Czasopismo Prawa Karnego i Nauk Penalnych
Journal of Criminal Law and Penal Studies
ISSN: 1506-1817 e-ISSN: 2719-6569
Rok/Volume XXV: 2021 Zeszyt/No. 3

Andrzej Porębski[1]

Analiza skupień zorganizowanych grup przestępczych jako przykład zastosowania metod statystycznych w kryminologii

Cluster Analysis of Organized Crime Groups as an Example of the Employment of Statistical Methods in Criminology

1. Wprowadzenie. O bagatelizowaniu roli metod statystyki matematycznej w badaniach kryminologicznych

Jeśli chciałoby się stworzyć listę dziedzin nauki w swej istocie interdyscyplinarnych, z całą pewnością musiałaby się znaleźć na niej kryminologia. Istnieje wiele możliwości definiowania przedmiotu kryminologii. Wszystkie z nich, a także sama etymologia słowa (łac. *crimen*, gr. *logos*), wskazują na związek kryminologii z badaniami przestępczości[2]. Te natomiast nie mogłyby osiągnąć satysfakcjonującego poziomu, gdyby ograniczono się w nich do przyjęcia tylko jednej z możliwych perspektyw. Zjawisko przestępczości cechuje się zbyt wysoką złożonością i wielowymiarowością, aby dało się je w sposób wystarczający analizować tylko w obrębie psychologii czy korzystając wyłącznie z metod socjologii. Wśród nauk, które powinny być wykorzystywane w badaniach kryminologicznych, wymienić można na przykład wspomniane już psychologię i socjologię, ale też pedagogikę, nauki medyczne i statystykę[3]. W obrębie

[1] Andrzej Porębski – mgr, Uniwersytet Jagielloński, Szkoła Doktorska Nauk Społecznych, Wydział Prawa i Administracji / M.A., Doctoral School in the Social Sciences, Jagiellonian University, Kraków; ORCID: 0000-0003-0856-5500; ✉ poreand@gmail.com.

[2] Zob. J. Błachut, A. Gaberle, K. Krajewski, *Kryminologia…*, s. 17 i n.; B. Hołyst, *Kryminologia…*, 2007, s. 39 i n.

[3] Szerzej na temat interdyscyplinarności kryminologii i jej związku z innymi dziedzinami nauki zob. np.: J. Błachut, A. Gaberle, K. Krajewski, *Kryminologia…*, s. 23–32; M. Kuć, *Kryminologia…*, s. 6–11; B. Hołyst, *Kryminologia…*, 2007, s. 65–78.

prezentowanego opracowania to relacja kryminologii z ostatnią z wymienionych nauk – statystyką – będzie najistotniejsza.

Bliskie powiązania między statystyką, rozumianą szeroko, jako nauka o analizowaniu danych[4], a kryminologią stają się łatwo widoczne, gdy zauważy się, że metodologia badań psychologicznych i socjologicznych w dużej mierze opiera się na metodach statystycznych. Jednocześnie rozmiar przestępczości musi być przecież w konkretny sposób mierzony. Niezbędna okazuje się więc statystyka, która pozwala we właściwy sposób interpretować dane ilościowe dotyczące zjawiska przestępczości. W kryminologii dane liczbowe pełnią bardzo ważną rolę – a gdzie występują dane, tam cenna jest pogłębiona ich analiza.

Roli statystyki dla kryminologii nikt w gruncie rzeczy nie neguje, wiele pozostawia jednak do życzenia wykorzystywanie jej metod przez badaczy działających w obrębie tej dziedziny. Wydaje się, że mnogość danych o przestępczości, zbieranych zarówno przez organy ścigania, jak i przez naukowców w ramach badań empirycznych, jest wystarczająca, aby była możliwa ich wnikliwa, zaawansowana analiza. Powinna ona dalece wykraczać poza najprostsze statystyki opisowe, elementarne testy statystyczne i podstawowe wizualizacje danych. Nie wydaje się jednak, aby polskie badania kryminologiczne zmierzały w tym kierunku.

Kryminologia Brunona Hołysta – bodaj najobszerniejszy podręcznik przedmiotu – w wydaniu z 2007 r. oprócz informacji na temat podstawowych cech systemu gromadzenia danych o przestępczości zawiera również niezwykle przystępne omówienie podstaw analizy statystycznej zjawisk społecznych[5], a także część dotyczącą wykorzystania metod statystycznych w badaniach kryminologicznych[6]. Nie powinno szczególnie dziwić, że na opis „zaawansowanych metod statystycznych" w ramach wspomnianych części poświęcono zaledwie cztery strony – autor jasno zaznacza, że w rozdziałach tych nie tworzy kompletnego wykładu metod statystycznych, a rozpatrując je, nie przyjmuje perspektywy statystyka – profesjonalisty[7]. Bardziej niepokoi fakt, że prezentacja zupełnych

[4] Istotne jest tutaj pokreślenie, że określenia „statystyka" używa się nie tylko wobec nauki, ale również w odniesieniu do danych liczbowych opisujących jakieś zjawisko, a także jako funkcji matematycznej określonej na przestrzeni statystycznej. W tej pracy istotne będą pierwsze dwa znaczenia, a najistotniejsze – znaczenie pierwsze.

[5] B. Hołyst, *Kryminologia*..., 2007, s. 131–188.

[6] B. Hołyst, *Kryminologia*..., 2007, s. 189–206.

[7] B. Hołyst, *Kryminologia*..., 2007, s. 131 i 189.

podstaw statystycznej analizy danych na kilkudziesięciu stronach podręcznika Hołysta stanowi wciąż jedno z obszerniejszych rozważań tego typu obecnych w polskiej literaturze kryminologicznej. Warto tutaj zaznaczyć, że istotna dla prawdziwości poprzedniego zdania jest wcześniej przytoczona data wydania. Wydanie IX rozszerzone było ostatnim, w którym omówiono zagadnienia metodologii badań kryminologicznych (a wraz z nimi, elementy nauki statystyki) – w wydaniach X oraz XI część ta już się nie pojawia[8].

Ze względu na mnogość publikacji naukowych dotyczących kryminologii nie sposób dokonać generalizacji dotyczącej niedostatecznego w nich wykorzystania statystyki jako nauki. Jednak nawet pobieżne prześledzenie prac odwołujących się w swojej treści do nowych wyzwań cywilizacyjnych wskazuje jasno, że wykorzystanie współczesnych metod analizy danych nie kojarzy się kryminologom z wyzwaniami współczesności. Wyzwań tych dopatrują się oni raczej na zewnątrz uprawianej dziedziny. Na przykład – w monografii *Kryminologia wobec współczesnych wyzwań cywilizacyjnych* z 2010 r. wyrazy pokrewne ze słowem „statystyka" padają szesnaście razy i w zdecydowanej większości dotyczą danych statystycznych, a nie metod statystycznych[9]. Jedyna wzmianka o „obliczeniach statystycznych" dotyczy obliczenia prostego wskaźnika (skądinąd, podany w pracy wzór nijak ma się do wykonanych obliczeń – jest po prostu błędny[10]).

Z kolei w pracy *Kryminologia. Współczesne aspekty* z 2014 r.[11] szczegółowo omówiono pewne zagadnienia z zakresu gromadzenia statystyk o przestępczości i problemów się z nimi wiążących, ale już nie współczesne aspekty wykorzystania posiadanych danych. Można odnieść generalne wrażenie, że w polskiej kryminologii znaczącą uwagę poświęca się problemom związanym z pozyskiwaniem danych[12]. Na drugi (czy nawet dalszy) plan odsuwa się natomiast to, co z pozyskanymi danymi uczynić, aby należycie wykorzystać drzemiący w nich potencjał.

Spostrzeżenia te nie stanowią oczywiście krytyki doboru tematów przez autorów rozważanych dzieł. Mają ono jedynie ukazać być może

---

8 Zob. B. Hołyst, *Kryminologia*..., 2009; B. Hołyst, *Kryminologia*..., 2016.

9 *Kryminologia...*

10 G. Kędzierska, *Kryminologiczna...*, s. 28.

11 J. Wójcik, *Kryminologia...*

12 Zob. np. J. Wójcik, *Kryminologia...*, s. 103–133; J. Błachut, *Problemy...*

niedostatecznie analityczny klimat myśli kryminologicznej, zarysować bieżące postrzeganie „współczesnych zagadnień", z którymi mierzyć ma się kryminologia i zasygnalizować problem zbyt małej uwagi poświęcanej dostępnym obecnie technologiom analizy danych.

