PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 468 • 2017

Taksonomia 28
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Marek Sobolewski
Politechnika Rzeszowska
e-mail: msobolew@prz.edu.pl

Andrzej Sokołowski
Uniwersytet Ekonomiczny w Krakowie
e-mail: sokolows@uek.krakow.pl

GRUPOWANIE METODĄ *K*-ŚREDNICH Z WARUNKIEM SPÓJNOŚCI[1]

CLUSTERING USING *K*-MEANS METHOD WITH COHERENCE PROPERTY

DOI: 10.15611/pn.2017.468.22
JEL Classification: C38, O18

Streszczenie: Metody taksonomiczne są często wykorzystywane do grupowania jednostek administracyjnych (gmin, powiatów, regionów, państw). Ich przedmiotem są najczęściej: jakość życia mieszkańców, atrakcyjność inwestycyjna regionów, poziom rozwoju infrastruktury lub inne bezpośrednio niemierzalne zjawiska. W klasycznych procedurach grupowania nie uwzględnia się powiązań przestrzennych jednostek administracyjnych, co prowadzi zazwyczaj do braku spójności przestrzennej otrzymanych grup. Interesujące byłoby opracowanie modyfikacji metody *k*-średnich pod kątem zapewnienia spójności przestrzennej uzyskiwanych podziałów jednostek terytorialnych. W pracy omówiono zmodyfikowany algorytm *k*-średnich oraz zaprezentowano jego implementację w postaci rozszerzenia programu *STATISTICA*. Przedstawiono przykładowe wyniki grupowania na spójne podzbiory dla danych dotyczących powiatów. Porównano uzyskany podział do klasycznej klasyfikacji, w której nie uwzględniono warunku spójności.

Słowa kluczowe: metoda *k*-średnich, grupowanie z warunkiem spójności, analiza regionalna.

Summary: Clustering methods are often used to group administrative units (municipalities, counties, regions, countries). Their most common subjects are: quality of life, investment attractiveness of regions, the level of development of infrastructure or other directly immeasurable phenomena. In the classic procedures for clustering one does not consider the spatial relationships of administrative units, which can lead to some cognitive dissonance when the isolated groups are characterized by the lack of spatial coherence. In the paper the modified algorithm of the *k*-means method with coherence property and its implementation in the form of an extension of the STATISTICA software was presented. Some examples of clustering

[1] Publikacja została dofinansowana ze środków przyznanych Wydziałowi Zarządzania Uniwersytetu Ekonomicznego w Krakowie w ramach dotacji na utrzymanie potencjału badawczego.

results into coherent subsets for the data on districts were presented. The obtained division was compared to the classic division of the classification, which does not take into account the coherence condition.

Keywords: *k*-means method, clustering with coherence condition, regional analysis.

1. Wstęp

Grupowanie jednostek administracyjnych (państw, regionów, województw, powiatów itd.) metodami analizy skupień daje zwykle podział na podzbiory niespójne terytorialnie, co znacząco zmniejsza możliwość wykorzystania tych rezultatów w praktyce. W klasycznych metodach grupowania uwzględnia się jedynie informacje o wartościach cech diagnostycznych w badanej zbiorowości, nie biorąc w ogóle pod uwagę relacji przestrzennych. Dlatego też użyteczne byłoby opracowanie takich metod, które pozwolą na utworzenie skupień maksymalnie jednorodnych pod względem wartości zmiennych diagnostycznych, zapewniając jednocześnie ich spójność przestrzenną.

W ostatnich latach opublikowano kilka prac zawierających propozycje uwzględniania relacji przestrzennych w analizach taksonomicznych. Kilku autorów, niezależnie od siebie, zaproponowało korektę miernika syntetycznego w metodach porządkowania liniowego ze względu na relacje przestrzenne [Antczak 2013; Sobolewski, Migała, Mentel 2014; Pietrzak 2014; Kuc 2015]. Uzasadnieniem przedstawionych koncepcji było spostrzeżenie, iż jednostki administracyjne oddziałują na siebie wzajemnie, a żaden obszar nie może być traktowany jako zupełnie izolowany. Kwestia ta jest istotna zwłaszcza dla relatywnie niewielkich obszarów (gmin czy powiatów). Jeśli chodzi o metody grupowania, zaproponowano uwzględnienie relacji przestrzennych w algorytmie Warda [Markowska, Sobolewski 2014].

