Kontrola i audyt

Analityka danych w audycie i kontroli

Wykorzystanie analizy skupień i jej prosta implementacja

W artykule zaprezentowano jedną z metod stosowanych w audytorskiej analityce danych (ang. *Audit Data Analytics* – ADA) oraz w algorytmach uczenia maszynowego bez nadzoru, jaką jest analiza skupień. Znalazła ona praktyczne zastosowanie w bardzo różnych dziedzinach działalności wymagających podejmowania decyzji (np. w marketingu, bankowości, ubezpieczeniach, badaniach społecznych, medycynie, biologii), natomiast jej wykorzystanie w audycie i kontroli jest nadal dość ograniczone. Wynika to w dużej mierze z postrzegania analizy przez środowisko audytorskie i kontrolerskie jako zbyt skomplikowanej pod względem matematycznym i narzędziowym. W artykule omówiono jej poszczególne etapy, dokonano doboru miar i metod pod kątem ich przydatności w badaniach oraz zaproponowano prostą implementację wybranych, zapewniającą automatyzację analiz. Zwrócono uwagę na możliwości szerokiego zastosowania analizy skupień na różnych etapach audytu i kontroli, wychodzącą poza typowe wykorzystanie do wykrywania anomalii w audycie finansowym. Może być szczególnie przydatna dla najwyższych organów kontroli ze względu na szeroki zakres podmiotowy i przedmiotowy badań prowadzonych przez te instytucje. Dotyczy to również kontroli NIK, a w szczególności kontroli wykonania zadań.

WIESŁAW KARLIŃSKI

Założenia i podstawowe etapy analizy skupień

Analiza skupień (klasteryzacja) to proces grupowania zbioru obiektów danych w wiele grup (klastrów), tak aby obiekty w obrębie klastra były do siebie bardzo podobne, ale różniły się od obiektów w innych klastrach. Różnice i podobieństwa ocenia się na podstawie wartości zmiennych opisujących obiekty i często wykorzystuje się do tego miary odległości[1]. Metoda zaliczana jest do kategorii metod uczenia maszynowego bez nadzoru, a w odniesieniu do audytu do analityki danych (ang. *Audit Data Analitics* – ADA).

J. Korzeniowski[2] zwraca uwagę, że z wieloletnich doświadczeń w zakresie rozwoju i zastosowania analizy skupień wynika podział pełnej analizy na następujące etapy:

- wybór obiektów i zmiennych,
- wizualizacja obiektów (lub zmiennych),
- normalizacja zmiennych,
- wybór miary odległości pomiędzy obiektami,
- wybór metody grupowania obiektów,
- ustalenie liczby skupień,
- grupowanie obiektów – właściwy etap analizy skupień,
- ocena wyników grupowania,
- opis i profilowanie klas.

Wybór obiektów jest uzależniony od celów oraz zakresu badania i o ile w niektórych zastosowaniach (np. w marketingu) badanie jest oparte na próbie losowej, to w wypadku audytu i kontroli będziemy mieli najczęściej do czynienia z analizą kompletnego zbioru transakcji, spraw, podmiotów. Ze zbioru należy usunąć obiekty, w odniesieniu do których występują braki danych lub błędne dane, a to oznacza, że wybór obiektów jest ściśle powiązany z wyborem zmiennych.

Wybór zmiennych opisujących obiekty jest jednym z najistotniejszych i najtrudniejszych etapów badania. M. Walesiak[3] proponuje, aby podzielić go na dwie fazy:

- Faza I – ustalenie wstępnej listy na podstawie znajomości przedmiotu badania, dostępności stosownych danych i przy ewentualnej współpracy z ekspertami z danej dyscypliny;
- Faza II – redukcja wstępnej listy przez usunięcie zmiennych o małym stopniu zmienności oraz zmiennych wzajemnie powiązanych (o wysokim współczynniku korelacji liniowej r Pearsona).

Można również tworzyć nowe zmienne, eliminując w ten sposób powiązanie zmiennych oryginalnych (np. w miejsce wagi i wzrostu posłużyć się wskaźnikiem BMI) lub uzyskując istotne dla celów badania informacje (np. określić terminowość na podstawie dat lub wyznaczyć istotne wskaźniki). Określając docelowy zestaw

[1] J. Han, J. Pei, H. Tong: *Data Mining. Concepts and Techniques* (4th edition), Morgan Kaufmann Publishers, Cambridge, MA, USA, 2023.
[2] J. Korzeniowski: *Metody selekcji zmiennych w analizie skupień. Nowe metody*, wyd. UŁ, 2012.
[3] M. Walesiak: *Problemy selekcji i ważenia zmiennych w zagadnieniu klasyfikacji, Taksonomia 12*, „Prace Naukowe AE we Wrocławiu" nr 1076/2005.

zmiennych, należy również mieć na uwadze typ zmiennych: ilościowe (metryczne) albo jakościowe (porządkowe lub nominalne), a także ich ilość (przy liczbie zmiennych powyżej 4 i liczbie obserwacji rzędu setek tysięcy, obliczenia mogą być czasochłonne).

Wizualizacja danych na etapie poprzedzającym właściwą analizę pozwala na optyczne oszacowanie liczby skupień oraz ich struktury, co może mieć wpływ na dobór metod analizy. Dla zestawu dwóch zmiennych czytelny obraz uzyskujemy na wykresie punktowym (rozrzutu). W wypadku trzech zmiennych (X, Y i Z) można konstruować trójwymiarowy wykres rozrzutu (o ile dysponujemy stosownymi narzędziami) lub trzy wykresy płaskie (X-Y, X-Z i Y-Z). Uzyskanie trzeciego wymiaru dzięki wykorzystaniu wykresu bąbelkowego nie będzie skuteczne przy dużej liczbie obserwacji. Jeśli liczba zmiennych przekracza trzy, sprawa wizualizacji komplikuje się jeszcze bardziej i w efekcie może być niewykonalna na tym etapie.

Istotnym elementem większości metod wykorzystywanych w analizie skupień w odniesieniu do zmiennych o charakterze metrycznym jest pomiar odległości pomiędzy poszczególnymi obserwacjami. Spośród wielu miar odległości często stosuje się prostą w definicji odległość euklidesową, która dla pary obserwacji (A i B) opisanych za pomocą n zmiennych jest wyrażona wzorem:

$$d = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \cdots + (A_n - B_n)^2} \quad (1)$$

Ponieważ dla każdej zmiennej (1 do n) sumuje się kwadraty różnic wartości obserwacji, to w wypadku gdy zmienne prezentowane są na skalach o różnej rozpiętości (różniącej się często o jeden lub nawet kilka rzędów wielkości), zmienne o dużej rozpiętości rzutowałyby na cały wynik, a nawet dochodziłoby do sytuacji, że byłby on uzależniony od jednostek pomiarowych (np. od tego, czy zmienną podajemy obszar w m kw., km kw., czy w ha). Uniknięciu takiej sytuacji służy typowa normalizacja danych, czyli sprowadzenie ich do porównywalnej skali pomiarowej. Niektórzy badacze (np. G. Milligan[4]) uznają co prawda, że normalizacja danych nie jest obowiązkowa, a nawet może zaburzyć strukturę skupień, ale w zastosowaniu audytowym i kontrolnym, gdzie skale zmiennych mogą być bardzo różne, wydaje się być nieodzowna. Jedynie wtedy, gdy mamy do czynienia ze zmiennymi wyrażonymi w takich samych jednostkach i posiadającymi zbliżoną skalę wartości, można zrezygnować z normalizacji i dokonać klasteryzacji danych surowych. W literaturze przedmiotu (patrz m.in. praca M. Walesiaka[5]) prezentuje się różne metody normalizacji danych. Generalnie wzór na normalizację można zapisać w postaci formuły:

$$z_i = \frac{x_i - a}{b} \quad (2)$$

[4] G. Milligan: *Clustering Validation: Results and Implications for Applied Analyses, Clustering and Classification*, P. Arabie, L. Hubert, G. de Soete (red.), World Scientific, Singapore, 1996.