Poczynione uwagi w połączeniu z wiedzą na temat potencjału nowoczesnych metod analizy danych prowadzą do pesymistycznej konstatacji – niewielka rola bardziej zaawansowanych zastosowań statystyki w polskich badaniach kryminologicznych ogranicza te badania. Z jednej strony, w wielu przypadkach ilość informacji uzyskana z prezentowanych danych mogłaby być zwielokrotniona, gdyby wykonać pogłębioną ich analizę. Ze strony drugiej, pewne tematy i podejścia do badań kryminologicznych – między innymi wszelkiego rodzaju predykcje oraz statystyczne modelowanie wpływu określonych czynników na cechy szczególnie dla kryminologii interesujące, takie jak popełnienie czy powrót do przestępstwa – nie zyskają należnej sobie pozycji tak długo, jak długo kryminolodzy stronić będą od zaawansowanych metod analizy danych.

W prezentowanej pracy przedstawiona zostanie jedna z metod analizy danych, która wydaje się być cenna w badaniach kryminologicznych – analiza skupień. Ważnym celem tego tekstu jest nie tylko zaprezentowanie analizy skupień, ale też prześledzenie potencjalnych korzyści z jej stosowania, także w praktycznym kontekście, czyli w badaniach nad zorganizowanymi grupami przestępczymi (ZGP). Z tego względu w pracy nie pojawią się szersze rozważania o matematycznych podstawach i zaawansowanych aspektach rozpatrywanej metody, a precyzyjniej opisana zostanie tylko jedna jej odmiana. To sprawia, że tekstu nie można traktować jako samodzielnego omówienia analizy skupień, a jedynie jako prezentację jej podstaw w określonym potencjalnym zastosowaniu.

Najważniejszy cel przyświecający powstawaniu tej pracy to próba wskazania na korzyści płynące ze stosowania metod statystycznych w badaniach kryminologicznych i pokazania, że nie warto systemowo stronić od tych metod. W tekście prezentowana jest wyłącznie analiza skupień, ale po pewnym przekonstruowaniu duża część uwag o jej użyteczności mogłaby być uogólniona na inne metody statystyczne.

2. Analiza skupień

2.1. Charakterystyka metody

Analiza skupień (klasteryzacja, grupowanie – te nazwy będą używane zamiennie; ang. *cluster analysis*, *clustering*) stanowi jedną z metod statystycznych, której wykorzystanie ma umożliwić wyodrębnienie ze zbioru danych określonej liczby grup podobnych do siebie obiektów (obserwacji) na podstawie przyjętej miary podobieństwa[13]. Cel jej stosowania – i zarazem przypadek optymalny, z którym oczywiście nie zawsze będzie się miało do czynienia – to wykrycie w zbiorze danych grup (skupień) naturalnie obecnych w ich strukturze, wynikających z charakterystyki danych, których wyodrębnieniu można przypisać określoną, sensowną interpretację. Innymi słowy, chodzi o wyodrębnienie z wyjściowego zbioru danych o obiektach takich podzbiorów obiektów, że elementy różnych podzbiorów różnić się będą od siebie bardziej niż elementy należące do tego samego podzbioru[14].

Warto przytoczyć w tym miejscu nieco bardziej formalną definicję analizy skupień: „narzędzie analizy danych służące do grupowania *m* obiektów, opisanych za pomocą wektora *p* cech w *K* niepustych, rozłącznych i możliwie «jednorodnych» grup-skupień”[15]. Należy zaakcentować dwa elementy tej definicji. Po pierwsze, liczba grup, na które podzielony zostanie zbiór obiektów, nie jest odgórnie określona przez sam algorytm i decyduje o niej analityk. Po drugie, obiekty grupowane są wedle posiadanych o nich danych, czyli danych, które je opisują. To niezwykle istotne, aby mieć świadomość, że klasteryzacja obiektów dotyczy każdorazowo wybranego ich przedstawienia. Inne przedstawienie obiektów, czyli inny dobór interesujących w badaniu cech, może sprawić, że powstałe skupienia okażą się być zupełnie inne. Algorytm grupujący nie ma w końcu dostępu do rzeczywistej natury obiektów, a tylko do tych danych o nich, które zdecydowano się sformalizować.

Najczęściej spotykane metody klasteryzacji można podzielić na hierarchiczne – opierające się na łączeniu lub dzieleniu obserwacji – oraz

[13] Przedstawiany tutaj opis analizy skupień można także znaleźć w wersji anglojęzycznej i dostosowanej do kontekstu analizy danych przestrzennych – A. Porębski, *Application…*, s. 99–104.

[14] Zob. S. Wierzchoń, M. Kłopotek, *Algorytmy…*, s. 19–20; M. Krzyśko, W. Wołyński *et al.*, *Systemy…*, s. 345–346.

[15] M. Krzyśko, W. Wołyński *et al.*, *Systemy…*, s. 345.

kombinatoryczne – w których optymalizuje się określoną funkcję jakości grupowania. Metody hierarchiczne dzielą się na aglomeracyjne (na początku każdy obiekt stanowi osobne skupienie, następnie są one łączone) oraz podziałowe (startuje się od jednego skupienia, w którego skład wchodzą wszystkie obiekty, które w kolejnych krokach jest dzielone na skupienia coraz mniejsze). Inne, rzadziej stosowane i bardziej wysublimowane, to między innymi metody relacyjne i grafowe. Z kolei najpopularniejsza metoda kombinatoryczna nosi nazwę metody k-średnich lub centroidów. Należy mieć na uwadze, że dalsza część tego punktu powstała z myślą o przedstawieniu metod hierarchicznych aglomeracyjnych. Jednocześnie uwagi dotyczące formalizacji obiektów pozostają w pełni aktualne również w odniesieniu do innych wskazanych wyżej metod.

2.2. Formalizacja obiektów rzeczywistych

Zagadnieniu formalizacji obiektów rzeczywistych należy poświęcić więcej miejsca. Odpowiedni dobór rozpatrywanych cech badanych obiektów, czyli adekwatna do przedsięwziętego celu formalizacja obiektów, będzie ważnym czynnikiem odróżniającym od siebie analizę danych przemyślaną od tej przeprowadzonej raczej „na oślep". Formalnie sprawa ma się zupełnie prosto[16]. Dany jest zbiór n obiektów, który przedstawiony jest jako macierz:

$X = (x_1, \ldots, x_n)^T$,

z których każdy obiekt opisywany jest p-wymiarowym wektorem (p – liczba cech):

$x_i = (c_{i1}, \ldots, c_{ip})$, gdzie $i \in \{1, \ldots, n\}$.

Wtedy c_{ij} jest j-tą cechą i-tego obiektu, a wektor x_i to obraz obiektu (czyli wektor cech obiektu).

Aby częściowo zdeformalizować powyższe ustalenia, warto przedstawić macierz w formie tabelarycznej. Niech wiersze będą kolejnymi obiektami, a kolumny kolejnymi ich cechami:

[16] To ujęcie jest wystarczające do celów prezentowanej pracy. Bardziej rozbudowany opis formalizacji obiektów oraz pełniejszą formalizację całego zagadnienia analizy skupień znaleźć można w S. Wierzchoń, M. Kłopotek, *Algorytmy…*, s. 20–23.

Cecha / Obiekt	Cecha 1	…	Cecha *p*
Obiekt 1	Wartość cechy 1 obiektu 1	…	Wartość cechy ***p*** obiektu 1
…	…	…	…
Obiekt ***n***	Wartość cechy 1 obiektu ***n***	…	Wartość cechy ***p*** obiektu ***n***

Tabela 1. Uogólniona macierz *p* cech dla *n* obiektów.
Źródło: opracowanie własne.

Z oczywistych względów – w związku z założeniem niepustości i rozłączności skupień – *n* musi być nie mniejsze od liczby grup *K*. W praktyce interesujące będą jedynie te sytuacje, w których *n* będzie zauważalnie przewyższać *K*.

Formalizacja obiektów pozwala na przedstawienie ich w formie punktów w przestrzeni *p*-wymiarowej, w której kolejne wymiary są kolejnymi ich cechami.