W pracy przedstawiono propozycję modyfikacji algorytmu grupowania metodą *k*-średnich, tak by określane za jego pomocą grupy jednostek terytorialnych były terytorialnie spójne. W efekcie zastosowania zmodyfikowanego algorytmu *k*-średnich zyskujemy spójność przestrzenną utworzonych skupień, dzięki czemu wyniki łatwiej można wykorzystać w zarządzaniu regionalnym. Z drugiej strony, nakładając ograniczenia na proces grupowania, tracimy częściowo spójność ekonomiczną – powstałe skupienia będą mniej jednorodne ze względu na wartości zmiennych diagnostycznych.

2. Algorytm metody *k*-średnich z warunkiem spójności

Metoda *k*-średnich jest szeroko stosowana w analizach danych z różnych dziedzin. Zaletą metody *k*-średnich jest intuicyjność i prostota podstawowej idei obliczeniowej. Przedmiotem analizy jest zbiór n obiektów (np. jednostek terytorialnych) opisywanych przez m cech (tzw. zmiennych diagnostycznych). Celem analizy jest

znalezienie optymalnego podziału wyjściowego zbioru obiektów na *k* podzbiorów, przy czym kryterium jakości podziału jest maksymalizacja sumy wariancji międzygrupowej zmiennych diagnostycznych (równoważnie – minimalizacja wariancji wewnątrzgrupowej). W metodzie *k*-średnich odległości pomiędzy obiektami określa się za pomocą odległości euklidesowej lub jej kwadratu (specyfika algorytmu powoduje, że wyniki w obu przypadkach są takie same).

Algorytm *k*-średnich można opisać w trzech punktach:

1. Punktem wyjścia jest podział danego zbioru obiektów na *k* podzbiorów (najczęściej generowany poprzez przypisanie każdego elementu do „najbliższego" wstępnie wybranego przedstawiciela *k* grup).
2. Wyznaczane są środki ciężkości każdej grupy w przestrzeni zmiennych diagnostycznych.
3. Przypisujemy każdy element do najbliższego środka ciężkości, po czym powracamy do punktu 2, jeżeli choć jeden element został przeniesiony do innej grupy.

Algorytm metody *k*-średnich można traktować jako swego rodzaju „odwrotność" analizy wariancji. Za jego pomocą znajdujemy taki podział badanej zbiorowości na *k* grup, który maksymalizuje wariancję międzygrupową i w konsekwencji statystykę *F*.

Wprowadzenie do metody *k*-średnich warunku spójności przestrzennej (każdy obiekt graniczy przynajmniej z jednym obiektem z tej samej grupy) można rozpatrywać jako nałożenie warunków ograniczających na etap algorytmu opisany w punkcie 3. Pewnych modyfikacji dokonuje się również w punkcie 1, zakładając, iż wyjściowy podział składa się ze spójnych podzbiorów. Oto opis algorytmu *k*-średnich z warunkiem spójności.

1. Punktem wyjścia jest podział danego zbioru obiektów na *k* spójnych podzbiorów (może on być generowany na podstawie odpowiedniego algorytmu bądź być dany *a priori* – na przykład otrzymany za pomocą grupowania metodą Warda z warunkiem spójności, opisaną w pracy [Markowska, Sobolewski 2014]).
2. Wyznaczamy środki ciężkości poszczególnych skupień.
3. Uwzględniamy tylko obszary graniczne, i to takie, których usunięcie z danego skupienia nie narusza jego spójności, i szukamy wśród nich takich obszarów, które leżą bliżej środka ciężkości graniczącego z nimi skupienia niż środka ciężkości „własnego" skupienia – jeżeli nie ma takich obszarów kończymy obliczenia, w przeciwnym wypadku przesuwamy do odpowiedniego skupienia ten obszar, który daje największy przyrost zmienności międzygrupowej dla zmiennych diagnostycznych i wracamy do p. 2.

Reasumując, różnice w stosunku do klasycznego algorytmu *k*-średnich są dwie: po pierwsze, ograniczamy możliwość przenoszenia obiektów tylko do tych, które nie powodują powstawania niespójnych grup – czyli do części obiektów leżących przy granicy skupień; po drugie, po każdym przeniesieniu jednego obiektu, wyliczamy od początku środki ciężkości – w algorytmie klasycznym przenosi się od razu wszystkie obiekty, które są bliższe środka ciężkości jakiejkolwiek innej grupy niż własnej.