[5] M. Walesiak: *Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej*, „Przegląd Statystyczny", R. LXI, zeszyt 4/2014.

gdzie x_i – oryginalna (surowa) wartość zmiennej, a z_i – wartość po normalizacji, a, b – parametry zależne od rodzaju normalizacji

Najprostsze metody normalizacji, z jakimi spotykamy się w analizie skupień, to:

- normalizacja min-max, zwana unitaryzacją zerową, gdzie parametrowi *a* przypisuje się minimum zmiennej x, czyli a=min(X), parametrowi *b* jej rozstęp, czyli b=max(X)-min(X), a wynik mieści się w przedziale [0,1];
- klasyczna standaryzacja (z-score), gdzie parametr *a* to średnia a=m(X), parametr *b* odchylenie standardowe b=sd(X), a zmienna wystandaryzowana ma średnią m = 0 i odchylenie standardowe sd = 1.

Normalizacja min-max jest najbardziej czytelna, ale w sytuacji gdy występują obserwacje silnie odstające (ekstremalne), a w badaniu pełnych zbiorów danych w audycie i kontroli ma to często miejsce, staje się ona mało przydatna dla celów klasteryzacji. W efekcie jej zastosowania obserwacja ekstremalnie duża przyjmie po normalizacji wartości 1, a zdecydowana większość pozostałych obserwacji będzie skupiona w pobliżu wartości 0 (albo odwrotnie przy obserwacji ekstremalnie małej), czyli będą one w bardzo małym stopniu zróżnicowane. Dla przykładu, gdyby jedną ze zmiennych była liczba ludności, a populacją 314 gmin województwa mazowieckiego, to po unitaryzacji zerowej ok. 87% gmin uzyskałoby wartość poniżej 0,01.

Klasyczna standaryzacja radzi sobie z obserwacjami odstającymi znacznie lepiej, ale również w tym wypadku parametry (średnia, odchylenie standardowe), a tym samym wynik normalizacji są wrażliwe na dane ekstremalne.

Metodą normalizacji odporną na występowanie danych odstających jest tzw. standaryzacja pozycyjna. W tej metodzie normalizacji parametrem *a* jest mediana a=Me (X), a parametrem *b* – bezwzględne przeciętne odchylenie od mediany MAD(X) (ang. *median absolute deviation*), które często jest dodatkowo korygowane o współczynnik 1,4826, czyli b=1,4826MAD(X).

Do obliczania MAD stosuje się zależność:

$$MAD_{med} = Me(|x_i - Me|) \quad (3)$$

Zastosowany w powyższej formule dopisek „*med*" został wprowadzony celowo z myślą o czytelniku, który zetknął się być może z zupełnie odmiennym pojęciem MAD (ang. *mean absolute deviation*) stosowanym w analityce śledczej, a w szczególności w analizie Benforda[6].

Normalizacja pozycyjna oparta na MAD_{med} nie jest tak popularna jak dwie pozostałe, ale niektórzy autorzy podkreślają jej szczególną przydatność przy występowaniu danych odstających w analizie skupień w takich obszarach, jak analiza obrazów, biostatystyka[7] czy analizy finansowe

[6] M. Nigrini: *Forensic Analytics. Methods and Techniques for Forensic Accounting Investigation*. John Willey & Sons, New Jersey, 2011.

[7] D. M. Rocke & B. Durbin: *A model for measurement error for gene expression arrays*, „Journal of Computational Biology" nr 8/2001.

i ekonomiczne[8]. Metoda jest również stosowana w sytuacji, gdy chcemy integrować dane metryczne (ilościowe) z danymi o skali porządkowej[9].

Wybór metody normalizacji jest zależny od rozkładu danych (brak danych odstających, pojedyncze dane odstające czy rozkład asymetryczny) oraz celu badania (segmentacja obiektów, wykrycie pojedynczych anomalii czy tylko wizualizacja lub ranking). Zagadnienie to będzie omówione szerzej w dalszej części przy identyfikacji tzw. anomalii globalnych.

Według klasyfikacji zaproponowanej przez J. Han i in.[10] podstawowe metody analizy skupień podzielić można na trzy grupy:

- podziałowe (występujące często w literaturze pod nazwą niehierarchicznych), grupujące obiekty wokół środków klastrów (centroidów);
- hierarchiczne, polegające na tworzeniu drzewa klastrów (dendrogramu), pokazującego relacje między obiektami na różnych poziomach szczegółowości;
- oparte na gęstości, identyfikujące klastry jako obszary o wysokiej gęstości punktów.

Z kolei do grupy zaawansowanych autorzy zaliczają m.in. metody: modelowe i probabilistyczne, oparte na grafach oraz o wysokiej liczbie wymiarów. Za najprostsze w realizacji oraz najbardziej wydajne przy analizie dużych zbiorów danych uznawane są metody podziałowe.

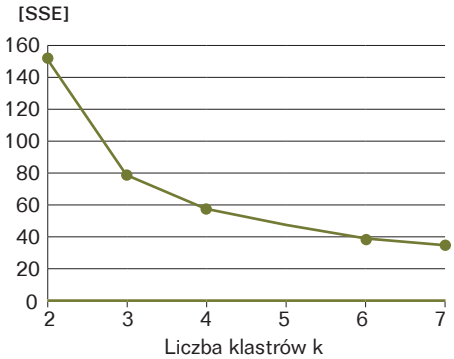
Proste metody analizy skupień w kontroli i audycie

W grupie metod podziałowych największą popularność w zastosowaniu do danych metrycznych (ilościowych) zdobyła metoda k-średnich (ang. *k-means*). Przypisanie obiektów do zadanej liczby k klastrów realizuje się w niej przenosząc obiekty między skupieniami, aż do momentu zoptymalizowania zmienności wewnątrz skupień i pomiędzy nimi. Proces ma charakter iteracyjny i realizowany jest w kilku krokach[11]:

1. Ustalenie liczby skupień;
2. Ustalenie warunku zatrzymania procedury (brak przesunięć obiektów między skupieniami oraz maksymalna liczba iteracji);
3. Wybór metryki, tj. sposobu pomiaru odległości pomiędzy obiektami;
4. Ustalenie środków skupień (centroidów): w pierwszej iteracji np. losowo, a w kolejnych np. jako średnia arytmetyczna współrzędnych punktów należących do danego skupienia;
5. Obliczenie odległości obiektów od środków skupień;
6. Przypisanie obiektów do skupień – porównujemy odległości każdej obserwacji od każdego ze skupień i przypisujemy ją do skupienia, do którego środka ma najbliżej;
7. Sprawdzenie warunku zatrzymania, a jeśli nie nastąpił – powrót do punktu 4.