Lepsze zrozumienie kluczowego dla analizy skupień zagadnienia matematycznego przedstawiania rzeczywistych obiektów powinno zapewnić zauważenie, że każdy taki obiekt – na przykład zwierzę, człowiek, organizacja, przedmiot użytkowy – może być w pewnym stopniu opisany poprzez określenie, także w sposób liczbowy, jego interesujących charakterystyk. Choć liczba wszystkich cech dowolnego rzeczywistego obiektu jest ogromna – lub nawet nieskończona – to liczba cech, których wyszczególnienie wystarczy do osiągnięcia sprecyzowanego celu często będzie jak najbardziej skończona i możliwa do zawarcia w niewielkim zbiorze (zbiorze o małej mocy).

Na przykład: sprzedawca w sklepie, który ma wątpliwości, czy może sprzedać klientowi alkohol, nie potrzebuje informacji o jego kolorze włosów, wzroście czy płci. Sprzedawcy wystarczyłby formalny obraz klienta uwzględniający dwie jego cechy: wiek oraz to, czy zachowanie osoby wskazuje, że znajduje się ona w stanie nietrzeźwości. Na podstawie tylko tych dwóch cech sprzedawca świadom obowiązujących unormowań będzie w stanie zakwalifikować klienta do osób, które mogą nabyć alkohol lub tych, które nie mogą tego uczynić.

Omówiona powyżej formalizacja jest o tyle trywialna, że bazuje na ściśle określonej ustawowej normie. Materiał normatywny zapewnia kompletne informacje co do tego, na jakie cechy osoby powinno się zwrócić uwagę w danej sytuacji. W klasteryzacji grup przestępczych nie

będzie miało miejsca takie ułatwienie procesu formalizacji – to analityk musi zdecydować, które cechy organizacji przestępczej są istotne dla badań i charakteru jej działalności. Powrót do kwestii doboru właściwego obrazu grupy przestępczej nastąpi w dalszej części tekstu. Przed zakończeniem tego podpunktu warto rozważyć jeszcze dwa przykłady.

Niech pierwsza formalizacja uwzględnia cechy ZGP: *liczebność* oraz *odsetek obcokrajowców wśród członków*. Druga formalizacja niech opiera się na zmiennych binarnych (dwuwartościowych, przyjmujących wartości „prawdy" – 1 lub „fałszu" – 0) opisujących przedmiot działalności: *czy grupa zajmuje się handlem narkotykami?*, *czy grupa zajmuje się kradzieżami samochodów?*, *czy grupa zajmuje się paserstwem?* (wartość 0 dla każdej zmiennej będzie oznaczać *inny rodzaj działalności*). Zależnie od przyjętej formalizacji obiektu, jakim jest grupa przestępcza, zupełnie inne jego cechy stają się znaczące w procesie grupowania. Wobec tego, przy formalizacji pierwszej liczebność ZGP oraz narodowość jej członków będą wpływały na to, czy dwie grupy znajdą się w jednym skupieniu, a przedmiot działalności nie będzie miał znaczenia. Przy formalizacji drugiej sytuacja będzie dokładnie odwrotna – tylko przedmiot działalności wpłynie na zawartość skupień.

Istotna dla prawidłowej interpretacji wyników analizy skupień jest więc świadomość, w oparciu o jakie cechy obiektów tworzone były formalizacje. Koniecznie trzeba też pamiętać, że analizowane dane nie są perfekcyjnym odwzorowaniem rzeczywistych obiektów, a zaledwie ich obrazem (zarówno w określonym wyżej znaczeniu formalnym, jak i potocznym).

2.3. Miara podobieństwa (odmienności)

Formalizacja obiektów to nie wszystko. W samą definicję klasteryzacji wpisane jest poszukiwanie skupisk obiektów podobnych, odróżniających się (odmiennych) od obiektów w innych skupiskach. Poszukiwanie to byłoby niemożliwe bez określenia matematycznych kryteriów podobieństwa, dzięki któremu możliwe będzie przeliczenie danych zawartych w wektorach cech obiektów na podobieństwo między nimi. To kryterium nazywane będzie miarą podobieństwa – lub odmienności; wbrew pozorom nazw tych można używać zamiennie, gdyż maksymalna odmienność obiektów to minimalne podobieństwo i *vice versa*, a tym samym maksymalizacja podobieństwa jest równoważna minimalizacji odmienności.

W ujęciu matematycznym miara podobieństwa to spełniająca pewne (nieistotne tu) warunki funkcja $s: X \times X \rightarrow \mathbb{R}^{+} \cup \{0\}$, czyli przyporządkowanie każdej parze uporządkowanej kwadratu kartezjańskiego rozpatrywanego zbioru dokładnie jednej nieujemnej liczby rzeczywistej.

Jak to już zostało wspomniane, kiedy „posiada" się już sformalizowane obiekty, można umieścić punkty je reprezentujące w p-wymiarowej przestrzeni euklidesowej. Jest to niezwykle przydatne ujęcie, gdyż pozwala na intuicyjne przedstawienie podobieństwa (odmienności) jako odległości między obiektami w tak sformułowanej przestrzeni. Jakkolwiek można rozważać i wprowadzać rozliczne miary podobieństwa (spośród których warte wspomnienia są różne warianty miar korelacyjnych, powiązanych ze współczynnikiem korelacji Pearsona, Spearmana lub Kendalla), najczęściej stosowanymi są funkcje odległości, nazywane również metrykami (w literaturze funkcjonują oba określenia). Aby miarę uznać za odległość (metrykę), funkcja pretendująca do tego miana $d: X \times X \rightarrow \mathbb{R}^{+} \cup \{0\}$ (na tym etapie widoczna jest już nieujemność jej zbioru wartości) musi spełniać zdania:

1) $d(x, y) = 0 \Leftrightarrow x = y$,
2) $d(x, y) = d(y, x)$,
3) $d(x,y) \leq d(x, z) + d(z, y)$.

Zdefiniowanie miary podobieństwa jako odległości jest o tyle przydatne, że umożliwia osiągnięcie wysokiej intuicyjności powstałego systemu formalnych pojęć. Obiekty „bliższe" sobie, czyli takie, pomiędzy którymi odległość jest mniejsza, będą też mniej od siebie odmienne, czyli będą sobie bardziej podobne.

Za najpopularniejszą i najbardziej podstawową odległość stosowaną w analizie skupień dla zmiennych ilościowych można uznać odległość euklidesową:

$$d_E(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2},$$

gdzie:
x_i – wartość i-tej cechy obiektu x,
y_i – wartość i-tej cechy obiektu y,
p_i – liczba cech rozpatrywanych obiektów.

Inną szeroko stosowaną jest odległość Manhattan (miejska, taksówkowa):

$$d_M(x, y) = \sum_{i=1}^{p} |x_i - y_i|.$$

Istotna dla prezentowanych później badań jest również miara Sokala i Miechnera, bazująca na równości poszczególnych cech zmiennych między obiektami. Jest ona szczególnie cenna przy analizie obiektów o obrazach zawierających tylko zmienne nominalne wielostanowe:

$$d_{SMW}(x, y) = \frac{p - p_r}{p},$$

gdzie:
p_r – liczba cech przybierających tożsame wartości w obiekcie x oraz y,
p – liczba wszystkich zmiennych (cech).

Dla celów prezentowanej pracy niepotrzebne są szersze rozważania na temat miar podobieństwa[17]. Warto jednak zasygnalizować istnienie przydatnej w identyfikacji obserwacji odstających odległości Mahalanobisa[18] oraz znajdującej zastosowanie w badaniach nad dokumentami tekstowymi odległości cosinusowej. Należy zdawać sobie sprawę z mnogości miar podobieństwa, bo ich właściwy dobór, uwzględniający charakter posiadanych danych, cel i założenia badania, to ważny element analizy skupień.

Po określeniu, jaka miara podobieństwa będzie stosowana w danej analizie skupień, możliwe staje się uzyskanie macierzy zawierającej zbiorczą informację o podobieństwie dwóch dowolnych rozpatrywanych obiektów względem siebie. Jeśli miara podobieństwa była odległością, macierz ta nazywana będzie macierzą odległości. Jest ona podstawą do wykonania klasteryzacji hierarchicznej.

---

[17] Szerzej na temat miar podobieństwa zob. np.: M. Walesiak, *Pomiar...*, s. 71–85; S. Wierzchoń, M. Kłopotek, *Algorytmy...*, s. 136–139; M. Krzyśko, W. Wołyński *et al.*, *Systemy...*, s. 23–34.