Komentarza wymaga definiowanie wyjściowego podziału. W klasycznej metodzie *k*-średnich zaproponowano kilka sposobów generowania początkowego podziału – może być on uzależniony od wskazanych arbitralnie elementów skupień, może być optymalizowany pod kątem jednorodności. W spójnym algorytmie *k*-średnich proponujemy wykorzystać jedną z opcji:

- arbitralny podział administracyjny (na przykład przy klasyfikacji województw na 6 skupień mogą być to regiony Polski na poziomie NUTS-1);
- podział wygenerowany za pomocą spójnej metody Warda;
- spójny podział uzyskany poprzez dołączanie kolejnych elementów sąsiadujących ze wstępnie zdefiniowanymi elementami skupień.

3. Implementacja algorytmu

Algorytm grupowania metodą *k*-średnich z warunkiem spójności został zaimplementowany jako autorskie rozszerzenie pakietu *STATISTICA*[2]. Opracowany program jako wejścia do analizy wymaga dwóch plików – z wartościami zmiennych diagnostycznych oraz z informacją o relacjach przestrzennych (czyli macierzy sąsiedztwa). W wyniku obliczeń w wyjściowym arkuszu danych pojawia się nowa zmienna, zawierająca wyniki grupowania spójną metodą *k*-średnich.

Trudność techniczną przy opracowywaniu algorytmu stanowiła kontrola spójności obszarów na każdym etapie algorytmu. Czasochłonne jest sprawdzanie, czy wykluczenie danego elementu ze skupienia nie narusza jego spójności. Okazuje się jednak, iż na początek wystarczy sprawdzić spójność zbioru wszystkich sąsiadów usuwanego obiektu. Jeżeli „półpierścień" sąsiadów danego obiektu jest spójny, jego usunięcie nie spowoduje naruszenia spójności całego zbioru. Ponieważ większość obszarów zwykle spełnia ten warunek, czas obliczeń można dość znacznie skrócić.

W programie zaimplementowano dwie możliwości określania wyjściowego podziału. Może być to informacja o spójnym przestrzennie podziale określona w zmiennej wejściowego arkusza danych bądź też efekt działania algorytmu doboru wejściowego podziału określonego w programie. W tym drugim przypadku użytkownik proszony jest o podanie numerów *k* obiektów, będących reprezentantami każdego ze skupień. Wśród ich sąsiadów przestrzennych wyszukiwany jest obiekt najbliższy w sensie odległości ekonomicznej, który jest dołączany do odpowiedniego wyjściowego elementu. Ta procedura jest powtarzana aż do uzyskania kompletnego podziału – na każdym jej etapie do pewnego z *k* skupień „dolepiany" jest obiekt z nim sąsiadujący i najbliższy w sensie odległości ekonomicznej.

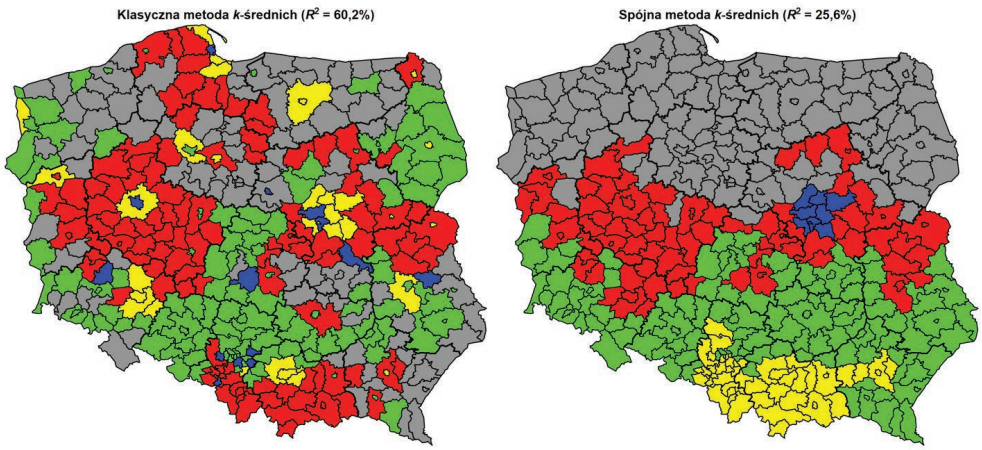
---

[2] Ponieważ celem poniższego opracowania jest rozpowszechnienie idei grupowania „spójnego" program ten będzie udostępniany wszystkim zainteresowanym osobom. Na pewno przyczyni się to do krytycznej oceny zarówno samej implementacji, jak i praktycznej wartości algorytmu grupowania spójną metodą *k*-średnich.