[8] C. C. Aggarwal: *Outlier Analysis*. Springer, 2013.
[9] P. Filzmoser, R. Maronna, M. Werner: *Outlier detection in high dimensions*, „Computational Statistics & Data Analysis" nr 3/2008.
[10] J. Han, J. Pei, H. Tong: *Data Mining*..., op.cit.
[11] A. Królak-Nowak, K. Kotarba: *Podstawy uczenia maszynowego*, wyd. AGH, Kraków, 2022.

Rysunek 1. **Dobór liczby klastrów w metodzie *k-means* dla zbioru iris (dane surowe)**



Źródło: Obliczenia własne z zastosowaniem algorytmu opisywanego w dalszej części.

Metoda jest dość prosta w realizacji i szybka, natomiast wymaga wskazania *a-priori* liczby skupień, ma mniejszą skuteczność w wypadku skupień o kształcie niecentrycznym, jest wrażliwa na występowanie danych odstających, a ponadto daje mało stabilne wyniki, w związku z losowym doborem centroidów w pierwszej iteracji. Powoduje to konieczność jej wielokrotnego zastosowania i wyboru najkorzystniejszego wariantu.

W większości publikacji proponuje się, aby liczbę skupień k dobierać w sposób eksperymentalny, stosując tzw. zasadę łokcia (ang. *elbow*). Jest to heurystyczna metoda polegająca na obliczeniu sumy kwadratów odległości (SSE) każdego punktu danych od centroidu klastra dla różnych wartości k, a następnie prezentacji wyników na prostym wykresie. SSE maleje wraz ze wzrostem liczby skupień, natomiast za optymalną uznaje się taką wartość k, poniżej której tempo spadku SSE zmniejsza się. Rysunek obok prezentuje zastosowanie metody *elbow* dla zbioru danych iris[12], który jest klasycznym zbiorem używanym w statystyce i uczeniu maszynowym do testowania algorytmów, zawierającym opis trzech gatunków irysów (setoza, versicolor i virginica) z zastosowaniem czterech zmiennych. Jakkolwiek zbiór nie jest optymalny do zastosowania metody k-średnich, to widać, że punkt elbow odpowiada liczbie klastrów k=3.

Uniknięcie wpływu obserwacji skrajnie odstających (anomalii globalnych) na wynik k-means jest możliwe przez wykluczenie takich obserwacji po normalizacji, a przed właściwą procedurą klasteryzacji. W zastosowaniu audytorskim nie mamy do czynienia z usunięciem obserwacji ze zbioru, a jedynie z ich oflagowaniem, wyłączeniem z analizy skupień i poddaniem analizie szczegółowej w innym trybie. Do identyfikacji obserwacji odstających przy standaryzacji klasycznej i pozycyjnej można zastosować prostą w realizacji odległość euklidesową w przestrzeni standaryzowanej. Dla każdej wystandaryzowanej obserwacji oblicza się sumę kwadratów odchyleń od punktu centralnego, który po standaryzacji ma wartość zerową. Jako orientacyjne kryterium selekcji można przyjąć wartości progowe, korzystając z kwantyli rozkładu χ^2 (Chi-kwadrat). Kryterium selekcji obserwacji odstających (anomalii globalnych) dla n zmiennych można zatem zapisać w postaci:

[12] <https://ocw.mit.edu/courses/15-097-prediction-machine-learning-and-statistics-spring-2012/resources/iris/> (dostęp: 22.1.2026).

$$D_i^2 = z_{1i}^2 + z_{2i}^2 + \cdots + z_{ni}^2 > \chi^2(n, \alpha) \quad (4)$$

Próg istotności α można określić na poziomie równym 0,01.

Zaproponowany algorytm bazuje na pomiarze odległości od wartości typowej w całym zbiorze, którą jest średnia (w standaryzacji klasycznej) albo mediana w standaryzacji pozycyjnej.

Algorytmu nie można stosować przy klasteryzacji danych surowych oraz podlegających normalizacji min-max ze względu na brak, w sensie probabilistycznym, wartości typowej. W obu wymienionych przypadkach należałoby w miejsce odległości euklidesowej zastosować tzw. odległość Mahalanobisa od punktu centralnego, a następnie użyć w stosunku do niej przedstawionego powyżej kryterium selekcji opartego na rozkładzie Chi-kwadrat. Odległość Mahalanobisa jest dość popularną, choć nieco bardziej skomplikowaną miarą stosowaną w analizie skupień, uwzględniającą korelację między zmiennymi, dlatego zrezygnowałem z prezentacji jej aparatu matematycznego.

Wybór różnych metod normalizacji w kontekście anomalii globalnych można podsumować następująco. Standaryzacja pozycyjna skutecznie wyeliminuje wpływ pojedynczych wartości odstających, zachowując czytelne relacje pomiędzy pozostałymi obserwacjami, natomiast w wypadku gdy zbiór jest mocno asymetryczny, a górny „ogon" rozkładu obejmuje 10% i więcej obserwacji, może zaklasyfikować zbyt dużo obserwacji jako odstające. Przy standaryzacji klasycznej liczba anomalii globalnych będzie niższa, ponieważ dla tego typu rozkładów wartość średniej jest wyższa od wartości mediany. W badaniach audytowych i kontrolnych warto zastosować adaptacyjny schemat normalizacji, uzależniony od charakteru rozkładów zmiennych. W wypadku rozkładów zbliżonych do symetrycznych należałoby wykorzystać klasyczną standaryzację z-score. Z kolei dla danych o dużych wartościach odstających lub silnej asymetrii, typowej dla zmiennych finansowych i operacyjnych, warto zastosować standaryzację pozycyjną. Jeżeli udział obserwacji oznaczonych jako odstające w wyniku standaryzacji pozycyjnej będzie niewielki, to można potraktować je jako anomalie globalne. W przeciwnym razie, tę identyfikację należałoby zinterpretować jako wstępną separację jednostek o odmiennej skali działalności, a dalszą segmentację przeprowadzić na pozostałych obserwacjach, co poprawi jakość grupowania. Jeśli nie zamierzamy wykonywać wstępnej separacji obserwacji, a udział obserwacji zaklasyfikowanych jako anomalie globalne przekroczy użyteczny operacyjnie odsetek (np. 20%), lepiej zastąpić ją standaryzacją klasyczną.

Unitaryzację zmiennych (normalizację min-max) można natomiast stosować do wizualizacji oraz w wypadku gdy znaczenie ma pozycja, a nie skala wartości i w zbiorze nie ma wartości odstających. Użycie do klasteryzacji danych surowych dotyczy szczególnych sytuacji, gdy skala wartości zmiennych jest porównywalna, a jednostki takie same.

Jeśli chodzi o zastrzeżenia dotyczące stabilności wyników, to rozwiązaniem może być zastosowanie modyfikacji metody *k-means* optymalizującej dobór punktów startowych. Stosowny algorytm nazwany *k-means++* został zaproponowany

w 2007 r.[13] i daje większą powtarzalność wyników. Zadowalający efekt uzyskamy również przy zwiększeniu liczby losowań w ramach klasycznej metody *k-means*, co zastosowano w opisywanych poniżej procedurach.