[18] Zob. np. artykuł, w którym wprowadzono tę metrykę: P. Mahalanobis, *On the generalised...*

2.4. Algorytm grupowania

Jak to już zostało wspomniane, spotkać można się z różnymi typami algorytmów mających na celu pogrupowanie obiektów. Dla realizacji założonego celu pracy – ogólnego przedstawienia możliwości zastosowania analizy skupień w kryminologii na przykładzie badań nad ZGP – najstosowniejsze jest rozważenie metody klasteryzacji hierarchicznej, aglomeracyjnej[19]. Po pierwsze, wspomniane badania z konieczności dotyczyć muszą stosunkowo niewielkiej liczby obiektów (liczba możliwych do przebadania w ramach jednego projektu badawczego grup przestępczych nie jest duża), zatem klasteryzacja tą metodą jest możliwa nawet bez użycia znacznych mocy obliczeniowych, a utworzony w grupowaniu dendrogram – wykres prezentujący skupienia – może być wystarczająco klarowny, aby zapewniać dodatkowe informacje. Po drugie, metody hierarchiczne, szczególnie aglomeracyjne, są prostsze do mniej sformalizowanego przedstawienia, a ich sposób działania łatwiejszy do intuicyjnego zrozumienia. Metody takie można uznać za najpopularniejsze i najczęściej stosowane, zwłaszcza w badaniach niepowiązanych z dużymi liczbami obserwacji (niedotyczących szeroko rozumianego *big data*). Po trzecie, w przeciwieństwie do kombinatorycznej metody k-średnich, klasteryzacja hierarchiczna nie wymaga odgórnego zakładania liczby tworzonych skupień, co jest ułatwieniem dla badacza i pozwala na łatwe modyfikowanie decyzji, na ile grup podzielony ma być zbiór danych. Ostatecznie, metody aglomeracyjne, inaczej od metody k-średnich, są deterministyczne (nielosowe) i „monotoniczne". Wynik za ich pomocą uzyskany jest zawsze taki sam, a zmiana liczby skupień o *m* doprowadzi do zmiany skupienia tylko przez obiekty *m* grup.

Algorytm aglomeracyjnej odmiany klasteryzacji hierarchicznej rozpoczyna się od stworzenia *n* skupień, z których każde zawiera pojedynczy obiekt. W następnym kroku, w oparciu o macierz odległości (czy

[19] Więcej informacji na temat hierarchicznych i niehierarchicznych algorytmów analizy skupień znaleźć można m.in. w M. Krzyśko, W. Wołyński *et al.*, *Systemy...*, s. 345–361, a także w obszernym i w całości poświęconym tej tematyce opracowaniu S. Wierzchoń, M. Kłopotek, *Algorytmy...* W Internecie bezpłatnie dostępne są wartościowe anglojęzyczne publikacje dotyczące m.in. analizy skupień: podstawowe ujęcie tematu wraz z przykładami algorytmów grupujących stworzonych w języku R obecne jest w J. Gareth, D. Witten *et al.*, *An Introduction...*, s. 385–401, 404–407, 410–413; bardziej zaawansowane i sformalizowane ujęcie zagadnień klasteryzacji znaleźć można w T. Hastie, R. Tibshirani, J. Friedman, *The Elements...*, s. 501–528.

też: podobieństw; w dalszej części określenia te będą używane zamiennie – ich synonimizacja, mimo występowania formalnych niuansów, nie powinna prowadzić do nieporozumień), znajdywane są dwa obiekty najbardziej do siebie podobne, co jest równoznaczne ze znalezieniem skupień najbardziej do siebie podobnych – na tym etapie każde skupienie składa się wszak z dokładnie jednego obiektu. Skupienia sobie najbliższe są łączone i w ten sposób powstaje pierwsze skupienie dwuelementowe (które zastępuje dwa skupienia, z których zostało utworzone). W kolejnych krokach ponownie łączone są skupienia sobie najbliższe, aż do momentu, zależnie od wersji algorytmu:

a) utworzenia jednego skupienia, zawierającego (łączącego) wszystkie *n* obiektów lub
b) utworzenia liczby skupień równej *K* – czyli pierwotnie zadeklarowanej szukanej liczby skupień.

Wzbudzić zastanowienie może to, wedle jakiego kryterium określa się odległość (podobieństwo) skupień od siebie, skoro tylko na samym początku algorytmu macierz odległości skupień jest równoważna macierzy odległości. Pojawienie się tej wątpliwości jest jak najbardziej uzasadnione – miara podobieństwa międzyskupieniowego musi zostać zdefiniowana, aby proces grupowania mógł zostać ukończony. Algorytm klasteryzacyjny – i wynik jego zastosowania – będzie odmienny w zależności od przyjętej definicji. Wyróżnić można wiele odmiennych sposobów określenia miary podobieństwa między skupieniami[20]. Dla celów tego opracowania wystarczy podanie trzech najbardziej podstawowych:

a) metoda pojedynczego wiązania[21] (ang. *single linkage*; także: metoda minimum) definiuje odległość między skupieniami jako najmniejszą z odległości między dwoma obiektami należącymi do różnych skupień,
b) metoda pełnego wiązania (ang. *complete linkage*; także: metoda maksimum) definiuje odległość między skupieniami jako największą z odległości między dwoma obiektami należącymi do różnych skupień,

[20] W literaturze specjalistycznej często wymienianych jest siedem algorytmów. Zob. np. S. Wierzchoń, M. Kłopotek, *Algorytmy…*, s. 35; F. Murtagh, P. Contreras, *Algorithms…*, s. 88–89.

[21] Szczegółowy, nieformalny opis zastosowania tej metody grupowania znaleźć można w T. Marek, C. Noworol, *Wprowadzenie…*, s. 35–37.

c) metoda średniego wiązania (ang. *average linkage*; także: UPGA, UPGMA) definiuje odległość między skupieniami jako średnią odległość między wszystkimi parami obiektów należących do różnych skupień.

Posiadając określoną miarę podobieństwa między skupieniami, można nieformalnie sformułować uproszczony algorytm postępowania dla hierarchicznej aglomeracyjnej klasteryzacji:

1) każdy obiekt potraktuj jako osobną grupę,
2) znajdź dwie najbliższe sobie grupy i połącz je w jedną grupę,
3) powtarzaj krok 2 do momentu, aż wszystkie obiekty znajdą się w jednej grupie[22].

3. Badania przestępczości zorganizowanej a możliwość zastosowania analizy skupień

We wcześniejszej części tekstu określony został cel wykorzystywania analizy skupień. W tym punkcie omówione zostaną wybrane korzyści, które wiążą się z zastosowaniem tej metody w badaniach nad przestępczością zorganizowaną, a także niektóre z trudności, które mogą wystąpić przy jej stosowaniu.

Obiektami w klasteryzacji wykorzystywanej w badaniach nad przestępczością zorganizowaną będą ZGP. Trudno wyróżnić jeden, powszechnie przyjmowany sposób ich definiowania. W literaturze można się spotkać z wielością określeń i podejść. Na gruncie tej pracy ZGP rozumiane będą możliwie szeroko, jako grupa trzech lub więcej osób prowadzących działalność przestępczą, posiadająca strukturę wewnętrzną, zawiązana na dłuższy lub nieokreślony czas, mająca na celu działanie w celu uzyskania korzyści lub władzy[23].

Jak nietrudno zauważyć, pojęcie ZGP nie precyzuje, jakiego rodzaju działalności aktywność konkretnej grupy dotyczy. Termin ten określa

[22] Alternatywnie: Potarzaj krok 2 do momentu, aż utworzona zostanie liczba skupień równa *K*.