4. Przykładowe wyniki

W przedstawionym poniżej przykładzie analizowano poziom życia w powiatach w roku 2013. Przykład ma mieć charakter poglądowy, dlatego też liczbę zmiennych diagnostycznych ograniczono do czterech wskaźników: wynagrodzenia, stopy bezrobocia, wskaźnika rodności, wskaźnika wybudowanych mieszkań (na tys. mieszk.).

Dane poddano klasycznej procedurze standaryzacji, a następnie dokonano podziału na 5 skupień za pomocą klasycznej metody *k*-średnich oraz jej wersji z warunkiem spójności przestrzennej. Uzyskane wyniki przedstawiono na rys. 1.



Rys. 1. Porównanie wyników grupowania powiatów ze względu na wartości wybranych wskaźników ekonomiczno-społecznych uzyskanych za pomocą klasycznej metody *k*-średnich i metody *k*-średnich z warunkiem spójności

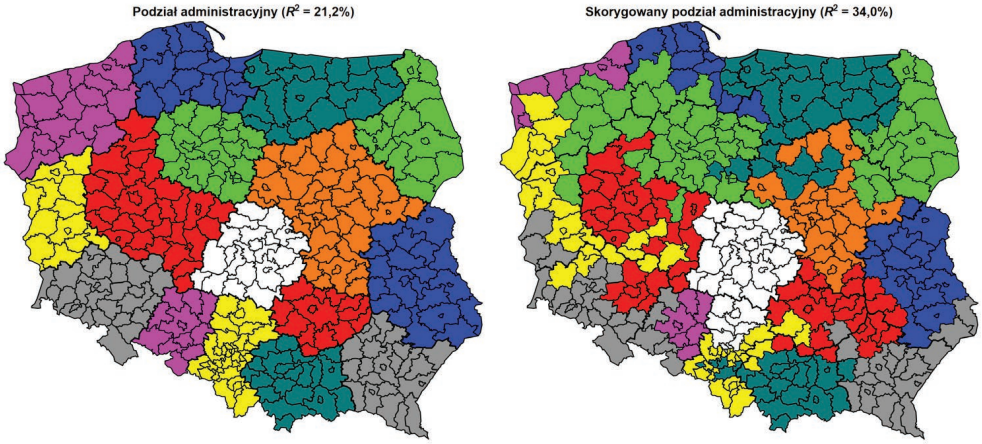
Źródło: opracowanie własne.

Aby określić stopień jednorodności powstałych skupień, wykorzystano współczynnik determinacji (R^2) obliczony na podstawie wyników jednoczynnikowej analizy wariancji. Ponieważ mamy cztery zmienne diagnostyczne, wyznaczono średnią wartość R^2. Podział uzyskany za pomocą klasycznej metody *k*-średnich charakteryzuje się dość wysoką jednorodnością skupień, bowiem uśredniony po zmiennych diagnostycznych współczynnik determinacji wynosi 60%. Jednakże z praktycznego punktu widzenia problemem jest chaotyczne rozmieszczenie elementów poszczególnych skupień. Jak bowiem wykorzystać w praktyce fakt przynależności do jednego skupienia (oznaczonego na rysunku kolorem zielonym) powiatu sanockiego, kołobrzeskiego, jeleniogórskiego i augustowskiego, które są położone na czterech krańcach Polski. Po zastosowaniu spójnej metody *k*-średnich otrzymano podział o mniejszej jednorodności ekonomicznej (R^2 = 26%), ale uporządkowanym układzie terytorialnym. Jako ciekawy fakt uznać należy równoleżnikowy układ utworzonych

skupień. Oczywiście, ocena układu uzyskanych skupień może prowadzić do decyzji o zmianie ich liczby – na przykład w spójnym podziale, w skupieniu oznaczonym kolorem czerwonym (rys. 1) widoczne jest wyraźne „przewężenie", co może sugerować konieczność dokonania podziału na 6, a nie 5 grup. Kończąc omawianie, w skrótowy z konieczności sposób, wyników grupowania, podkreślić należy, iż od praktycznego celu prowadzonej analizy zależy, czy bardziej przydatne będą wyniki klasycznego czy też zmodyfikowanego algorytmu.

Drugi przykład dotyczy tej samej zbiorowości i tych samych zmiennych diagnostycznych, lecz inny jest cel analizy. Punktem wyjścia jest podział administracyjny na 16 województw, zaś algorytm *k*-średnich z warunkiem spójności wykorzystano do oceny jednorodności województw ze względu na wartości rozważanych cech diagnostycznych. Obrazowo rzecz ujmując, jeżeli któryś z powiatów granicznych nie pasuje do „swojego" województwa w sensie jednorodności ekonomicznej, przenoszono go do województwa „bliższego".