Zastosowanie metody *k-means* w audycie proponują m.in. S. Thiprungsri i M. Vasarhelyi[14], którzy wykorzystali ją w praktyce do badania odszkodowań w grupowym ubezpieczeniu na życie, bazując na danych o charakterze ilościowym. Charakter danych ma tutaj kluczowe znaczenie, gdyż metoda *k-means* opiera się na pomiarze odległości (różnicy), a zatem może być zastosowana jedynie dla danych metrycznych (ilościowych). Dla danych porządkowych nie można wyznaczyć różnicy wartości, a jedynie relacje mniejsza/równa/większa, natomiast dla danych nominalnych można jedynie zastosować operację równości. W pewnych sytuacjach (jeśli pojedyncza zmienna jest istotna w badaniu i ma charakter porządkowy, a kafeteria odpowiedzi jest odpowiednio liczna, np. 5 lub więcej) można dopuścić uwzględnienie jej w procedurze *k-means* na równi z danymi ilościowymi. Należałoby jednak zastosować normalizację pozycyjną (z zastosowaniem MAD_{med}), a przy pomiarze odległości (wzór 1) można wprowadzić dla tej zmiennej odpowiednią wagę mniejszą od 1. Najlepszym rozwiązaniem jest jednak użycie metody przeznaczonej dla zmiennych jakościowych (porządkowych lub nominalnych), jaką jest metoda *k-modes*.

Metoda *k-modes* działa podobnie jak metoda k-średnich (patrz m.in. J. Han i in.[15]), przy czym:

- oryginalne dane nie muszą mieć postaci numerycznej i nie podlegają normalizacji;
- miarą podobieństwa obiektu A i B nie jest odległość euklidesowa (A-B), a liczba zmiennych, dla których Ai=Bi;
- przy aktualizacji centroidów używa się mody (dominanty), czyli najczęściej występującej kategorii w klastrze, w miejsce wartości średniej.

Algorytm można opisać następująco:

1. Wybiera się k początkowych centroidów, po jednym dla każdego klastra;
2. Przypisuje się obiekt do klastra, którego centroid jest mu najbliższy, używając miary podobieństwa;
3. Dla każdego klastra aktualizuje się centroid, obliczając modę (dominantę) każdej cechy w obrębie tego klastra;
4. Powtarza się kroki 2 i 3, aż nastąpi brak zmian w przynależnościach obiektów do klastrów lub zostanie przeprowadzona określona liczba iteracji.

Miarą jakości doboru w metodzie *k-modes* nie jest wartość SSE, a minimalizacja tzw. funkcji kosztu (J), zdefiniowanej jako łączna liczba niezgodności cech pomiędzy obserwacjami, a dominantami klastrów. Funkcję kosztu wykorzystuje się również do doboru optymalnej

[13] D. Arthur, S. Vassilvitskii: *k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 2007.

[14] S. Thiprungsri, M. Vasarhelyi: *Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach*, „The International Journal of Digital Accounting Research", Vol. 11, 2011.

[15] J. Han, J. Pei, H. Tong: *Data Mining*..., op.cit..

liczby klastrów k, stosując analogiczną jak w *k-means* zasadę *elbow*.

W praktyce audytorskiej spotkamy się często z sytuacją, gdy w zbiorze danych opisujących jakiś proces występują zarówno zmienne metryczne (ilościowe), jak i zmienne jakościowe. Istnieje co prawda podejście hybrydowe nazywane *k-prototypes*, które łączy obie metody (*k-means* i *k-modes*) w jednym, ale stosowanie takiego rozwiązania rodzi szereg komplikacji (np. kwestia doboru optymalnej liczby klastrów), a uzyskane wyniki są mniej czytelne dla odbiorcy. Lepszym rozwiązaniem jest zatem podział zmiennych na dwie grupy (ilościowe oraz jakościowe) i odrębne zastosowanie klarownych metod dla obu grup. Podział na dwie kategorie zmiennych zastosowali D. Wei i in.[16] przy badaniu kompletu zapisów księgowych. Dla zmiennych ilościowych proponują oni metodę LOF (ang. *Local Outlier Factor*), a dla zmiennych jakościowych metodę grupowania *k-modes*. Metoda LOF jest oparta na analizie gęstości i nie należy do klasycznych metod analizy skupień, a służy do identyfikacji wartości odstających, co odpowiada celowi badania, jakim było wykrycie wartości odstających na poziomie transakcji.

Po zastosowaniu dwóch algorytmów (w tym wypadku *k-means* oraz *k-modes*) uzyskujemy dla każdej obserwacji przypisanie do klastra ilościowego oraz niezależne przypisanie do klastra jakościowego. Oba wyniki można połączyć tworząc macierz klastrów, co jest mechanizmem powszechnie stosowanym w audycie, np. przy analizie ryzyka (macierz ryzyka). Należy jednak pamiętać, że numery klastrów uzyskane po operacji grupowania nie mają znaczenia merytorycznego, dlatego warto przed utworzeniem macierzy przypisać im etykiety (numery porządkowe lub nazwy) odpowiadające znaczeniu klastra (ang. *post-hoc cluster labeling*). Dla danych numerycznych wyznacznikiem znaczenia klastra mogą być np. duże wartości średnie jakichś zmiennych w klastrze, a dla danych jakościowych np. określone typy spraw. Poprawnej interpretacji wyników oraz identyfikacji rzadkich kombinacji cech będzie służyło podanie w macierzy *k-means x k-modes* liczebności klastrów z ewentualnym uzupełnieniem o dane procentowe. Taka tabela jest czytelna i nie wymaga dodatkowych narzędzi do interpretacji.

Gdyby okazało się, że w zbiorze danych występuje pojedyncza zmienna ilościowa, a pozostałe mają charakter jakościowy, to można przyjąć rozwiązanie polegające na konwersji zmiennej ilościowej drogą stratyfikacji do postaci porządkowej i następnie zastosować wyłącznie metodę *k-modes*.

Analizę skupień wykorzystuje się często w audycie do identyfikacji anomalii. S. Thiprungsri i M. Vasarhelyi[17] zwracają uwagę, że za anomalie można uznać:

---

[16] D. Wei, S. Cho, M. Vasarhelyi, L. Te-Wierik: *Outlier Detection in Auditing: Integrating Unsupervised Learning within a Multilevel Framework for General Ledger Analysis*, „Journal of Information Systems", nr 2/2024, American Accounting Association.

[17] S. Thiprungsri, M. Vasarhelyi: *Cluster Analysis for Anomaly*..., op.cit.

• obserwacje, które nie należą do żadnego klastra;
• obserwacje najbardziej odległe od centroidu klastra;
• obserwacje tworzące małe lub rzadkie klastry.