[23] Przedstawiona definicja, będąca syntezą obecnych w tekstach poświęconych tej problematyce podejść, podkreśla tylko najważniejsze cechy definiowanego pojęcia i nie rości sobie prawa do oddawania bogactwa obecnych w literaturze poglądów. Definiowanie pojęć przestępczości zorganizowanej i zorganizowanych grup przestępczych to temat często poruszany. Zob. np.: B. Hołyst, *Kryminologia*..., 2007, s. 414–419; W. Mądrzejowski, *Przestępczość*..., s. 31–36; O. Krajniak, *Zorganizowane*..., s. 35 i n.; K. Laskowska, *Rosyjskojęzyczna*..., s. 18 i n.; A. Michalska-Warias, *Pojęcie*...

tylko minimalną liczebność takiej grupy. Nie odnosi się do poziomu jej hierarchizacji ani nie formułuje sposobu występowania wielu innych cech, które mogą znacząco się różnić w przypadku konkretnych zorganizowanych grup przestępczych, wpływając na ich charakterystykę. Wszystko to sprawia, że wśród wszystkich desygnatów pojęcia ZGP znaleźć można te bardzo sobie podobne, ale i zupełnie odmienne. Oczywiście pełna odmienność z jednej, a tożsamość lub bliskie tożsamości podobieństwo z drugiej strony, to tylko krańce *continuum*. Różne pary rozpatrywanych grup zajmować mogą na nim dowolne miejsce. W trakcie badań zbioru grup przestępczości zorganizowanej interesujący powinien być nie tylko wgląd w każdą z nich osobno, ale też w relacje, jakie między nimi występują – zatem także podobieństwa lub różnice pomiędzy ich strukturą czy sposobem ich działania. Ma to między innymi związek z pytaniami takimi jak: które grupy są do siebie najbardziej podobne, a które najbardziej odmienne pod badanymi względami? Czy można w rozpatrywanym zbiorze wyodrębnić podzbiory spójne lub zbliżone w obrębie pewnych cech? Jak kształtuje się częstość występowania poszczególnych charakterystyk działalności? A może dany zbiór zawiera ZGP zupełnie od siebie odmienne, które nie sposób rozpatrywać wspólnie? Pytania tego typu można mnożyć i wcale nie będzie to działanie bezcelowe. Spojrzenie na całą strukturę posiadanych danych, na relacje i powiązania występujące – lub nie – między poszczególnymi obserwacjami, na poziom makro, a nie mikro, pozwala uzyskać informacje, których nie sposób wydobyć z jednostkowego analizowania rozpatrywanych obiektów.

W poszukiwaniu odpowiedzi na przytoczone pytania i w szeroko rozumianym analizowaniu powiązań między charakterystykami ZGP niezwykle cenna może być analiza skupień. Z punktu widzenia badania przestępczości zorganizowanej możliwość wyodrębnienia skupień ZGP mocno zbliżonych do siebie, spójnych pod pewnymi względami, jest niewątpliwie pożądana. Analiza taka wykazać może w sposób wolny od subiektywizmu istnienie pewnych schematów w strukturze tworzenia i działania przestępczości zorganizowanej. Innymi słowy, może ona prowadzić do automatyzacji procesu klasyfikowania, dając jednocześnie matematyczne, uściślone podstawy procesowi wyróżniania typów grup przestępczych. Ponadto algorytmy hierarchicznego grupowania umożliwiają precyzyjne określenie bliskości poszczególnych skupień i wizualizację tych zależności w formie dendrogramu. To z kolei pozwala na dostrzeżenie,

jak w badanym zbiorze kształtują się powiązania analizowanych cech poszczególnych grup przestępczych z innymi grupami – jak wiele obiektów jest do siebie zbliżonych i jak mocne jest to podobieństwo; które obiekty blisko wiążą się z innymi, a które są odmienne od większości pozostałych.

Wartą podkreślenia korzyścią, która idzie za stosowaniem komputerowych algorytmów klasteryzacji, jest możliwość analizowania nawet bardzo dużych zbiorów obiektów. We współczesnym świecie stosunkowo łatwo dostępne są znacznych rozmiarów zbiory danych. Jednak chcąc je analizować, koniecznie posłużyć się trzeba metodami statystycznymi. Trudno przypuszczać, że z wystarczającą wydajnością i precyzją duży zbiór byłby w stanie analizować człowiek niewspomagany komputerowymi technologiami. Oczywistą metodą analizy bliskości charakteru ZGP w 100- czy 200-elementowych zbiorach[24], których elementy rozpatrywane są pod względem większej liczby cech, będą algorytmy klasteryzacji. Choć można by oczywiście takie zbiory systematyzować, korzystając na przykład wyłącznie ze statystyk opisowych, nie byłoby to działanie o skuteczności porównywalnej z algorytmami analizy skupień. Na koniec należy przypomnieć, że nieosiągalna dla człowieka precyzja i szybkość działania komputerowych algorytmów to znaczące zalety obecne również w analizie mniejszych zbiorów danych – takich jak ten analizowany w prezentowanej pracy.

4. Przykładowe zastosowanie metody analizy skupień do klasteryzacji grup przestępczych

W tej części zaprezentowana zostanie przeprowadzona z użyciem środowiska R[25] klasteryzacja czternastu grup przestępczych, które zostały

[24] Liczby te nie są wcale przesadzone – w samym tylko 2020 r. wszczęto 131 postępowań w związku z art. 258 § 1 oraz art. 258 § 2 Kodeksu karnego [zorganizowana grupa i związek przestępczy] (ustawa z dnia 6 czerwca 1997 r. – Kodeks karny, Dz.U. 2018, poz. 1600, tekst jedn. ze zm.), a średnioroczna liczba wszczętych postępowań z lat 2016–2020 to 108 – zob. *Zorganizowana grupa i związek przestępczy (art. 258)*, Statystyki Policji, < http://statystyka.policja.pl/st/kodeks-karny/przestepstwa-przeciwko-13/63615,Zorganizowana-grupa-i-zwiazek-przestepczy-art-258.html >. Według raportu Centralnego Biura Śledczego Policji w 2017 r. rozbito 176 grup przestępczych – zob. *Efekty pracy CBŚP w 2017 roku*, Centralne Biuro Śledcze Policji, < http://www.cbsp.policja.pl/cbs/aktualnosci/153520,Efekty-pracy-CBSP-w-2017-roku.html >.

[25] Informacja o użytych pakietach znajduje się w bibliografii. Wszystkie znajdujące się w pracy wykresy stanowią opracowanie własne.

opisane w dziele *Rosyjskojęzyczna przestępczość zorganizowana. Studium kryminologiczne* Katarzyny Laskowskiej[26]. Zastosowanie to jest wyłącznie ilustracją i nie należy go pojmować jako samodzielnego projektu badawczego ani najważniejszej części prezentowanej pracy. Możliwości aplikacyjne omawianej metody są dużo szersze, jednak obszerniejsza ich prezentacja przekraczałaby ramy tego tekstu.

W pierwszym z kroków analizy każda z czternastu uwzględnionych ZGP (które dalej nazywane będą „obiektami") została sformalizowana pod względem sześciu cech:

1) przedmiot działalności [*PRZED*, 1 – przemyt lub nielegalny obrót towarami; 2 – nielegalny przerzut osób; 3 – rozboje i wymuszenia],
2) posiadanie „wspólnej kasy" przez grupę [*WK*, 1 – tak; 0 – nie],
3) stosowanie przemocy [*PRZEM*, 1 – tak; 0 – nie],
4) współpraca z innymi grupami przestępczymi [*WSP*, 1 – tak; 0 – nie],
5) liczba członków [*LC*, zmienna ilościowa],
6) liczba narodowości [*LN*, zmienna ilościowa].

Liczba cech jest niewielka ze względu na skrótowy charakter opisów grup w powołanym źródle. Jej zwiększenie doprowadziłoby do tego, że nie każdy obiekt mógłby być scharakteryzowany pod względem każdej cechy, czyli wystąpiłyby braki w danych. Liczbę cech można by oczywiście rozszerzać, jeśli dysponowałoby się odpowiednimi danymi[27].

Istotne jest, że zmienna PRZED jest zmienną kategoryczną (jakościową) wielowartościową, umieszczoną na skali nominalnej. Nie może być ona traktowana jako zmienna ilościowa (umieszczona na skali przedziałowej lub interwałowej). Użycie w tej sytuacji odległości euklidesowej jako miary podobieństwa byłoby błędem metodologicznym, gdyż powodowałoby błędne wrażenie, iż obiekty o PRZED = 1 są bardziej podobne do tych o PRZED = 2 niż do tych o PRZED = 3. Taka zależność oczywiście nie występuje.

Opracowane dane prezentuje tabela 2.

---

[26] K. Laskowska, *Rosyjskojęzyczna...*, s. 336–346.