Na rys. 2 zamieszczono wyjściowy podział administracyjny Polski na województwa i podział skorygowany pod kątem jednorodności ekonomicznej województw. Zróżnicowanie województw, oceniane jako średnia wartość współczynnika R^2 dla zmiennych diagnostycznych, wynosi 21,2%, zaś po korekcie za pomocą metody *k*-średnich więcej, bo ok. 34%. W sumie dokonano przeniesienia 108 powiatów, które w przestrzeni rozważanych zmiennych diagnostycznych bardziej „pasują" do sąsiednich województw.



Rys. 2. Przykładowy wynik grupowania spójną metodą *k*-średnich na podstawie arbitralnie określonego podziału wyjściowego – ocena jednorodności województw w podziale na powiaty pod względem wartości wybranych wskaźników ekonomiczno-społecznych

Źródło: opracowanie własne.

Objętość artykułu nie pozwala na przytoczenie dokładniejszych wyników, warto jednak podać, z których województw „ubyło" najwięcej powiatów, a z których najmniej – otóż najmniejszą jednorodnością ekonomiczną charakteryzowało się woj.

zachodniopomorskie (po korekcie pozostało w nim ok. 40% powiatów), na drugim biegunie sytuuje się województwo łódzkie, które nie „straciło" żadnego powiatu, „zyskując" kilkanaście nowych (głównie z woj. śląskiego).

5. Podsumowanie

Wprowadzenie do algorytmu *k*-średnich warunku spójności pozwala uzyskać koherentne przestrzennie podziały jednostek terytorialnych klasyfikowanych według wartości zmiennych diagnostycznych. W pracy przedstawiono propozycję takiego algorytmu, jego implementację oraz rezultaty przykładowych analiz dla przestrzeni regionalnej Polski na poziomie powiatów.

Wprowadzając warunek spójności, poprawiamy obraz przestrzenny, ale pogarszamy jakość dokonanego podziału, w sensie jednorodności powstałych skupień. Z praktycznego punktu widzenia spójność przestrzenna wydzielanych obszarów wydaje się jednak bardzo istotna – takimi obszarami po prostu łatwiej zarządzać. Optymalne rozwiązanie to przeprowadzenie grupowania na dwa sposoby – z warunku spójności i bez tego warunku – oraz porównanie obu podziałów. Jeżeli utrata zmienności międzygrupowej w przestrzeni zmiennych diagnostycznych jest niewielka, na pewno warto wybrać podział spójny przestrzennie.

Zaproponowany algorytm pozwala na przenoszenie obszarów granicznych – można i należy rozbudować proces przeszukiwania tak, by sięgał „w głąb" przekształcanych skupień – to na pewno zwiększyłoby jego efektywność i jakość końcowych podziałów.

Przedmiotem dalszych rozważań może być też związek korelacji przestrzennych zmiennych diagnostycznych z jakością podziału uzyskiwanego w wyniku grupowania spójnego. Można postawić roboczą hipotezę, iż w przypadku występowania silnych dodatnich autokorelacji przestrzennych wszystkich zmiennych diagnostycznych wyniki uzyskiwane za pomocą metody z warunkiem spójności będą zbliżone do rezultatów klasycznej metody grupowania.

Literatura

Antczak E., 2013, *Przestrzenny taksonomiczny miernik rozwoju*, Wiadomości Statystyczne, nr 7, s. 37-53.

Kuc M., 2015, *Wpływ sposobu definiowania macierzy wag przestrzennych na wynik porządkowania liniowego państw Unii Europejskiej pod względem poziomu życia ludności*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 384, s. 163-170.

Markowska M., Sobolewski M., 2014, *Wrażliwość regionalnych rynków pracy Unii Europejskiej na kryzys ekonomiczny. Klasyfikacja metodą Warda z warunkiem spójności*, Acta Universitatis Lodziensis, Folia Oeconomica 6 (308), s. 79-94.

Pietrzak M.B., 2014, *Taksonomiczny miernik rozwoju (TMR) z uwzględnieniem zależności przestrzennych*, Przegląd Statystyczny, zeszyt 2, s. 181-201.

Sobolewski M., Migała-Warchoł A., Mentel G., 2014, *Ranking poziomu życia w powiatach w latach 2003-2012 z uwzględnieniem korelacji przestrzennych*, Acta Universitatis Lodziensis, Folia Oeconomica 6 (308), s. 147-159.