W opisanej powyżej metodzie analizy obserwacje odstające na poziomie globalnym (nienależące do żadnego klastra) identyfikujemy stosując mechanizm wyrażony wzorem (4). Identyfikacja małych klastrów jest prosta, gdyż odbywa się przez zliczenie liczby obserwacji w każdym klastrze. Klasyczna metoda *k-means* nie radzi sobie natomiast z identyfikacją obserwacji odległych od cetroidów (lokalnych anomalii) oraz rzadkich klastrów. Problemowi temu można zaradzić rozszerzając standardową funkcjonalność *k-means* o pomiar i interpretację odległości. Dla każdej obserwacji uzyskujemy wówczas, poza wskazaniem numeru klastra, również wartość odległości obserwacji (d_i) od centroidu. Należy w tym miejscu podkreślić, że w algorytmie *k-means* centroidy mają charakter punktów teoretycznych i nie muszą odpowiadać żadnej rzeczywistej obserwacji.

Jeśli chcielibyśmy zastosować miarę odległości porównywalną dla różnych klastrów, to należałoby dokonać swoistej normalizacji wartości d_i, np. dzieląc ją przez medianę[18] obliczaną dla każdego klastra k:

$$d'_i = \frac{d_i}{Me(d_k)} \qquad (5)$$

Proces identyfikacji anomalii lokalnych można wówczas zautomatyzować, przyjmując zasadę:
• $d'_i < 2$ – obserwacja typowa,
• $d'_i = 2 - 3$ – obserwacja nietypowa,
• $d'_i = 3$ i więcej – obserwacja odstająca.

W sytuacji gdy liczba zmiennych przekracza 5 warto podwyższyć zaproponowane powyżej wartości progowe (2 i 3), wprowadzając współczynnik korygujący, np. $(n/3)^{1/2}$, gdzie n – liczba zmiennych.

Implementacja analizy skupień w audycie i kontroli

P. Byrnes[19] zwraca uwagę, że realizacja poszczególnych etapów klasteryzacji krok po kroku jest dużym wyzwaniem dla typowych audytorów. Należałoby jego zdaniem zmierzać do automatyzacji tego procesu, co pozwoliłoby skupić się na interpretacji uzyskanych wyników. Istotna jest też kwestia znajomości odpowiednich narzędzi analitycznych. Komercyjne pakiety statystyczne czy też popularne otwarte narzędzia informatyczne udostępniają procedury i biblioteki przeznaczone do analizy skupień (w Pythonie biblioteki Scikit-learn, NumPy i Pandas, natomiast w R –pakiety cluster, stats i flexclust). Są to jednak narzędzia na razie niezbyt popularne wśród audytorów i kontrolerów. Założeniem artykułu jest prezentacja możliwie prostych rozwiązań dotyczących analizy skupień, funkcjonujących półautomatycznie i zaimplementowanych w środowisku znanym audytorom i kontrolerom,

[18] Można również rozważyć normalizację odstępem międzykwartylowym.
[19] P. Byrnes: *Automated Clustering for Data Analytics*, *Journal of Emerging Technologies in Accounting* nr 2/2019, American Accounting Association.

jakim jest MS Excel. Dlatego podjąłem się próby opracowania w języku VBA i przetestowania procedur o roboczej nazwie *Audit_kMeans, Audit_kMeans_LocalRisk* i *Audit_kModes*, które realizują opisywane wcześniej metody *k-means*, *k-means* z identyfikacją anomalii lokalnych oraz *k--modes*.

Procedury zakładają, że dane wejściowe zapisano w arkuszu Excela w kolumnach (stosownie do liczby zmiennych), począwszy od kolumny A, natomiast pierwszy wiersz zawiera nazwy zmiennych.

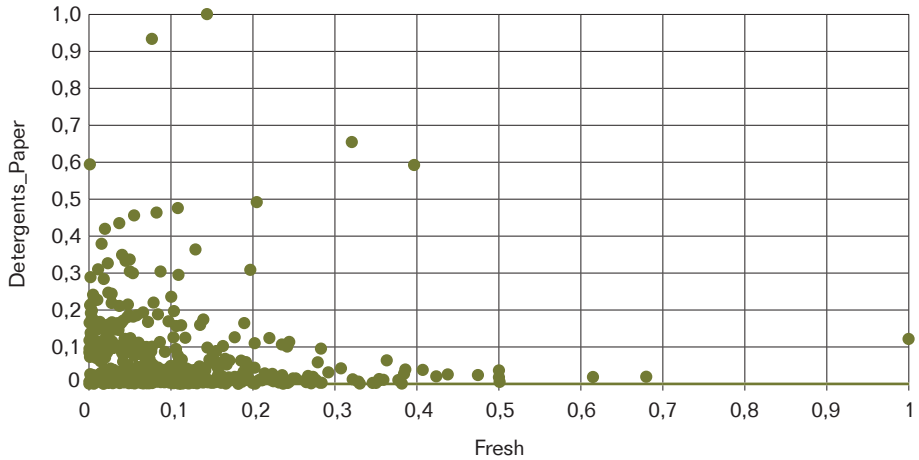
Procedura Audit_*kMeans* wymaga podania parametrów wejściowych: liczby zmiennych, metody normalizacji (klasyczna, pozycyjna, min-max albo dane surowe) oraz maksymalnej liczby skupień (k_{max}). Dane podlegają automatycznej normalizacji oraz weryfikacji pod kątem występowania obserwacji odstających (anomalii globalnych), a adnotacja o wyniku weryfikacji zapisywana jest w arkuszu. Przy normalizacji min-max oraz przy danych surowych anomalie globalne wyznaczane są z użyciem odległości Mahalanobisa. Operacja klasteryzacji wykonywana jest metodą k-means dla kolejnej liczby klastrów (k), począwszy od dwóch, a skończywszy na wskazanej wartości maksymalnej (k_{max}). Z analizy wyłączone są automatycznie obserwacje uznane za anomalie globalne. Dla każdej wartości k dokonuje się określonej liczby losowań (domyślnie 10) i zapisuje w arkuszu ich najlepszy wynik oraz odpowiadającą im wartość błędu SSE. Maksymalna liczba iteracji w każdym losowaniu wynosi domyślnie 100. Cała operacja jest automatyczna, a po wykonaniu procedury wystarczy sporządzić prosty wykres liniowy zależności SSE od k (patrz rysunek 1) i zdecydować, którą wartość k z zakresu 2 - k_{max} oraz odpowiadający jej wynik (przypisanie obserwacji do klastrów) uznać za optymalny.

Po wyborze optymalnej liczby klastrów k warto zastosować procedurę *Audit_kMeans_LocalRisk*. Procedura jest wykonywana dla wskazanej liczby zmiennych, typu normalizacji (metody j.w.) oraz liczby klastrów. W wyniku jej wykonania uzyskujemy dla każdej obserwacji: przypisanie do numeru klastra, wartość odległości od centroidu, znormalizowaną medianowo wartość odległości oraz adnotację dotyczącą typowości/nietypowości (typowy/ nietypowy/ odstający). Dodatkowo w procedurze przewidziano dla celów prezentacyjnych możliwość zapisu do arkusza danych znormalizowanych.

Procedura *Audit_kModes* jest obsługiwana w analogiczny sposób jak *Audit_kMeans*: tj. należy podać liczbę zmiennych jakościowych oraz maksymalną liczbę klastrów, ale bez wskazywania metody normalizacji. Po zakończeniu obliczeń prezentowane są najlepsze wyniki losowań (domyślnie po 10 losowań) oraz wartości funkcji kosztu (J) dla każdej liczby klastrów. Również w wypadku *k-modes* nie zdecydowano się na automatyczny dobór liczby klastrów k, pozostawiając to do decyzji użytkownika.

