawie mapy można zauważyć, że powstały klasy o podobnej liczebności, a także, że nie występuje skupienie, do którego trafiłaby większość gmin. Widać również zależność, że większość gmin miejskich trafiła do tej samej grupy oznaczonej numerem 2, i również delikatną zależność, że gminy w północnej części województwa trafiają do jednego skupienia, a gminy z południowej części do innego. Sąsiadujące ze sobą obiekty najczęściej wchodzą do tej samej klasy.

Na rysunku 3 przedstawiono średnie dla zmiennych w poszczególnych skupieniach. Standardowy przebieg wartości średnich w grupach zaobserwowano dla skupień, w których znalazły się tylko Świeradów-Zdrój i Wrocław. Trafiły one odpowiednio do klas 9 i 10.

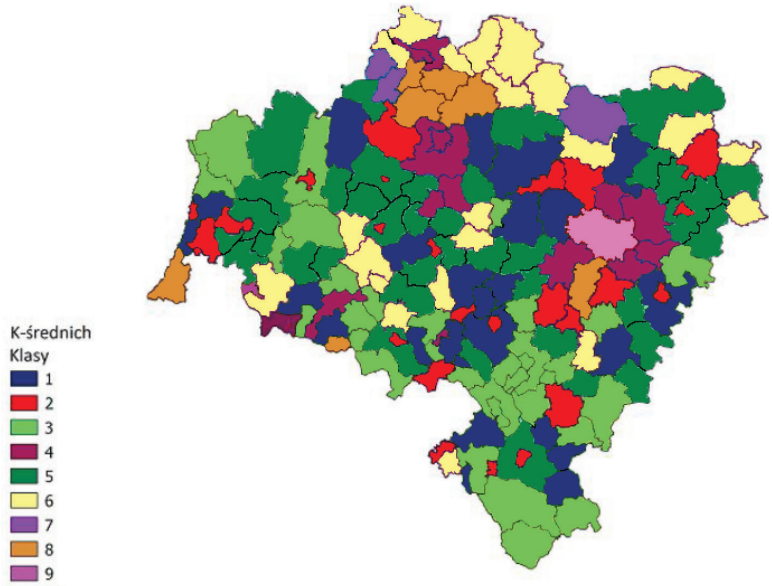


Rysunek 3. Wykres średnich dla podziału zgodnie z metodą medoidów

Źródło: opracowanie własne.

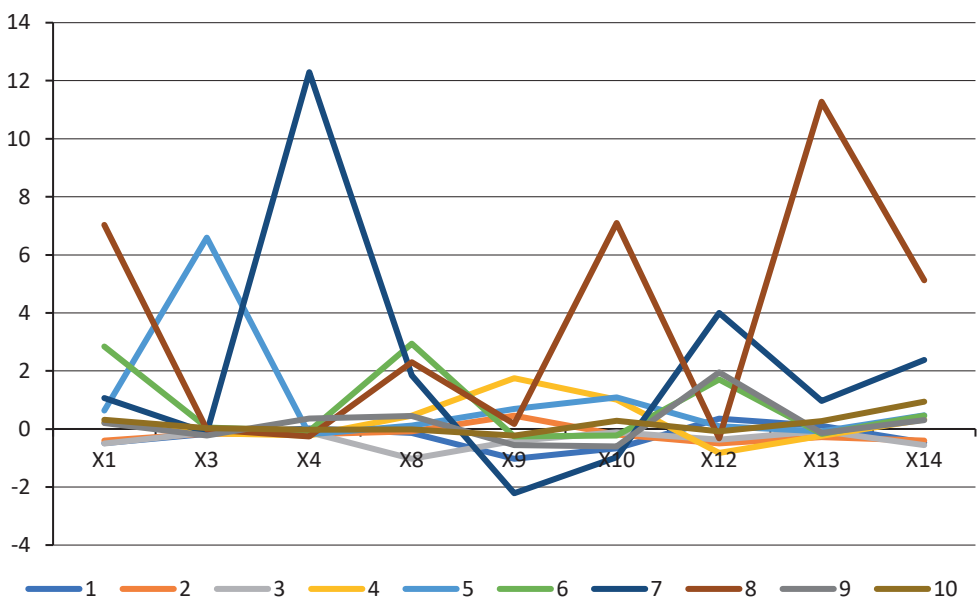
Na rysunku 4 zaznaczono przynależność gmin do klastrów po podziale wykonanym dla losowo wybranych początkowych środków skupień.

Analizując mapę (zob. rys. 4), można zauważyć, że podział jest podobny do tego przedstawionego na rysunku 3, ale nie jest identyczny. Ponownie występuje zależność, że sąsiadujące gminy mają tendencję do łączenia się w jedno skupienie. Do tej samej klasy trafiła również większość gmin miejskich. Ponownie osobne klastry tworzą dwie gminy: Wrocław i Świeradów-Zdrój. W tym podziale również powstały grupy o podobnej liczebności.