[27] W rzeczywistości występuje szereg sposobów radzenia sobie z klasteryzacją opartą na danych niepełnych, zawierających brakujące wartości zmiennych. Zagadnienie to wykracza poza zakres prezentowanej pracy, wydaje się jednak istotne dla badań nad przestępczością ze względu na liczne trudności w zdobywaniu danych, z jakimi badania te się wiążą. Szerzej na ten temat np.: A. Matyja, K. Simiński, *Comparison...*; K. Wagstaff, *Clustering...*

OBIEKT	PRZED	WK	PRZEM	WSP	LC	LN
A	2	0*	0	0	3	1
B	1	1	0	0	7	2
C	1	0	0	1	15	3
D	2	0	0	1	7	2
E	1	1	0	0	13	2
F	2	0	0	1	4	2
G	1	1	0	0	3	2
H	2	0	0	0	6	2
I	2	0	0	1	7	3
J	3	0*	1	0	6	2
K	3	1	1	0	9	2
L	3	1	1	1	8	2
M	2	0	0	0	5	2
N	1	0	0	0	6	3

* – brak jasnej informacji co do występowania „wspólnej kasy", prawdopodobnie jej brak.

Tabela 2. Dane o analizowanych obiektach – grupach przestępczych.
Źródło: opracowanie własne na podstawie K. Laskowska, *Rosyjskojęzyczna*..., s. 336–346.

Kolejno zostanie rozważona klasteryzacja dla czternastu obiektów z uwzględnieniem zbioru zmiennych $Z_1 = \{PRZED, WK, PRZEM, WSP\}$ oraz $Z_2 = \{LX, LN\}$. Pierwszy zbiór został dobrany w ten sposób, aby zawierał tylko zmienne nominalne, co pozwoli na zastosowanie miary podobieństwa zbudowanej dla zmiennych nominalnych. Drugi zbiór zawiera dwie występujące w badaniu zmienne ilościowe.

4.1. Wariant Z_1

Mimo powyższych uwag co do kategorycznego charakteru zmiennej *PRZED*, nie będzie konieczna jej binaryzacja[28]. Uda się jej uniknąć ze względu na zastosowanie odpowiedniej miary podobieństwa – omówionej w punkcie 2.3 miary Sokala i Miechnera. Zastosowanie tej miary do zbioru danych pozwala obliczyć ujętą w tabeli 3 macierz podobieństwa między obiektami.

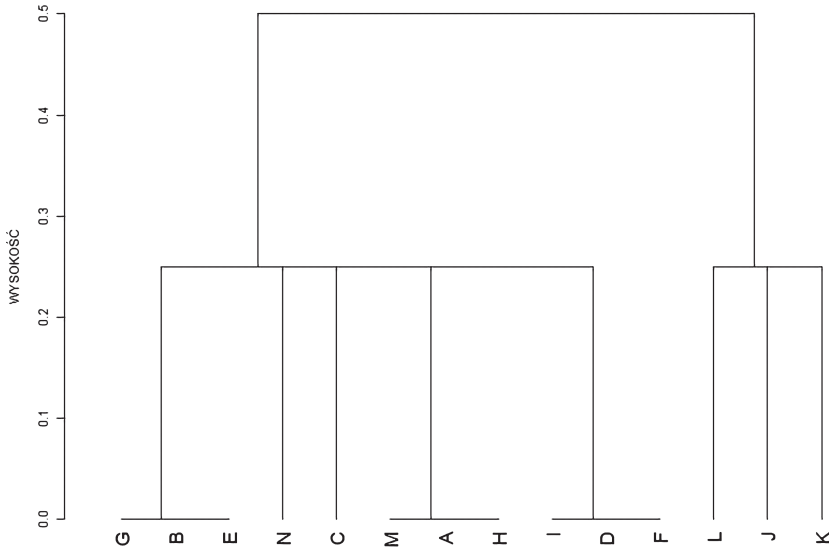
[28] Binaryzacja zmiennej to przekształcenie zmiennej nominalnej na odpowiednią liczbę zmiennych binarnych (dychotomicznych). Najprostszym przykładem binaryzacji jest zamiana zmiennej PŁEĆ przyjmującej wartości {Kobieta, Mężczyzna} na zmienną $PŁEĆ_B$ przyjmującą wartości {1, 0}, w której jako 1 zakodowano wartość Kobieta, a jako 0 wartość Mężczyzna.

	A	B	C	D	E	F	G	H	I	J	K	L	M
B	0,50												
C	0,50	0,50											
D	0,25	0,75	0,25										
E	0,50	**0,00**	0,50	0,75									
F	0,25	0,75	0,25	**0,00**	0,75								
G	0,50	**0,00**	0,50	0,75	**0,00**	0,75							
H	**0,00**	0,50	0,50	0,25	0,50	0,25	0,50						
I	0,25	0,75	0,25	**0,00**	0,75	**0,00**	0,75	0,25					
J	0,50	0,75	0,75	0,75	0,75	0,75	0,75	0,50	0,75				
K	0,75	0,50	**1,00**	**1,00**	0,50	**1,00**	0,50	0,75	**1,00**	0,25			
L	**1,00**	0,75	0,75	0,75	0,75	0,75	0,75	**1,00**	0,75	0,50	0,25		
M	**0,00**	0,50	0,50	0,25	0,50	0,25	0,50	**0,00**	0,25	0,50	0,75	**1,00**	
N	0,25	0,25	0,25	0,50	0,25	0,50	0,25	0,25	0,50	0,50	0,75	**1,00**	0,25

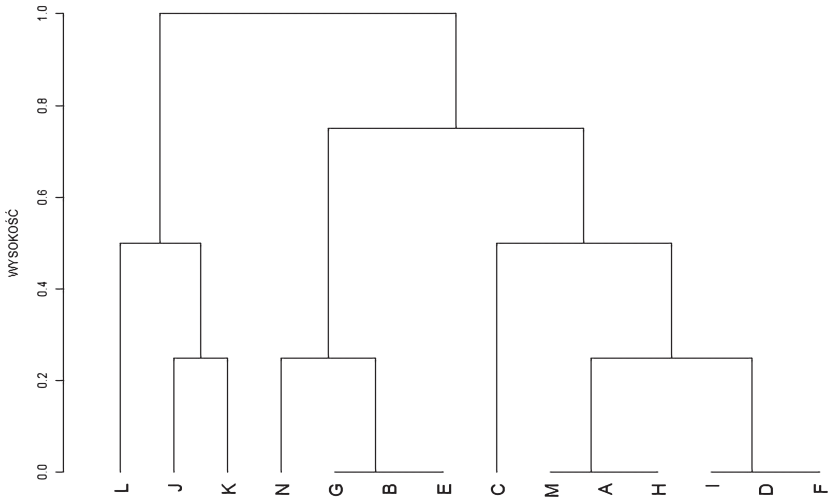
Tabela 3. Macierz podobieństwa obiektów dla zbioru cech Z1.
Źródło: opracowanie własne.

Jak widać, miara podobieństwa przyjmuje tylko cztery różne wartości. Definicja zastosowanej miary powoduje, że przybiera ona tyle wartości, ile cech (zmiennych) jest rozpatrywanych. Wartość 1 (wyróżniona w tabeli na czerwono) oznacza całkowitą odmienność w każdej z cech, natomiast 0 (wyróżniona na jasnoniebiesko) całkowitą tożsamość w obrębie rozpatrywanych zmiennych. Można określić 91 różnych par grup przestępczych, spośród których 9 par będzie łączyło obiekty tożsame, a 8 par łączyło obiekty zupełnie odmienne. Już na tym etapie można zidentyfikować, że: $(A =_{Z1} H =_{Z1} M)$, $(B =_{Z1} E =_{Z1} G)$, $(D =_{Z1} F =_{Z1} I)$, gdzie $=_{Z1}$ jest relacją tożsamości w obrębie cech zbioru Z_I.

Posiadając obliczoną macierz odległości (podobieństwa), stworzyć można dendrogram. Zależnie od wyboru metody grupowania jego wygląd będzie odmienny (wybrane metody zostały omówione w podpunkcie 2.4). Dla celów prezentacyjnych przedstawione zostają dwa dendrogramy, pierwszy – stanowiący wykres 1 – utworzony metodą pojedynczego wiązania (minimum), a drugi – obecny na wykresie 2 – metodą kompletnego wiązania (maksimum).



Wykres 1. Dendrogram dla metody minimum, dla zbioru cech Z_1.
Źródło: opracowanie własne.