Procedury przetestowano na kilku rodzajach danych referencyjnych. Dla wspomnianego wcześniej zbioru testowego iris uzyskano metodą *k-means*, przy normalizacji klasycznej, zgodność na poziomie: setoza 50/50, versicolor 39/50 i virginica 36/50, natomiast przy użyciu danych surowych, na poziomie: setoza 50/50, versicolor 48/50 i virginica 36/50. Użycie danych surowych było uzasadnione z uwagi na podobną skalę

Rysunek 2. **Rozrzut zmiennych Fresh oraz Detergents_Paper po unitaryzacji**



Źródło: Opracowanie własne.

wartości zmiennych i dało nieco lepszą trafność typowania, choć zbiór jest specyficzny, gdyż klastry dla versicolor i virginica nie są kuliste i nakładają się na siebie.

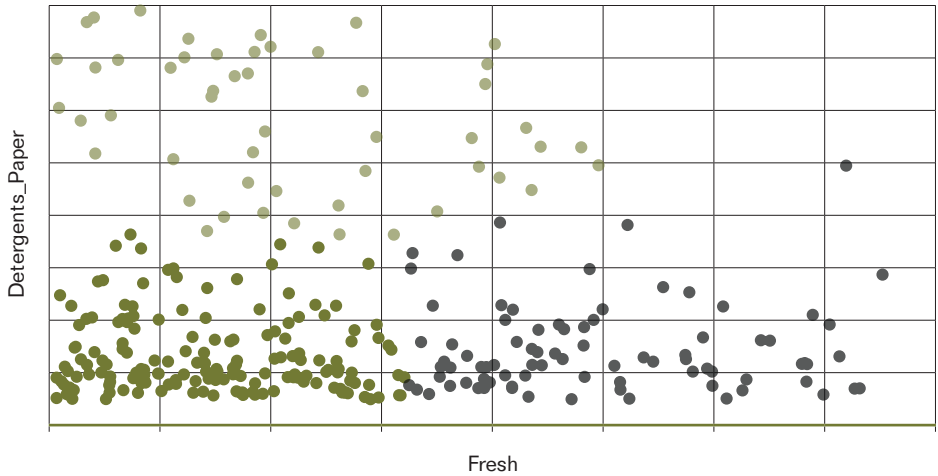
W celu oszacowania wydajności procedury *Audit_kMeans_LocalRisk* przeprowadzono analizę na zbiorze zapisów księgi głównej zawierającym trzy zmienne ilościowe i około 230 000 obserwacji, uzyskując czas analizy poniżej 2 min, co powinno być satysfakcjonujące dla audytorów finansowych.

Do demonstracji skuteczności różnych rozwiązań analitycznych w ramach analizy skupień wykorzystano zbiór danych *Wholesales_customer* dostępny w repozytorium uczenia maszynowego UCI[20]. Zbiór zawiera dane dotyczące 440 odbiorców 6 różnych kategorii towarów (*Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen*). Dodatkowo w zbiorze zawarto dwie zmienne jakościowe (*Channel i Region*). Przyjęto założenie, że celem badania jest segmentacja odbiorców na etapie przygotowania audytu lub raportu końcowego.

Zmienne ilościowe poddano analizie pod kątem korelacji wzajemnej (opcja „Korelacja" w dodatku Excela „Narzędzia analizy danych"), identyfikując silną korelację pomiędzy niektórymi z nich. Dla celów demonstracyjnych zdecydowano o wyborze dwóch zmiennych, które wskazują na rodzaj działalności odbiorcy oraz intensywność operacyjną (*Fresh* oraz *Detergents_Paper)* i nie są skorelowane. Wykres rozrzutu dla zmiennych *Fresh* oraz *Detergents_Paper* po unitaryzacji zerowej przedstawiono na rysunku nr 2.

[20] <https://archive.ics.uci.edu/dataset/292/wholesale+customers> (dostęp: 19.2.2026).

Rysunek 3. **Podział odbiorców na klastry (z wyłączeniem jednostek o bardzo dużych obrotach)**



Źródło: Opracowanie własne.

Przy zastosowaniu do tych zmiennych dwóch wymienionych wcześniej metod normalizacji uzyskano następującą liczbę anomalii globalnych: 18 – dla standaryzacji klasycznej, 135 – dla standaryzacji pozycyjnej. Bardzo duży odsetek (ok. 34%) obserwacji odstających przy standaryzacji pozycyjnej wynika z asymetrii rozkładów zmiennych i wskazuje na jednostki o odmiennej skali działalności. Do celów segmentacji odbiorców zdecydowano się uznać je za odrębną kategorię, a nie anomalie globalne. Taka operacja poprawiła jakość grupowania pozostałych odbiorców (bardziej jednoznacznie wyartykułowany punkt *elbow*, wskazujący na trzy klastry). W efekcie zastosowania procedury *Audit_kMeans_LocalRisk* dokonano grupowania pozostałych 305 odbiorów na trzy klastry, które stosownie do właściwości nazwano: 1- mix zakupów (poziom umiarkowany), 2- przewaga chemii, 3- przewaga żywności. Równocześnie procedura wykazała, że występuje 6 obiektów nietypowych w klastrze 1 oraz 5 obiektów nietypowych i 1 odstający w klastrze 3. Efekt grupowania odbiorów przedstawiono graficznie na rysunku 3.

W stosunku do dwóch zmiennych jakościowych znajdujących się w zbiorze *Wholesales_customer* (*Channel* i *Region*) podjęto próbę zastosowania analizy skupień metodą *k-modes*. Procedura *Audit_kModes* wskazała na optymalną liczbę 5 skupień, podczas gdy obie zmienne zawierały łącznie 2x3=6 różnych wartości. Analiza skupień nie wniosła zatem istotnej redukcji struktury względem klasycznej tabeli kontyngencji, co wynika z niewielkiej liczby kategorii oraz ich jednoznacznych kombinacji. Wobec powyższego ograniczono się do użycia jednej zmiennej jakościowej

w celu warstwowania wyników *k-means*. Wybrano zmienną, która opisuje tryb działania procesu, a w tym zbiorze najlepiej spełnia to zmienna odpowiadająca za kanał dystrybucji (*Channel*). Wyniki analizy można zaprezentować w postaci poniższej tabeli zawierającej liczebności poszczególnych kategorii.

Tabela 1. **Zestawienie wyników klasteryzacji ze względu na zmienne ilościowe i jakościowe**

k-means/ k-modes	Kanał 1	Kanał 2	Razem
1-mix zakupów	166	5	171
2-przewaga chemii	23	28	51
3-przewaga żywności	79	4	83
Bardzo duże obroty	30	105	135
Ogółem	298	142	440

Źródło: Opracowanie własne.

Tabela pokazuje, że wstępna separacja jednostek o bardzo dużej skali działalności umożliwiła uzyskanie stabilnej i interpretowalnej segmentacji pozostałych obserwacji, tj. na oddzielenie profilu zakupowego od efektu skali. Widać również zróżnicowanie kanałów dystrybucji: kanał 2 jest skoncentrowany na dużych obrotach, a kanał 1 zróżnicowany strukturalnie. Taki podział można wykorzystać przy ocenie ryzyka, planowaniu strategii audytu, doborze jednostek czy prezentacji wyników audytu.