Rysunek 4. Przydział gmin do klas na mapie województwa dolnośląskiego zgodnie z losowym przydziałem środków skupień

Źródło: opracowanie własne.



Rysunek 5. Wykres średnich dla podziału zgodnie z losowym przydziałem środków skupień

Źródło: opracowanie własne.

Dla tego podziału również policzono średnie zestandaryzowanych wartości zmiennych dla każdej z klas. Wyniki przedstawiono na rysunku 5. Przebieg poszczególnych krzywych na wykresie jest zbliżony do tego, jaki uzyskano metodą medoidów (zob. rys. 3), chociaż podziały te nie są identyczne. Wskazuje to na fakt, że w każdej z tych metod szczególnie zostały wyodrębnione obiekty o nietypowych wartościach. Pozostałe klasy charakteryzują się średnimi o podobnych wartościach. Może to sugerować, że pomimo wskazań metod określających liczbę klas zmienne te zostały podzielone na zbyt dużo klas.

Do oceny jakości klasyfikacji zastosowano indeks Silhouette.

Tabela 1. Wartości indeksu ogólnej jakości klasyfikacji

	Początkowe środki skupień	
	Medoidy	Losowe
Jakość	0,155	0,142

Źródło: opracowanie własne.

W tabeli 1 przedstawiono wartości indeksu ogólnej jakości klasyfikacji dla podziałów wykonanych za pomocą metod niehierarchicznych.

Indeks ten przyjął bardzo niskie wartości, świadczące o bardzo złym podziale. W podziałach obiema metodami wystąpiły jedynie dwie klasy jednoelementowe, przez co wpływ wartości 0 uwzględnianych w obliczaniu indeksu ogólnej jakości klasyfikacji nie jest tak znaczący. Na podstawie tego można wywnioskować, że nie znaleziono prawidłowej struktury klas.

7. Zakończenie

Celem zaprezentowanego badania było przedstawienie algorytmu *k*-średnich. Po przeprowadzeniu badania szczególnie uwypuklone zostały jego wady. Jedną z najpoważniejszych jest niejednoznaczność przy wyborze liczby klas. Istnieje wiele metod, za pomocą których możliwe jest wyznaczenie sugerowanej liczby klas. Jednak, jak przedstawiono w tym badaniu, za pomocą różnych metod można uzyskać różną proponowaną liczbę klas. Dlatego ostateczna decyzja musi zostać podjęta przez badacza, jednak tak jak to wystąpiło w tej analizie, przez niejednoznaczne przesłanki może nie być ona optymalna. Istnieją jednak metody porównywania podziałów. Wymaga to jednak znajomości kolejnych metod lub programów.

Brak możliwości wizualizacji zbiorów danych nie pozwala również określić, czy wykonanie grupowania dla wybranego zbioru danych jest sensowne. Dla niektórych zbiorów, zwłaszcza takich, w których obiekty są bardzo zbliżone do siebie, grupowanie może okazać się zbędne.

Wadą metody *k*-średnich, która przez łatwość implementacji jest częściej stosowaną metodą grupowania, jest losowy wybór początkowych środków skupień. Etap ten przeprowadza się na początku algorytmu, dlatego przez błędny podział początkowy prawdziwa struktura danych może nie zostać wykryta.

Niestety przez uzyskanie nieprawidłowych podziałów niemożliwe jest wskazanie, czy potwierdzono hipotezy badawcze związane z sytuacją finansową gmin. By określić ich prawidłowość, należałoby ponownie przeprowadzić grupowanie, jednak przy wykorzystaniu innej metody.

Nie udało się uzyskać zadowalających wyników analizy, gdyż występowanie znacznych wartości odstających sprawiło, że tylko one były wychwytywane. W takim przypadku badanie powinno zostać przeprowadzone ponownie, jednak z wykorzystaniem metody grupowania odpornej na wpływ wartości odstających.

Literatura

Adams, E. W., Fagot, R. F. i Robinson, R. E. (1965). A Theory of Appropriate Statistics. *Psychometrika*, *30*(2), 99-127. https://doi.org/10.1007/BF02289443

Caliński, R. B. i Harabasz, J. A. (1974). Dendrite Method for Cluster Analysis. *Communications in Statistics*, (3), 1-27. http://dx.doi.org/10.1080/03610927408827101

Gatnar, E. i Walesiak, M. (2004). *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych.* Wydawnictwo Akademii Ekonomicznej we Wrocławiu.