Wykres 2. Dendrogram dla metody maksimum, dla zbioru cech Z_1.
Źródło: opracowanie własne.

Im niższa jest wysokość, na której występuje połączenie w dendrogramie, tym wyższe jest podobieństwo skupień. Dlatego obiekty B, E oraz G, a także A, H, M oraz D, F, I łączą się na poziomie zerowym –

jak to już zostało wcześniej wspomniane, są one sobie tożsame w cechach grupy Z_1. Skonstruowawszy dendrogram, można go przyciąć do określonej wysokości lub tak, aby powstała zadana liczba grup. Tabela 4 prezentuje oparty o metodę maksimum wynik podziału obiektów na trzy grupy. Podział dla metody minimum nie będzie dalej prezentowany i opisywany, ponieważ – jak można zauważyć na dendrogramie – prowadzi on do wyklarowania się na niskiej wysokości dwóch bardzo obszernych skupień, przez co wnioski płynące z podziału dokonanego tą metodą nie byłyby zbyt ciekawe.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	2	1	1	2	1	2	1	1	3	3	3	1	2

Tabela 4. Wynik grupowania metodą maksimum dla zbioru cech Z_1, $K = 3$.
Źródło: opracowanie własne.

Wszystkie ZGP w pierwszej grupie nie stosowały w swojej działalności przemocy i nie posiadały wspólnej kasy. Wszystkie oprócz jednej trudniły się nielegalnym przerzutem osób. W drugiej grupie znalazły się zbliżone do siebie ZGP, których przedmiotem działalności był przemyt lub nielegalny obrót towarami, niestosujące przemocy, niewspółpracujące z innymi ZGP. Trzy z czterech obiektów tego klastra posiadały wspólną kasę. Obecne w trzecim skupieniu ZGP zespaja stosowanie przemocy w działalności polegającej na rozbojach i wymuszeniach.

4.2. Wariant Z_2

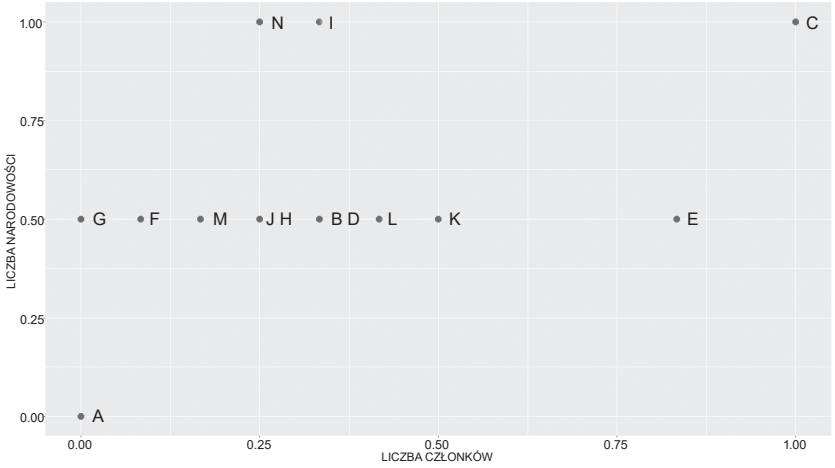
W tym wariancie obie występujące zmienne są ilościowe (choć w pewnych wypadkach można by rozważać inny charakter zmiennej LN). Przyjmują one jednak zupełnie inne wartości. Gdyby pozostawić je nieprzekształcone, zmienna LC wpływałaby w dużo większym stopniu od zmiennej LN na odległość obiektów od siebie. Aby tego uniknąć, konieczna jest normalizacja zmiennych[29]. Zostanie ona wykonana za

[29] Celem normalizacji zmiennych jest przede wszystkim uczynienie danych porównywalnymi. Jest ona bardzo ważnym etapem wstępnej obróbki danych, koniecznym w wielu metodach statycznych. Istnieją rozliczne formuły normalizacji, z których najpopularniejsze to wymieniona metoda minimum-maksimum oraz standaryzacja. Obszerny przegląd różnych metod normalizacji znaleźć można w M. Walesiak, *Przegląd...*

pomocą skalowania minimum-maksimum (ang. *min-max scaling*), zwanego też unitaryzacją zerową. Formuła unitaryzacji to:

$$X_{norma} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Ponieważ w tym wypadku rozpatrywane są tylko dwa wymiary (dwie cechy), dane można łatwo zwizualizować na typowym wykresie dwuosiowym. Zaprezentowany wykres 3 dotyczy danych znormalizowanych.



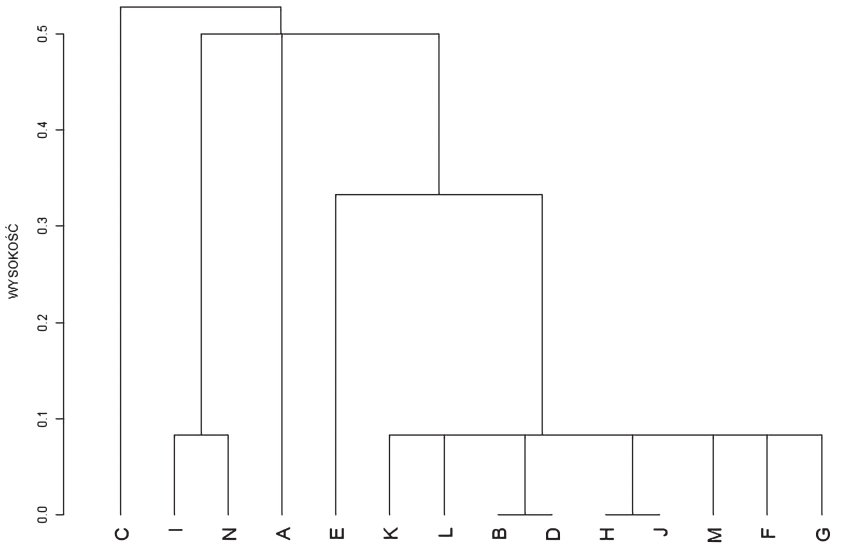
Wykres 3. Znormalizowane dane o cechach obiektów z grupy Z_2.
Źródło: opracowanie własne.

Jak widać, liczba punktów na wykresie wynosi tylko 12. Powodem tego są tożsamości ($J =_{Z2} H$) oraz ($B =_{Z2} D$). Spojrzenie na wykres pozwala przypuszczać, że w obrębie analizowanych obiektów obserwacje A, C oraz E są znacząco odmienne od pozostałych i będą miały tendencję do tworzenia odrębnych skupień. Ponowne stworzenie macierzy podobieństwa i wykonanie klasteryzacji umożliwi zweryfikowanie tych przypuszczeń. Macierz widnieje w tabeli 5, a dendrogramy na wykresach 4 (metoda minimum) i 5 (metoda maksimum). W macierzy wartość 0 wyróżniono kolorem niebieskim, a wartość największą – czerwonym.

Z kolei wpływ wyboru metody normalizacji na wyniki analizy statystycznej pokazano np. w M. Jarocka, *Wybór...*

	A	B	C	D	E	F	G	H	I	J	K	L	M
B	0,60												
C	**1,41**	0,83											
D	0,60	**0,00**	0,83										
E	0,97	0,50	0,53	0,50									
F	0,51	0,25	1,04	0,25	0,75								
G	0,50	0,33	1,12	0,33	0,83	0,08							
H	0,56	0,08	0,90	0,08	0,58	0,17	0,25						
I	1,05	0,50	0,67	0,50	0,71	0,56	0,60	0,51					
J	0,56	0,08	0,90	0,08	0,58	0,17	0,25	**0,00**	0,51				
K	0,71	0,17	0,71	0,17	0,33	0,42	0,50	0,25	0,53	0,25			
L	0,65	0,08	0,77	0,08	0,42	0,33	0,42	0,17	0,51	0,17	0,08		
M	0,53	0,17	0,97	0,17	0,67	0,08	0,17	0,08	0,53	0,08	0,33	0,25	
N	1,03	0,51	0,75	0,51	0,77	0,53	0,56	0,50	0,08	0,50	0,56	0,53	0,51

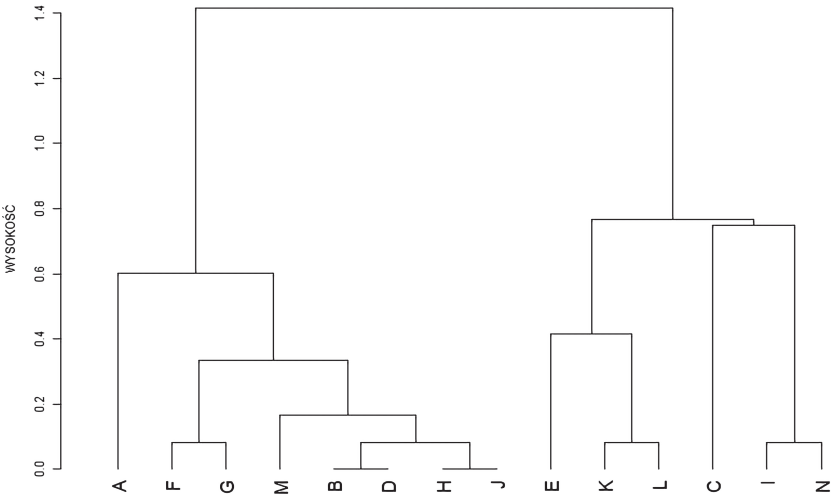
Tabela 5. Macierz podobieństwa obiektów dla zbioru cech Z_2.
Źródło: opracowanie własne.