W zamieszczonym powyżej przykładzie celowo, z uwagi na łatwość prezentacji graficznej, ograniczono liczbę zmiennych ilościowych do dwóch, co może budzić wątpliwości dotyczące zasadności stosowania metody i możliwości zastąpienia jej prostym filtrowaniem. Zasadniczą różnicę jakościową można zaobserwować dopiero przy większej liczbie zmiennych, kiedy filtrowanie jest zdecydowanie niewystarczające. Dodatkowym atutem metody jest automatyzacja całego procesu, pozwalająca audytorowi/kontrolerowi skupić się na zagadnieniach merytorycznych.

Obszary zastosowania analizy skupień

Spośród licznych zastosowań analizy skupień w biznesie, marketingu, biologii, medycynie, badaniach społecznych czy analizie obrazów przeważa cel segmentacji zbiorów danych do określonych celów lub wykrywania zachowań nietypowych. Natomiast w audycie i kontroli koncentruje się ona obecnie na wykrywaniu anomalii i dotyczy przede wszystkim audytu finansowego. Doświadczenia autora w obszarze metodyki planowania oraz prowadzenia kontroli i audytu wskazują, że takie ograniczenie zastosowania analizy skupień nie wykorzystuje potencjału tej metody. Analiza skupień w audycie i kontroli[21] może pełnić nie tylko funkcję detekcyjną, lecz także eksploracyjną,

[21] Na potrzeby artykułu nie różnicowano tych form działalności w odniesieniu do najwyższych organów kontroli (NOK), przyjmując, że chodzi tu o zewnętrzne badanie prowadzone przez nie w ramach posiadanych kompetencji, niezależnie od specyfiki każdego z nich wynikającej z krajowego ustawodawstwa. Prezentowana metoda ma bowiem wymiar uniwersalny i może dotyczyć każdej analizy odbywającej się w ramach zewnętrznego audytu/kontroli, zarówno w organach kontroli państwowej, jak i innych instytucjach zajmujących się audytem.

planistyczną i benchmarkingową, wspierając ocenę ryzyka, projektowanie procedur audytowych oraz interpretację złożonych populacji danych. Funkcjonalności te mogą okazać się szczególnie przydatne dla najwyższych organów kontroli z uwagi na szeroki zakres przedmiotowy i podmiotowy ich działalności kontrolnej. W wypadku Najwyższej Izby Kontroli analizę skupień można zastosowań do kontroli planowych, a w szczególności kontroli koordynowanych, ale również niektórych kontroli doraźnych, ukierunkowanych na zbadanie określonych problemów.

Potencjalne obszary wykorzystania metody w audycie i kontroli to:
1. Segmentacja procesów i jednostek: analiza skupień pozwala pogrupować jednostki podlegające kontroli w zależności od podobieństwa ich wskaźników finansowych i operacyjnych, dzięki czemu można wskazać grupy podmiotów o wysokim ryzyku (istotne w audycie poświadczającym), ale również odpowiednio dobrać do jednostki w kontroli wykonania zadań. Z kolei na etapie sporządzania raportu (informacji o wynikach kontroli) segmentacja (np. na podstawie danych uzyskanych w trakcie kontroli) pozwoli na prezentację bardziej wyważonych ocen;
2. Benchmarking i ocena efektywności: grupowanie jednostek wg kosztów jednostkowych, produktywności, wskaźników operacyjnych w celu uniknięcia fałszywych wniosków i porównywania tylko podobnych do siebie;
3. Tworzenie planu audytu/kontroli: grupowanie obszarów wg istotności, złożoności procesów, wyników wcześniejszych kontroli;
4. Identyfikacja anomalii i nieprawidłowości: wyłonienie obserwacji znacząco różniących się od reszty, nie tylko w audycie finansowym, ale również w kontroli zgodności (identyfikacja wzorców naruszeń) i wykonania zadań (np. identyfikacja podejrzanych transakcji w analizie kosztów);
5. Identyfikacja nietypowych procesów (nie transakcji) oraz wykrycie alternatywnych ścieżek procesów;
6. Analiza dostawców i zamówień publicznych: grupowanie dostawców wg liczby umów, trybu udzielania zamówień, powtarzalności kwot, relacji czasowych;
7. Analiza zachowań użytkowników systemów ERP: grupowanie pod kątem częstotliwości operacji, zakresu uprawnień, typów wykonywanych czynności.

Zestawienie potencjalnych obszarów wykorzystania analizy skupień z ukierunkowaniem na opisane metody (*k-means* i *k-modes*) przedstawia tabela 2, s. 23.

Z przedstawionego zestawienia wynika, że analizę skupień można zastosować w różnych rodzajach audytu i kontroli, a w szczególności w kontroli wykonania zadań. Może ona posłużyć do:
- identyfikacji jednostek, procesów lub programów funkcjonujących istotnie gorzej od pozostałych i w efekcie bardziej obiektywnego wytypowania obszarów wymagających pogłębionego badania;
- doboru jednostek i spraw do badania z uwzględnieniem heterogeniczności klastrów (dobór warstwowy optymalny Neymana);
- wykrywania ukrytej nieefektywności przez grupowanie jednostek ze względu na relacje nakłady – wyniki.

Interesująco wypada również wykorzystanie analizy skupień do benchmarkingu

Tabela 2. **Możliwe wykorzystanie procedur *k-means* i *k-modes* w różnych obszarach audytu i kontroli**

Obszar audytu/kontroli	Cel analizy skupień	Typ danych	Metoda	Efekt
Planowanie	Dobór jednostek do kontroli	Ilościowe	k-means	Priorytetyzacja obszarów ryzyka
Sporządzanie raportu	Segmentacja ex post	Mieszane	k-means k-modes	Uzyskanie wyważonych ocen
Badanie sprawozdań finansowych	Segmentacja transakcji	Ilościowe	k-means	Grupowanie podobnych transakcji, wsparcie doboru próby
Badanie sprawozdań finansowych	Wykrywanie anomalii	Ilościowe	k-means z analizą anomalii	Identyfikacja nietypowych obserwacji
Badanie procesów	Typologia ścieżek procesów	Jakościowe	k-modes	Wykrycie niestandardowych procesów
Audyt IT	Profilowanie użytkowników	Mieszane	k-means k-modes	Identyfikacja konfliktów uprawnień
Wykonanie zadań	Benchmarking jednostek	Ilościowe	k-means	Ocena efektywności względnej
Wykonanie zadań	Ocena efektywności	Ilościowe	k-means	Identyfikacja słabych punktów
Badanie zgodności	Analiza naruszeń	Nominalne	k-modes	Identyfikacja wzorców naruszeń
Badanie zgodności	Analiza zamówień	Mieszane	k-means k-modes	Wykrywanie obejścia progów
Audyt danych	Ocena jakości danych	Ilościowe	k-means	Identyfikacja systemowych błędów
Dobór prób	Stratyfikacja	Ilościowe	k-means	Większa reprezentatywność próby

Źródło: Opracowanie własne.

jednostek w zestawieniu z analizą obwiedni danych (DEA)[22]. Klasteryzacja nie tworzy granicy efektywności, nie wskazuje wzorców „idealnych", a skupia się na grupowaniu jednostek podobnych, jest bardziej odporna na obserwacje nietypowe, a wynikające z niej konkluzje są lepiej odbierane przez kontrolowanych. W kontroli sektora publicznego, w której celem jest nie tyle maksymalizacja efektywności, lecz identyfikacja obszarów wymagających poprawy, analiza skupień stanowi praktyczną i komunikacyjnie bezpieczną alternatywę dla DEA. Z drugiej strony, analiza skupień nie mierzy wprost efektywności technicznej, choć może dotyczyć relacji nakłady – wyniki. Stąd też warto rozważyć rozwiązanie hybrydowe, polegające na zastosowaniu DEA nie do całej populacji, a do jednostek funkcjonujących w podobnych warunkach, wyłonionych dzięki analizie skupień.