Główny Urząd Statystyczny. (b.d.). *Co to jest statystyka publiczna?.* Pobrano 2 lutego 2022 z https://stat.gov.pl/portal-edukacyjny/co-to-jest-statystyka-publiczna/

Hartigan, J. A. i Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, *28*(1), 100-108. https://doi.org/10.2307/2346830

Hubert, L. J. i Levine, J. R. (1976). Evaluating Object Set Partitions: Free Sort Analysis and Some Generalizations. *Journal of Verbal Learning and Verbal Behaviour*, (15), 549-570.

Jajuga, K. (1987). Statystyka ekonomicznych zjawisk złożonych – wykrywanie i analiza niejednorodnych rozkładów wielowymiarowych. *Prace Naukowe Akademii Ekonomicznej we Wrocławiu*, *371*(39).

Kajstura, A. (2013). *Metoda k-średnich*. Pobrano 27 lutego 2022 z https://www.statystyka.az.pl/analiza-skupien/metoda-k-srednich.php

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R*. STHDA. Pobrano 26 lutego 2022 z https://www.datanovia.com/en/wp-content/uploads/dn-tutorials/book-preview/clustering_en_preview.pdf

Kaufman, L. i Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, (20), 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Rozporządzenie Parlamentu Europejskiego i Rady (WE) Nr 763/2008 z dnia 9 lipca 2008 r. w sprawie spisów powszechnych ludności i mieszkań

Serafin, R. i Luściński, S. (b.d.). *Normalizacja kryteriów oceny dostaw w systemach zaopatrzenia*. Pobrano 10 lutego 2022 z http://www.ptzp.org.pl/files/konferencje/kzz/artyk_pdf_2016/T1/t1_1010.pdf

Sobczak, E. (2006). Skale pomiaru w międzynarodowych badaniach marketingowych. *Prace Naukowe Akademii Ekonomicznej we Wrocławiu*, (1100), 97-109. Pobrano 10 lutego 2022 z https://depot.

ceon.pl/bitstream/handle/123456789/19162/2006_Sobczak_Skale_pomiaru_w_miedzynarodowych_badaniach.pdf?sequence=1

Spis ludności. (b.d.). *Encyklopedia PWN*. Pobrano 7 lutego 2022 z https://encyklopedia.pwn.pl/haslo/spis-ludnosci;3978274.html

Standar, A. (2017). Ocena kondycji finansowej gmin oraz jej wybranych uwarunkowań na przykładzie województwa wielkopolskiego przy wykorzystaniu metody TOPSIS. *Wieś i Rolnictwo*, *2*(175), 69-92. https://doi.org/10.7366/wir022017/04

StatSoft. (b.d.). *Analiza skupień*. Pobrano 24 lutego 2022 z https://www.statsoft.pl/textbook/stathome_stat.html?https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstcluan.html

Steczkowski, J. i Zeliaś, A. (1981). *Statystyczne metody analizy cech jakościowych*. Państwowe Wydawnictwo Ekonomiczne.

Ustawa z dnia 29 czerwca 1995 r. *o statystyce publicznej* (Dz. U. z 1995 r. Nr 88, poz. 439).

Walesiak, M. (1990). Syntetyczne badania porównawcze w świetle teorii pomiaru. *Przegląd Statystyczny*, (1-2), 37-46.

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

Wiśniewski, M. (2011). Wyznaczniki sytuacji finansowej gminy – ocena istotności za pomocą analizy skupień. *Nauki o Finansach*, *4*(9), 110-119.

Zalewska, E. (2017). Zastosowanie analizy skupień i metody porządkowania liniowego w ocenie polskiego szkolnictwa wyższego*. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, (469), 234-242. https://doi.org/10.15611/pn.2017.469.24

Review and Evaluation of the Usefulness of Clustering Methods in Official Statistics

Abstract: The subject of this work is a review and evaluation of the usefulness of clustering methods in official statistics. The purpose of the analysis is to compare the grouping results obtained with the use of different variants of the methods. The focus was on presenting both advantages and disadvantages of their application on the official statistics data, supported by conclusions and images from the analysis on real data. This topic is developed in detail in three chapters. The first one presents the theory of official statistics and the concepts related to the data preparation. It also classifies the methods of multivariate statistical analysis, which is supported by examples of use. The second chapter describes in detail the compared grouping methods: hierarchical and non-hierarchical methods, as well as various variants of these algorithms at individual stages of the analysis. Additionally, there is a description of their advantages and disadvantages. In the last chapter, the results obtained with the use of selected methods were compared on the basis of actual data on the financial situation of communes in the Dolnośląskie Voivodeship in 2018. The agglomeration divisions made with the Ward, average linkage, and complete linkage methods were compared to those using non-hierarchical algorithms, based on randomly selected cluster centers or on selected as medoids.

Keywords: *k*-means method, financial situation of communes, official statistics