Wykres 4. Dendrogram dla metody minimum, dla zbioru cech Z_2.
Źródło: opracowanie własne.

Zgodnie z przypuszczeniami, w obu metodach grupowania obiekty A, C i E łączą się z dowolnym innym skupieniem na najwyższych wysokościach. Świadczy to o ich specyfice w obrębie analizowanych cech. Podział na cztery grupy w oparciu o metodę maksimum pokazuje

największą odmienność obiektu C, tworząc dla niego odrębną grupę. Jego wynik można zobaczyć w tabeli 6. Metoda minimum doprowadziła natomiast do powstania na małej wysokości jednego obszernego skupienia, przez co – podobnie jak dla Z_1 – podział tą metodą byłby mniej zróżnicowany (dla $K = 4$: jedna duża grupa, jedna dwuobiektowa i dwie jednoobiektowe) i dlatego nie jest opisywany.



Wykres 5. Dendrogram dla metody maksimum, dla zbioru cech Z_2.
Źródło: opracowanie własne.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	1	2	1	3	1	1	1	4	1	3	3	1	4

Tabela 6. Wynik grupowania metodą maksimum dla zbioru cech Z_2, $K = 4$.
Źródło: opracowanie własne.

We wspomnianej osobnej grupie – drugiej – znalazła się ZGP o największej liczbie członków, aż trzech różnych narodowości. Pierwsza grupa łączy obiekty mniej liczne (od trzech do siedmiu członków). Trzecia grupa zawiera ZGP o większej liczbie członków (od ośmiu do 13), ale reprezentujących tylko dwie narodowości. Z kolei dwie ZGP ujęte w grupie czwartej co prawda mają średnią liczbę członków (sześciu i siedmiu), ale pochodzących z trzech różnych państw.

5. Podsumowanie

Prezentowana praca jest próbą wykazania użyteczności metod analizy skupień do celów badań zorganizowanych grup przestępczych, a przy tym wskazania na potrzebę stosowania metod statystycznych w kryminologii. Wykorzystanie opisywanej metody może pozwolić na uzyskanie z posiadanych już danych wielu nowych informacji. Daje także podstawę do badań kryminologicznych, które bez zastosowania nowoczesnych metod analizy danych po prostu nie byłyby możliwe. Potencjał zawarty w opisywanej w tej pracy metodzie klasteryzacji znaleźć może zastosowanie również w kryminologii. Powinien on zostać dostrzeżony i szerzej wykorzystywany.

Summary

The paper is a methodological study of the cluster analysis method enriched with an empirical example of its application to the study of organized crime. The paper provides a formal (mathematical) and informal description of hierarchical cluster analysis and considers the application of this method to the study of empirical data on organized crime groups. The empirical part originally used qualitative data on 14 organized crime groups, formalized in terms of six variables. Using cluster analysis, clusters of similar criminal groups were distinguished. The study shows that the cluster analysis method can effectively distinguish interpretable clusters of similar criminal groups.

Keywords

cluster analysis, computational social science, criminology, organized crime, application of data analysis methods

Bibliography

Błachut J., *Problemy związane z pomiarem przestępczości*, Kraków 2007.
Błachut J., Gaberle A., Krajewski K., *Kryminologia*, Gdańsk 2007.
Gareth J., Witten D., Hastie T., Tibshirani R., *An Introduction to Statistical Learning with Applications in R*, New York 2017, < https://www-bcf.usc.edu/~gareth/ISL/ >.
Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, New York 2009, < https://web.stanford.edu/~hastie/Papers/ESLII.pdf >.
Hołyst B., *Kryminologia*, Warszawa 2007.
Hołyst B., *Kryminologia*, Warszawa 2009.
Hołyst B., *Kryminologia*, Warszawa 2016.

Jarocka M., *Wybór formuły normalizacyjnej w analizie porównawczej obiektów wielocechowych*, „Ekonomia i Zarządzanie" 2015, nr 1.
Kędzierska G., *Kryminologiczna i kryminalistyczna analiza wybranych elementów udziału kobiet w realizacji przestępstwa*, w: *Kryminologia wobec współczesnych wyzwań cywilizacyjnych*, red. G. Kędzierska, W. Pływaczewski, Olsztyn 2010.
Krajniak O., *Zorganizowane grupy przestępcze. Studium kryminalistyczne*, Warszawa 2011.
Kryminologia wobec współczesnych wyzwań cywilizacyjnych, red. G. Kędzierska, W. Pływaczewski, Olsztyn 2010.
Krzyśko M., Wołyński W., Górecki T., Skorzybut M., *Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*, Warszawa 2008.
Kuć M., *Kryminologia*, Warszawa 2015.
Laskowska K., *Rosyjskojęzyczna przestępczość zorganizowana. Studium kryminologiczne*, Białystok 2006.
Mahalanobis P., *On the generalised distance in statistics*, „Proceedings of the National Institute of Sciences of India" 1936, < https://insa.nic.in/writereaddata/UpLoadedFiles/PINSA/Vol02_1936_1_Art05.pdf >.
Marek T., Noworol C., *Wprowadzenie do analizy skupień*, Kraków 1983.
Matyja A., Simiński K., *Comparison of Algorithms for Clustering Incomplete Data*, „Foundations of Computing and Decisions Sciences" 2014, vol. 39, no. 2.
Mądrzejowski W., *Przestępczość zorganizowana. System zwalczania*, Warszawa 2008.
Michalska-Warias A., *Pojęcie przestępczości zorganizowanej – aspekty kryminologiczne*, „Studia Iuridica Lublinensia" 2003, t. 1.
Murtagh F., Contreras P., *Algorithms for hierarchical clustering: an overview*, „Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery" 2012, vol. 2.
Porębski A., *Application of Cluster Analysis in Research on the Spatial Dimension of Penalised Behaviour*, „Acta Universitatis Lodziensis. Folia Iuridica" 2021, vol. 94.
Wagstaff K., *Clustering with Missing Values: No Imputation Required*, w: *Classification, Clustering, and Data Mining Applications*, red. D. Banks, L. House, F.R. McMorris, P. Arabie, W. Gaul, Berlin 2004.
Walesiak M., *Pomiar podobieństwa obiektów w świetle skal pomiaru i wag zmiennych*, „Prace Naukowe Akademii Ekonomicznej we Wrocławiu" 2002, t. 10, nr 950.
Walesiak M., *Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej*, „Przegląd Statystyczny" 2014, t. 61, nr 4.
Wierzchoń S., Kłopotek M., *Algorytmy analizy skupień*, Warszawa 2015.
Wójcik J., *Kryminologia. Współczesne aspekty*, Warszawa 2014.

Oprogramowanie wykorzystane do analizy

Bittinger K., *usedist: Distance Matrix Utilities*, R package version 0.3.0, 2019, < https://CRAN.R-project.org/package=usedist >.
R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna 2019, < https://www.R-project.org/ >.
Wickham H., Francois R., Henry L., Müller K., *dplyr: A Grammar of Data Manipulation*, R package version 0.8.3, 2019, < https://CRAN.R-project.org/package=dplyr/ >.
Wickham H., *ggplot2: Elegant Graphics for Data Analysis*, New York 2016, < https://CRAN.R-project.org/package=ggplot2/ >.