[22] W. Karliński: *Metoda obwiedni danych (DEA) w kontroli wykonania zadań*, „Kontrola Państwowa" nr 4/2022.

Podsumowanie

Analiza skupień może być narzędziem wspomagającym decyzje organizacji audytorskich i kontrolnych oraz indywidualnych audytorów i kontrolerów dzięki segmentacji jednostek, identyfikacji ryzyka i anomalii oraz optymalizacji działań kontrolnych. Metoda może okazać się szczególnie przydatna dla instytucji funkcjonujących w sektorze publicznym, w tym najwyższych organów kontroli, ze względu na dużą liczbę i zróżnicowanie podmiotów podlegających ich ocenie.

Zastosowanie analizy skupień w audycie i kontroli powinno przyczynić się do:

- zwiększenia efektywności wykorzystania zasobów przez koncentrację na grupach wysokiego ryzyka;
- ułatwienia podejmowania decyzji dzięki syntetycznej charakterystyce grup obiektów;
- możliwości wykrycia struktur i powiązań w danych, które są trudne do zauważenia metodami tradycyjnymi;
- ograniczenia subiektywizmu w wyborze obiektów do audytu i kontroli;
- lepszej konstrukcji ocen w raportach z audytu i kontroli.

Brak wystarczająco dużej popularności analizy skupień w audycie i kontroli wynika w znacznej mierze z postrzegania jej jako zbyt skomplikowanej metodologicznie i narzędziowo oraz przeznaczonej do wąskich zastosowań (detekcja anomalii w audycie finansowym). W wypadku dużych organizacji audytorskich lub kontrolnych, posiadających odpowiednie działy analityczne, bariery metodologiczne i narzędziowe mają charakter drugorzędny, natomiast dla indywidualnych audytorów wydają się być kluczowe.

W artykule zwrócono uwagę na potencjalnie szerokie obszary zastosowań analizy skupień w audycie i kontroli, zaprezentowano istotne kwestie metodologiczne i propozycje praktycznych rozwiązań oraz zademonstrowano możliwość stosunkowo prostej implementacji wybranych metod analizy skupień na bazie popularnych narzędzi informatycznych.

dr inż. **WIESŁAW KARLIŃSKI**
specjalista z zakresu metodyki kontroli i zastosowania metod analitycznych

Słowa kluczowe: analiza skupień, wykorzystanie analizy skupień w audycie, metody segmentacji danych w audycie, identyfikacja anomalii, identyfikacja obszarów ryzyka

Bibliografia:

1. Aggarwal C. C. : *Outlier Analysis*. Springer 2013.
2. Arthur D., Vassilvitskii S.: *k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 2007.
3. Byrnes P.: *Automated Clustering for Data Analytics*, „Journal of Emerging Technologies in Accounting" nr2/20219, American Accounting Association.

4. Filzmoser P., Maronna R., Werner M.: *Outlier detection in high dimensions*, *Computational Statistics & Data Analysis* nr 3/2008.
5. Han J., Pei J., Tong H.: *Data Mining. Concepts and Techniques* (4th edition), Morgan Kaufmann Publishers, Cambridge, MA, USA, 2023.
6. Karliński W.: *Metoda obwiedni danych (DEA) w kontroli wykonania zadań*, „Kontrola Państwowa" nr 4/2022.
7. Korzeniowski J.: *Metody selekcji zmiennych w analizie skupień. Nowe metody*, wyd. Uniwersytetu Łódzkiego 2012.
8. Królak-Nowak A., Kotarba K.: *Podstawy uczenia maszynowego*, wyd. AGH, Kraków 2022.
9. Milligan G: *Clustering Validation: Results and Implications for Applied Analyses*, *Clustering and Classification*, P. Arabie, L.Hubert, G. de Soete (red.), World Scientific, Singapore 1996.
10. Nigrini M.: Forens*ic Analytics. Methods and Techniques for Forensic Accounting Investigation*, John Willey & Sons, New Jersey 2011.
11. Rocke D. M. & Durbin B.: *A model for measurement error for gene expression arrays*, „Journal of Computational Biology" nr 6/2001.
12. Thiprungsri S., Vasarhelyi M.: C*luster Analysis for Anomaly Detection in Accounting Data: An Audit Approach*, „The International Journal of Digital Accounting Research", Vol. 11/2011.
13. Walesiak M.: *Problemy selekcji i ważenia zmiennych w zagadnieniu klasyfikacji*, *Taksonomia 12*, „Prace Naukowe AE we Wrocławiu" nr 1076/2005.
14. Walesiak M.: *Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej*, „Przegląd Statystyczny", R. LXI, zeszyt 4/2014.
15. Wei D., Cho S., Vasarhelyi M.: *Te-Wierik L.: Outlier Detection in Auditing: Integrating Unsupervised Learning within a Multilevel Framework for General Ledger Analysis*, „Journal of Information Systems" nr 2/2024, American Accounting Association.

ABSTRACT

Use of Data Clustering and its Simple Implementation – Data Analytics in Auditing

Data clustering is the process of grouping a set of data objects into multiple groups (clusters) so that objects within a cluster are very similar to one another, but differ from objects in other clusters. Differences and similarities are identified based on the values of the variables describing the objects, and distance measures are frequently used for this purpose. The method belongs to the category of unsupervised machine-learning methods and, in the context of audit, to audit data analytics (ADA). Data clustering has found practical application in various decision-making areas (e.g. marketing, banking, insurance, social research, medicine, biology), while its application in auditing is still relatively limited. This is mostly because the audit community perceive it as overly complex, in terms of both mathematics and tools. The author discusses the subsequent stages of data clustering, presents selected measures and methods with a view to their application in auditing, and proposes a simple implementation of selected methods that allows for analysis automation. He also pays attention to a broad potential that

data clustering may have at different stages of an audit, beyond typical application of this method for detecting anomalies in financial audits. Such analysis may be particularly useful for Supreme Audit Institutions due to the wide thematic scope and coverage of audits they conduct. This also applies to audits carried out by NIK, especially performance audits. The issue is presented from a perspective of an auditor and engineer simultaneously, rather than from a purely academic one, and the proposed implementation is based on tools that are widely known to the community of auditors.

Eng. WIESŁAW KARLIŃSKI, PhD, specialist in the field of audit methodology and the application of analytical methods

Key words: data clustering, use of data clustering in auditing, data segmentation methods in auditing, anomaly identification, identification of risk areas, audit data analytics